Registry No. Pi, 9035-68-1; TSH, 9002-71-5; FSH, 9002-68-0; LH, 9002-67-9; CG, 9002-61-3; IGF-I, 67763-96-6; IGF-II, 67763-97-7; Cyt-*c*, 9007-43-6; Cyt-c_2, 9035-43-2; PI, 9001-84-7; RNase, 9001-99-4; Cyt-*c'*, 9035-41-0; insulin, 9004-10-8.

REFERENCES AND NOTES

(1) Davison, D.; Thompson, K. H. "A Non-Metric Sequence Alignment Program". *Bull. Math. Biol.* **1984**, *46*, 579-590, and references cited therein.
(2) Nishikawa, K.; Ooi, T. "Correlation of the Amino Acid Composition of a Protein to Its Structural and Biological Characters". *J. Biochem. (Tokyo)* **1982**, *91*, 1821-1824.
(3) Sneath, P. H. A.; Sokal, R. R. *Numerical Taxonomy*; W. H. Freeman: San Francisco, 1973.
(4) Ito, T.; Kodama, Y.; Toyoda, J. "A Similarity Measure between Patterns with Nonindependent Attributes". *IEEE Trans. Pat. Anal. Math. Intel.* **1984**, *PAMI-6*, 111-115.
(5) Hoel, P. G. *Introduction to Mathematical Statistics*, 4th ed.; Wiley: New York, 1971.
(6) Dayhoff, M. O. *Atlas of Protein Sequence and Structure*; National Biomedical Research Foundation: Washington, DC, 1972; Vol. 5 and subsequent supplements.
(7) Augston, J. G.; Minker, J. "An Analysis of Some Graph Theoretical Cluster Techniques". *J. Assoc. Comput. Mach.* **1970**, *17*, 571-588.

A New Algorithm for Selection of Synthetically Important Rings. The Essential Set of Essential Rings for Organic Structures

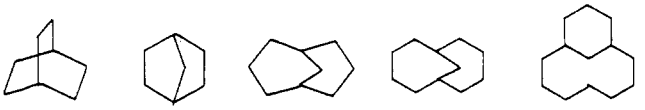
SHINSAKU FUJITA

Research Laboratories, Ashigara, Fuji Photo Film Co., Ltd., Minami-Ashigara, Kanagawa, 250-01, Japan

Received February 9, 1987

The concept of tied rings, multi-tied rings, and dependent rings is introduced, wherein transannular bonds and heterogeneity and abnormality of a ring are key classifiers. The essential set of essential rings (ESER) is defined as a set of rings other than tied, multi-tied, and dependent rings. An algorithm for detection of the ESER and its scope and limitations are discussed.

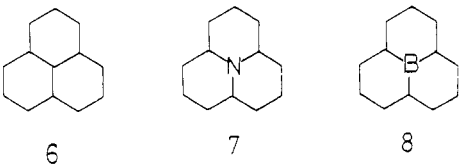
The perception of synthetically important rings is a crucial problem in the manipulation of organic structures by a computer. The smallest set of the smallest rings (SSSR) and its analogues have been widely adopted by computer systems for this purpose.[1] The SSSR is not unique in some cases when the equivalent sets are present in a given structural formula. For example, three 6-membered rings are equivalent in compound **1** and two rings are arbitrarily selected from the three. Corey's first criterion solved this difficulty by the concept of "collection of maximum proper covering sets of rings".[2] This approach is successful in obtaining all three rings of compound **1** but fails to select important rings for organic syntheses in some cases (e.g., **2**–**6**). The Corey's "synthetic subset" adopted



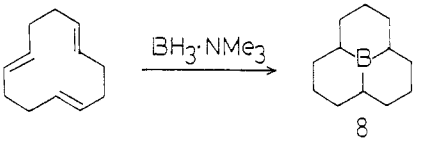
1 2 3 4 5

additional rings with six or fewer members.[3] This criterion is also successful in selecting a 6-membered ring along with two 5-membered rings (SSSR) from compound **1** and **2**. However, an 8-membered ring in compound **3** would be ignored by this procedure. Later, Wipke[4] chose the SSSR and all other rings with eight or fewer atoms. This principle, which is adequate for the purpose of abstracting 6- and 8-membered rings from **2** and **3**, respectively, is not fruitful in the cases of compounds **4** and **5**. A 10-membered ring in **4** and a 12-membered one in **5** are desirable to be adopted in a synthetic point of view. Since these rings in compounds **2**–**5** are in the same situation from the viewpoint of topology, they should be selected by a simple algorithm that meets our chemical sense. Fugamann's approach[5] gave satisfactory results in the above cases. But a more chemist-friendly algorithm is desirable.

A more delicate problem should be mentioned here. Three 6-membered rings should be selected from a carbocyclic compound (**6**) but a 12-membered one need not be chosen. However, the 12-membered rings of compounds **7** and **8** are



6 7 8



8

desirable to be selected, since the center atoms are a nitrogen and a boron atom, respectively. Let us consider that compound **8** is obtained from cyclododecatriene as follows. The 12-membered ring is important synthetically. Thus, a carbocyclic ring is to be preferred synthetically.

Although the importance of the concept of the SSSR is unchangeable now and in the future, a rational extension is desirable to solve the above-described problems. We propose here the essential set of essential rings (ESER), which is a simple algorithm to settle these problems.

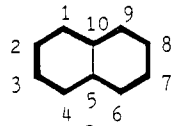
DEFINITION AND ALGORITHM OF ESER

Rings are classified as essential rings and nonessential rings. First, we define nonessential rings, which are tied rings, multi-tied rings, or dependent rings. Then ESER is defined as a set of rings other than nonessential rings.

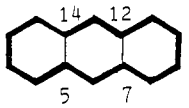
Tied Ring and Multi-Tied Ring. A tied ring is defined as a ring with one transannular bond that links directly two nonadjacent nodes of rings. For example, the 10-membered ring of compound **9** is a tied ring in which a bond between nodes 5 and 10 is a transannular bond defined as above. The tied rings are nonessential rings in any case, since they are

0095-2338/88/1628-0078$01.50/0

recognized as fused rings by chemists. Compound **9** is regarded usually to have a 6 + 6 fused ring and no 10-membered ring.



9



10

A multi-tied ring is defined as a ring that has two or more transannular bonds. A multi-tied ring is also nonessential and excluded. Three 6-membered, two 10-membered, and one 14-membered ring are detected in compound **10**. The two 10-membered rings are tied rings in accordance with the above definition and so are nonessential ones. The 14-membered ring, which has two transannular bonds (5–14 and 7–12), is a multi-tied ring and so is excluded. As a result, three six-membered rings remain as the ESER in the case of compound **10**.

Heterogeneity and Abnormality of Rings. We classify non-hydrogen atoms into three classes: (1) a carbon atom; (2) heteroatoms (N, O, S, and P); and (3) abnormal atoms (other than the above).[6] An index of heterogeneity (IH) of a ring is defined as the number of heteroatoms in the ring. An index of abnormality of a ring is the number of abnormal atoms in said ring. We classify rings in three categories: (1) a carbocyclic ring that contains neither heteroatoms nor abnormal atoms (IH = 0 and IA = 0); (2) a heterocyclic ring that involves at least one heteroatom and no abnormal atoms (IH > 0 and IA = 0); and (3) an abnormal ring that contains at least one abnormal atom (IA > 0). This classification provided chemist's intelligence to computer perception of rings to some extent as described below.

Dependent Rings. Let **B** be the set of all bonds in a given ring (R). Let T_r be the set of all tied rings.[7] The ring (R) is defined as a dependent ring when it is covered by the set (**T**) of tied rings[7] and the following covering conditions are fulfilled.

(1) Let **T** be a subset of T_r; **T** covers all bonds in **B**.

(2) All S ($\in$**T**) are the same size as or smaller than the ring (R).

(3) The intersection of all S ($\in$**T**) and R involves not less than half the bonds of S.

(4) All S ($\in$**T**) are of the same class (i.e., carbocyclic, heterocyclic, or abnormal) as said ring (R).

(5) All S ($\in$**T**) have the same or smaller value of IH (in the case of heterocyclic rings) or IA (in the case of abnormal rings) as compared with that of said ring (R).

A dependent ring is a nonessential ring. As shown in Figure 1, a 12-membered ring (**6-1**) is a dependent ring, since it is covered by two 10-membered rings that are selected from three of tied rings represented by heavy lines as **6-2**, **6-3**, and **6-4**. If the values of nodes of rings are ignored, as in this case, the above covering conditions are fulfilled. Therefore, the 12-membered ring **6-1** is a nonessential ring.

In the case of compound **7** (Figure 1), heterogeneity (IH) of the 12-membered ring **7-1** is different from that of 10-membered tied rings (**7-2** to **7-4**). Thus, the former ring (**7-1**) cannot be covered by the latter rings in any case. As a result, the 12-membered ring **7-1** is selected as an essential ring.

Similarly, the 12-membered ring **8-1** cannot be covered by three 10-membered tied rings (**8-2** to **8-4**) in light of abnormality (IA). That ring is also an essential ring.

Essential Set of Essential Rings. The essential set of essential rings (ESER) is defined as a set that contains all rings other than the above-defined tied, multi-tied, and dependent rings. Elements of the ESER are called *essential rings*. By this definition, 12-membered ring **6-1** does not belong to the

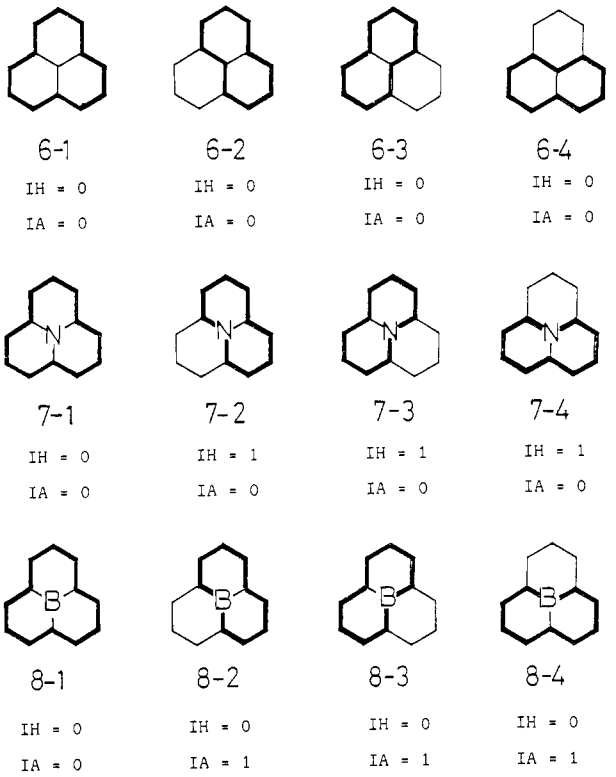


Figure 1. Heterogeneity (IH) and abnormality (IA) that affect selection of a dependent ring.

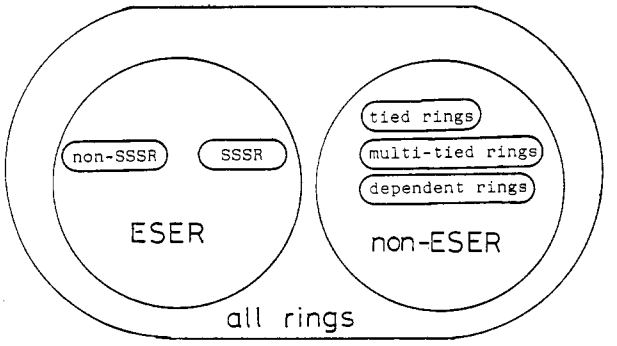


Figure 2. ESER, SSSR, tied rings, multi-tied rings, and dependent rings.

ESER. The corresponding rings **7-1** and **8-1** are selected as the elements of the ESER.

The ESER contains all rings of the SSSR (and, in some cases, several additional rings). This is proven as follows (Figure 2). No tied rings and no multi-tied rings are identical with the rings of the SSSR, since the former rings are recognized as fused rings of the corresponding components. No dependent rings are identical with the rings of the SSSR, since they are covered by several tied rings. Thus, the set of nonessential rings (not the ESER) contains no rings of the SSSR. Therefore, the remaining set (i.e., the ESER) contains all of the SSSR.

Algorithm for ESER. The basic strategy of selecting the ESER is composed of (1) detection of all rings and (2) exclusion of nonessential rings (tied, multi-tied, and dependent rings) and can be summarized in the following concise algorithm.

Step 1. Detect all rings.[8]

Step 2. Set the rings in ascending order of ring size.

Step 3. Calculate IH and IA.

Step 4. Detect and exclude tied rings and multi-tied rings.

Step 5. If a remaining ring is carbocyclic, compare this with carbocyclic tied rings (IH = 0, IA = 0). If the ring is covered

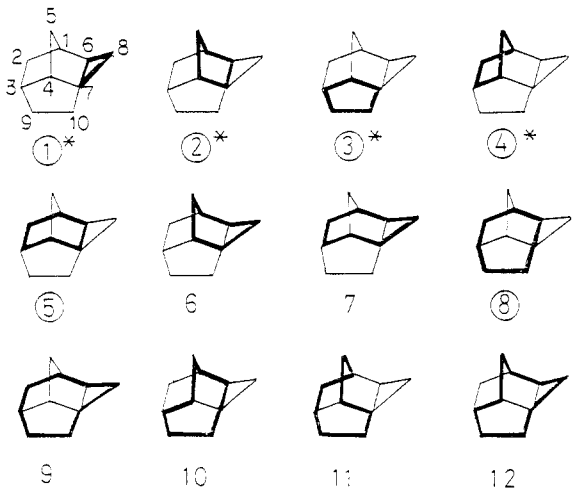


Figure 3. Twelve rings detected from compound **11**. The circled rings belong to the ESER. The rings marked by asterisks are of the SSSR.

by these in terms of the above-described covering conditions, exclude said ring as a dependent ring. Then go to step 8.

Step 6. If a remaining ring is heterocyclic, compare this with heterocyclic tied ring (IH > 0, IA = 0). If the ring is covered by these as above, exclude the ring as a dependent ring. Then go to step 8.

Step 7. If a remaining ring is an abnormal ring, compare this with abnormal tied rings (IA > 0). If the ring is covered as defined above, exclude it as a dependent ring. Then go to step 8.

Step 8. When examination is completed, all rings other than tied, multi-tied, and dependent rings are selected as the ESER. Done. Else go to step 5.

IMPLEMENTATION AND RESULTS

The present algorithm was programed in FORTRAN 77 and implemented on a VAX 11/750 (Digital Equipment Co.).[9]

In light of our ESER algorithm, the ESERs of compounds **1**–**5** are (6,6,6) for **1**, (5,5,6) for **2**, (6,6,8) for **3**, (7,7,10) for **4**, and (6,10,12) for **5**, wherein sizes of rings contained in the ESER are cited. Contrary to this, the SSSR are (6,6) for **1**, (5,5) for **2**, (6,6) for **3**, (7,7) for **4**, and (6,10) for **5**. The additional rings of the ESER as compared with the corresponding SSSR are bridged rings that are important synthetically.

Compound **11** has been reported to have 12 rings (Figure 3) from which 4 rings (marked with an asterisk) are selected as real rings.[2] These are identical with the SSSR in this case. On the other hand, six rings (circled rings of Figure 3) are selected to be members of the ESER. An output of the result is shown in Figure 4, where a ring of MLT = 0 is an element of the ESER but a ring of MLT ≠ 0 is nonessential. The newly assigned rings (no. 5 and 8) are bridged 6- and 7-membered rings, respectively. This selection is plausible if this case is compared with the above-described ESER of compounds **1**–**5**. It is noted that ring 8 cannot be covered by rings 10 and 11 because of the covering condition 2. The synthetic subset[3] of compound **11** would contain 6-membered rings 5 and 6 (Figure 3). However, ring 6 is regarded as a 3 + 5 fused ring. The method selecting all rings with eight or fewer atoms[4] would adopt rings 1–11. Among these, rings 6, 7, and 9–11 are recognized as fused rings by organic chemists.

From dodecahedrane (**12**), 1168 rings are detected, and 12 5-membered rings are selected as the ESER by the present algorithm. The SSSR of dodecahedrane contains 11 5-membered rings that are selected arbitrarily from the 12 equivalent 5-membered rings. The resulting 12 cases of the SSSRs are

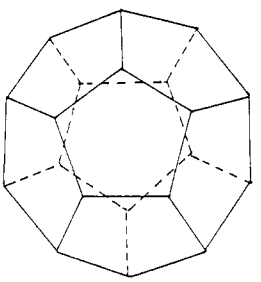
```
 NUMBER OF RINGS SELECTED:          11
 NUMBER OF TOTAL RINGS:             12

$$$$$$$$$$$$$$
$$$ITS RINGS$$$
$$$$$$$$$$$$$$
 %%%LIST OF RINGS%%%
 RING #  1;     SIZE = [  3]
   MEMBERS:   6   7   8
 RING #  2;     SIZE = [  5]
   MEMBERS:   1   5   4   7   6
 RING #  3;     SIZE = [  5]
   MEMBERS:   3   4   7  10   9
 RING #  4;     SIZE = [  5]
   MEMBERS:   1   2   3   4   5
 RING #  5;     SIZE = [  6]
   MEMBERS:   1   2   3   4   7   6
 RING #  6;     SIZE = [  6]
   MEMBERS:   1   5   4   7   8   6
 RING #  7;     SIZE = [  7]
   MEMBERS:   1   2   3   4   7   8   6
 RING #  8;     SIZE = [  7]
   MEMBERS:   1   2   3   9  10   7   6
 RING #  9;     SIZE = [  8]
   MEMBERS:   1   2   3   9  10   7   8   6
 RING # 10;     SIZE = [  8]
   MEMBERS:   1   5   4   3   9  10   7   6
 RING # 11;     SIZE = [  8]
   MEMBERS:   1   2   3   9  10   7   4   5

 %%%RING TYPE%%%
                          BOND TYPE      RING                 ESER
 RING #   SIZE       A     B     C      TYPE       MULTI      MLT
    1     [  3]      3     0     0   INVAR            0        0
    2     [  5]      5     0     0   INVAR            0        0
    3     [  5]      5     0     0   INVAR            0        0
    4     [  5]      5     0     0   INVAR            0        0
    5     [  6]      6     0     0   INVAR            0        0
    6     [  6]      6     0     0   INVAR            1        1
    7     [  7]      7     0     0   INVAR            1        1
    8     [  7]      7     0     0   INVAR            0        0
    9     [  8]      8     0     0   INVAR            1        1
   10     [  8]      8     0     0   INVAR            1        1
   11     [  8]      8     0     0   INVAR            1        1
```

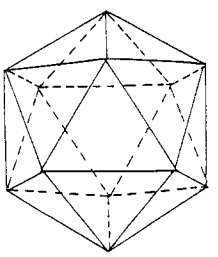
Figure 4. ESER detection of compound **11**. A ring of MLT = 0 is a member of the ESER. In this program, a multi-tied ring is neglected. Each ring corresponds to the counterpart listed in Figure 3.

equivalent to each other. On the other hand, the present ESER contains all 12 5-membered rings and thus gives the unique result



12

Icosahedron (**13**) involves 12878 rings, 20 3-membered rings of which are selected as the ESER, whereas the SSSR contains 19 nonunique 3-membered rings that are selected from those 20 3-membered rings.



13

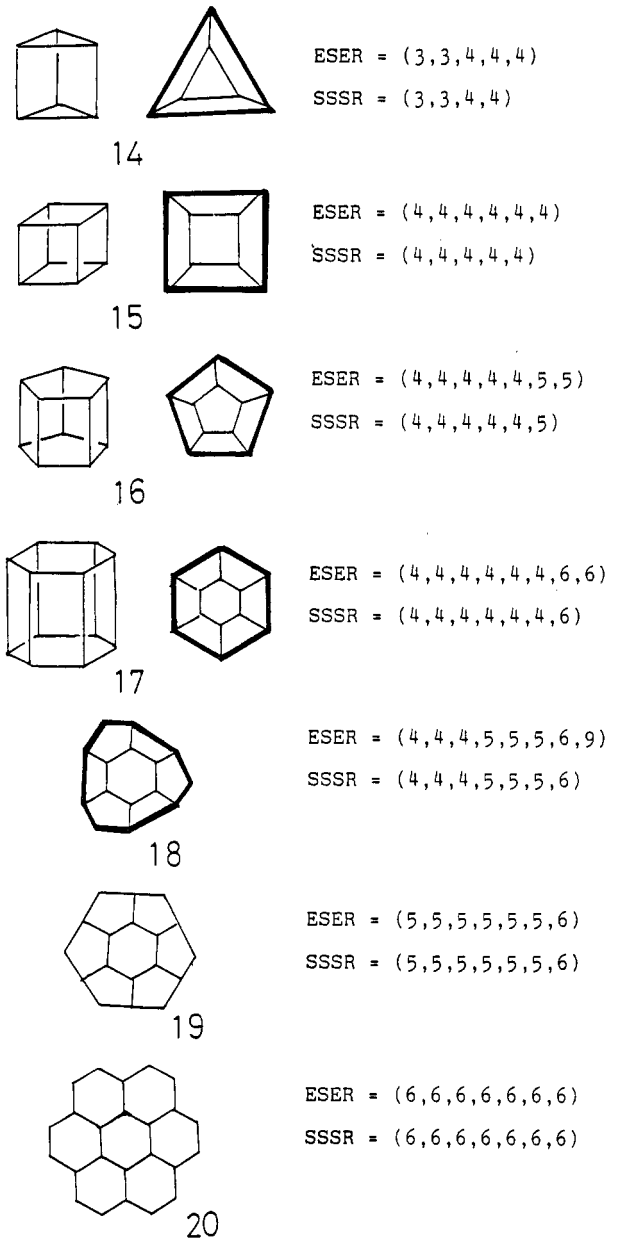


Figure 5. Scope and limitations of the ESER.

Cubane, or tetraprismane (**15**), has 28 rings in all, 6 4-membered rings of which are assigned to the ESER by the present algorithm (Figure 5). The ESER and SSSR of the related hydrocarbons are collected in Figure 5. Each rim (outer ring) shown by a heavy line is of the ESER in compounds **14–18**. However, the corresponding rings in compounds **19** and **20** do not belong to the ESER. It is noted that comparison between the two groups of compounds in Figure 5 would give insight into covering condition 3 as introduced above. Suppose that compounds **18** and **19** are bowls whose bottoms are hexagons. Compound **18** would be regarded as a deeper bowl than compound **19**. As a result, the rim (9-membered ring) is an ESER member, but the counterpart (12-membered ring) is not selected to be of the ESER.

Introduction of heterogeneity and abnormality of rings is effective in the selection of ESERs from compounds collected in Figure 6. Seven rings are detected as all rings in each case. Twelve-membered rings of compounds **6**, **21**, and **22** are nonessential rings by our ESER algorithm. On the other hand, the corresponding 12-membered rings of compounds **7**, **8**, and **23** belong to the respective ESER. These differences stem mainly from covering conditions 4 and 5.

Ferrocene is represented by structural formula **24** in

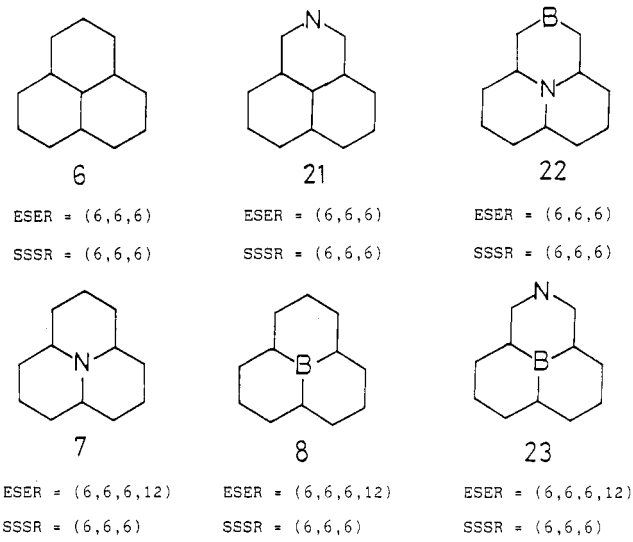
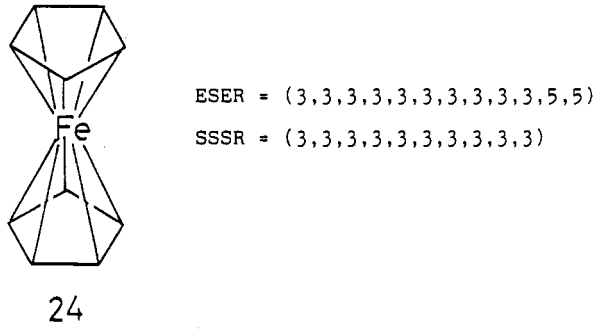


Figure 6. Comparison of the ESER with the SSSR. The ESER selects the synthetically important 12-membered rings from **7**, **8**, and **23**.

Chemical Abstracts.[10] This structure gives 10 3-membered, 10 4-membered, 12 5-membered, and 10 6-membered rings.



Among these rings, the 10 3-membered ones are selected as the SSSR. The two carbocyclic 5-membered rings are not selected by the SSSR algorithm, although they are important chemically. On the contrary, the present ESER approach affords two 5-membered rings along with the rings of the SSSR. The other 5-membered rings are omitted by the ESER algorithm, since they are unnecessary to perceive. The selection of the additional rings is also based on covering conditions 4 and 5.

Gasteiger and Jochum[1g] discussed in detail the problem of synthetically important rings. They concluded that, with other applications (e.g., synthetic strategy), a more careful analysis of the entire ring system is needed and that the task cannot be accomplished by an arbitrary selection of all rings up to a certain size. However, no practical algorithms were proposed for the task in their paper. The algorithm described in the present paper has solved several items indicated by them as well as other problems formulated here. The problem of synthetically important rings, however, cannot be fully solved if our consideration is restricted within the field of organic structure. We have discussed this point in the preceding paper.[11]

CONCLUSION

The definition of tied rings and multi-tied rings is given by counting transannular bonds. Dependent rings are derived from tied rings in light of heterogeneity and abnormality of the rings. Then the ESER is defined as a set of rings other than the above-defined three types of nonessential rings. An algorithm for selection of the ESER is discussed. The scope and limitations of the ESER are discussed as compared with the SSSR. The introduction of heterogeneity and abnormality

of a ring represents chemist's intelligence to a certain extent.

REFERENCES AND NOTES

(1) (a) Frèrejacque, M. *Bull. Soc. Chim. Fr.* **1939**, *6*, 1008. (b) Plotkin, M. *J. Chem. Doc.* **1971**, *11*, 60. (c) Zamora, A. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 40. (d) Bersohn, M. *J. Chem. Soc., Perkin 1* **1973**, 1239. (e) Esak, A. *J. Chem. Soc., Perkin 1* **1975**, 1120. (f) Schmidt, B.; Fleischhauer, J. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 204. (g) Gasteiger, G.; Jochum, C. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 43. (h) Roos-Kozel, B. L.; Jorgensen, W. L. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 101. (i) Hendrickson, J. B.; Grier, D. L.; Toczko, A. G. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 195.

(2) Corey, E. J.; Wipke, W. T. *Science (Washington, D.C.)* **1969**, *166*, 178.

(3) Corey, E. J.; Petersson, G. A. *J. Am. Chem. Soc.* **1972**, *94*, 460.

(4) Wipke, W. T.; Dyott, T. M. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 140.

(5) Fugamann, R.; Dolling, U.; Nickelsen, H. *Angew. Chem., Int. Ed. Engl.* **1967**, *6*, 723.

(6) The present paper tentatively introduced the terms *abnormal* and *abnormality* in order to describe a ring containing an atom other than C, N, O, S, and P. This convention intends to simulate a chemist's supposition that lies in his/her processes of ring perception.

(7) Multi-tied rings are not adopted here because of saving various arrays that are concerned with rings detected. However, another algorithm that considers both tied and multi-tied rings for covering conditions would be adopted for some purposes.

(8) Programs for detecting all rings have been reported. See: (a) ref 4. (b) Kudo, Y. In *Organic Synthesis and Computers* (R&D Report No. 39); CMC: Tokyo, 1983; p 49. (c) Balaban, A. T.; Filip, P.; Balaban, T.-S. *J. Comput. Chem.* **1985**, *6*, 316. We have used Kudo's algorithm after rewriting with FORTRAN 77.

(9) The comparison of the present program with various programs reported, especially with respect to their execution times, will be published later.

(10) (a) Ryan, A. W.; Stobaugh, R. E. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 22. (b) Freeland, R. G.; Funk, S. A.; O'Korn, L. J.; Wilson, G. A. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 94.

(11) By some extension, the present algorithm can be applied to the perception of ring-opening, ring-closure, and rearrangement reactions. See: Fujita, S. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 1. See also: Fujita, S. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 205; **1987**, *27*, 115.

Knowledge Representation Using an Augmented Planning Network: Application to an Expert System for Planning HPLC Separations

A. L. ANANDA* and S. M. FOO

Department of Information Systems and Computer Science, National University of Singapore, 10 Kent Ridge Crescent, Singapore 0511

HARI GUNASINGHAM

Department of Chemistry, National University of Singapore, 10 Kent Ridge Crescent, Singapore 0511

Received March 11, 1987

The augmented planning network (APN), which is modeled on the augmented transition network (ATN), is proposed as a convenient structure for representing the knowledge of an expert system for planning HPLC separations. The advantage of the APN is realized in its inherent flexibility in handling preconditions, meta-knowledge (or control knowledge), and procedural inferencing. The internal structure of the APN is also presented.

INTRODUCTION

The planning and optimization of separations in high-performance liquid chromatography (HPLC) have been the subjects of considerable interest. A number of quantitative approaches, including statistical techniques, have been employed, but these have been largely unsatisfactory. Although the theory of HPLC as it stands at present is well developed, when one is dealing with real separations, its application requires insight and experience. Also, experimental variability can pose limitations.

In the past two years, we have been concerned with developing an expert system approach to the problem.[1,2] The aim of the HPLC expert system is to guide the chemist in the selection of sample preparation technique, column (stationary phase), eluent (mobile phase), and detection technique by making use of heuristic knowledge derived from experts in the area of HPLC or from the literature. The motivation is that heuristic knowledge is based on experience that cuts across theoretical limitations.

The key element of the HPLC expert system then is the knowledge base, and, in our conception, the primitive elements of the knowledge base are rules. In the design of the expert system, we have adopted a hierarchical strategy where the selection process is divided into modular subtasks as shown in Figure 1. An important benefit of problem reduction in this way is that only those sections of the knowledge base relevant to the particular subtask being executed need be searched for a solution to the subtask. However, because the sections can be by themselves quite large, it is important that an efficient representational structure is chosen.

In the earlier design, the hierarchical structure is reduced to an AND/OR decision tree.[1] Each node of the tree is a rule that can be invoked if the conditional part of the rule is satisfied and the extraction of a solution was derived by a heuristic search of this tree. Although the simple decision tree is a convenient structure in that it is easily implemented, it has a number of limitations. The most significant is the tendency for the tree to proliferate as the number of possible solutions increases.

In this paper we seek to show that the augmented transition network (ATN) formalism is a more efficient structure for representing the knowledge base of the HPLC expert system. In the search for a solution, the exploration of alternatives is natural to the ATN's network form.

The ATN model has been widely used to represent grammars for question-answering systems such as the LUNAR program.[3] It has also been used in speech-understanding systems,[4] in the design of a separable transition-diagram compiler,[5] for modeling learning, and even for processing visual information.[6] ATTENDING[7] uses the ATN model as the basis of its critiquing analysis of anesthetia plans for patients.

A KNOWLEDGE REPRESENTATION MODEL FOR THE HPLC EXPERT SYSTEM

For this application, networks are called *augmented planning networks* (APNs). Each APN is composed of *states* (represented by circles) connected by *arcs*. Each network has an *initial state* from which a *path* can be traced through the

0095-2338/88/1628-0082$01.50/0