A Ring-Based Chemical Structural Query System: Use of a Novel Ring-Complexity Heuristic

RAMASWAMY NILAKANTAN,* NORMAN BAUMAN, KEVIN S. HARAKI, and R. VENKATARAGHAVAN

Medical Research Division, Lederle Laboratories, Pearl River, New York 10965

Received October 6, 1989

A system for extracting, organizing, storing, and retrieving ring systems contained in large structural databases has been developed. For each compound a unique set of ring systems is extracted by using a hash-coding scheme to eliminate duplicates. The set of all the different ring systems in the database is then classified by using a simple heuristic measure of ring complexity. The ring-system data are stored in a database that allows retrieval of compounds based on the size, number, or complexity of the rings they contain. The ring-system connection tables are stored in a separate database that can be used to do standard substructure searches as well as topological similarity searches on the ring systems. Such searches can be a powerful adjunct to conventional database searches.

INTRODUCTION

Pharmaceutical researchers often have to search for compounds "related" to known drugs. Such searches are usually conducted over large databases of diverse compounds, as are available publicly or as maintained by drug companies as proprietary resources.

Such databases can be searched in many different ways depending on one's definition of "related"; for example, one may search for a substructural fragment or look for compounds similar in a topological sense to a given fragment or compound.[1,2] If three-dimensional coordinates are available, one could also carry out pharmacophore searches[3-5] or molecular shape searches.[6,7] Here we present techniques for probing large databases in a ring-system-oriented manner. Several ring systems like the steroid nucleus or the benzodiazepine ring system are associated with important pharmacological activities. Furthermore, it is easy to convince oneself that much of the structural variety in chemistry arises from the different ring systems and combinations thereof. Indeed, in most organic chemical databases, no more than 20–25% of the compounds are acyclic.

METHODS

The method consists of four major steps: (a) extraction of ring systems; (b) elimination of duplicates using a hash-coding scheme; (c) classification of ring systems using a simple complexity index; (d) storage of ring data in an easily searchable manner.

(a) Extraction of Ring Systems. Each molecule is treated as a graph where the nodes represent non-hydrogen atoms and the edges represent the bonds. A spanning tree, constructed from this graph, includes all the edges in the original graph except those that close the cycles in the graph. These remaining edges constitute the ring-closure list. Using this representation of the graph, it is possible to identify all the cycles in the graph. Then all the acyclic edges are dropped, and a simple algorithm is used to identify the connected components. Each connected component is a ring system and is assigned an arbitrary ring number when it is first encountered. For an explanation of the graph-theoretical terms and a more detailed discussion of the method, please see Appendix I.

(b) Elimination of Duplicates. As the same ring system may occur more than once in the same molecule or in different molecules, it is necessary to identify each ring system uniquely, so that we may eliminate duplicates. We have used a novel hash-coding scheme to do this. As each ring is discovered, two hash codes are calculated for it from its connection table. These hash codes are calculated by using two different molecular descriptors known as atom-pair and topological torsion, which we use in structure activity correlation studies.[3,4] The hash-coding method is described fully in Appendix II.

The pair of calculated hash codes is compared with those of other rings found so far. If the hash-code pair has not been seen earlier, we assume that the ring system is new and assign a new ring number to it. If the hash-code pair has already been seen, the corresponding old ring number is assigned to the ring system. Our experiments with these hash codes have shown that, taken as a pair, they are sufficiently discriminatory so that we do not expect any chance occurrence of the same hash codes for two different ring systems.

(c) Classification of Ring Systems. Having obtained a set of rings from the database, we are left with the problem of classifying them in some reasonable way. The reason classification is needed can be appreciated when one considers the fact that in a large database of a few hundred thousand compounds, there may be several thousand different ring systems. Unless these are put into some order, it would be almost impossible to locate any ring system or even to browse through them systematically.

We have classified the ring systems by calculating a simple heuristic measure of complexity using

$$c = nb^2 - na^2 + na \qquad (1)$$

where c is the complexity measure and na and nb are the number of atoms and the number of bonds, respectively. It can be seen that this complexity measure is independent of atom and bond types and the connectivity.

The rationale behind the above equation is as follows. Since the number of cycles in a graph is $E - N + 1$ (where E is the number of edges and N is the number of nodes), the simplest ring complexity heuristic is of course just $E - N$. However, this would not be sufficiently discriminatory since all the graphs with the same number of cycles would have the same complexity. We should like the complexity number to increase as the number of atoms in the ring system increases. Accordingly, we might try the difference between the squares of the number of edges and number of nodes. But this is unsatisfactory for simple cycles, for which the complexity number would be 0 regardless of the number of nodes. To get around this problem, we add on the number of nodes and thus arrive at eq 1.

It can be seen that the smallest value that the ring-complexity number can have is 3, and this corresponds to the smallest simple ring, namely, the three-membered ring. There

* To whom correspondence should be addressed.

0095-2338/90/1630-0065$02.50/0

Table I. Comparison between the *Chemical Abstracts* Database and the *Fine Chemicals Directory*[a]

FCD	CA
6-membered	6-membered
5-membered	5-membered
6–6 fused	6–5 fused
6–5 fused	6–6 fused

[a] Most commonly found ring topologies in order of frequency of occurrence.

is really no upper limit to the complexity number, but in databases that we have examined, the highest value varies from about 1000 to 2000.

(d) Storage of Ring Data. We have written a FORTRAN program that reads in the connection table of each compound from a MACCS[8] database, extracts each ring system, and writes to a file the ring number, the two hash codes, the number of atoms, the number of bonds, the complexity, the number of occurrences in that compound, and the identification number of the compound in which the ring was found. Thus, we obtain a file that has, for each compound in the database, the list of ring systems and data associated with the rings. These data are loaded into a System 1032 dataset.[9] Further, whenever a new set of hash codes is obtained, the program writes the connection data of the ring system to another file that is then used to create a MACCS database in which each ring-system type has a unique ring number.

RESULTS AND DISCUSSION

Uses of the Ring-System Databases. The ring data, stored as discussed above, can be used to analyze the structural composition of a database. We give an example of the analysis of the *Fine Chemicals Directory* (FCD).[10] In this database, of a total of 66 754 compounds, 15 673 were acyclic. This represents about 23.5% of the database. The database contained 1877 different ring systems.

Recently, Stobaugh[11] has published a summary of statistics of the *Chemical Abstracts* database. In Table I we show a comparison of some aspects of that database with the *Fine Chemicals Directory*. The most commonly encountered ring topologies are shown in order of their frequency. It can be seen that the most common ring topology in both databases is the simple six-membered ring. Next most common is the simple five-membered ring. Then come the 6–6 fused and the 6–5 fused systems, although in reverse order in the two databases. We also note that *Chemical Abstracts* has only about 11% acyclic compounds.

Using the complexity index, one can classify the ring systems in a database. Figure 1 shows a random selection of ring systems from the *Fine Chemicals Directory* arranged in order of increasing complexity. It can be seen that the classification accords with our intuitive notions of ring complexity.

We can select the most complex ring systems in the database, complexity being measured as described earlier. It may be pointed out that such a question cannot easily be formulated within the framework of conventional structure searching techniques. In our example, it turned out that the highest complexity numbers were in the region of 750–900. Figure 2 shows the compounds containing these ring systems.

The System 1032 database can be used to generate ideas for the design of novel compounds. For example, consider the benzodiazepine ring system, which is a 6–7 fused system. Suppose we want to synthesize novel variants of this ring system; we might look at other ring systems of the same order of complexity. This is done by simply looking for all ring systems of the same complexity as a 6–7 fused system, i.e., complexity equal to 34. In the *Fine Chemicals Directory*, we found a few interesting ring topologies other than the 6–7 fused type that had the same complexity. These are illustrated in Figure 3. While some of these ring types may be irrelevant in the context of benzodiazepine-like anxiolytics, they do aid in the design process by identifying interesting ring topologies. Using this database, one could also identify compounds containing chosen numbers of chosen types of ring systems.

The MACCS database can be used for substructure searches on the ring systems. Such searches become useful when one needs to query by partially defined ring systems. For example, if we wish to locate all ring systems of which a naphthalene is an embedded part, we can do this with a simple substructure search on the ring-systems database. If such a search were done on the compound database, the removal of unwanted compounds such as simple substituted naphthalenes would be more difficult and time-consuming.

Further, we can conduct structural similarity searches on the ring systems. The idea here is to be able to find compounds with ring systems similar to a given ring system. If such a similarity search is done on the compound database, the calculation tends to get distorted by the presence of components other than the ring system of interest. These searches are done by extracting topological descriptors from the ring systems in the database and comparing these to the descriptors contained in any chosen probe molecule. We have accordingly generated descriptor files for the ring systems in the database. The details of the similarity search method have been published elsewhere.[1,2]

Ring systems are useful indicators of novelty. Thus, we can use ring systems as a basis for acquisition of novel compounds to enrich a collection. Our method can be used to identify all the ring systems contained in one database but not in another. This is done by running the ring perception program as an update, when only the new ring systems will be extracted. This can be used to reveal, for example, novel ring systems that are commercially available but not contained in a company's private collection. We have implemented this technique successfully in our laboratory.

Hash Codes. Our hash codes depend on structure only; they are independent of numbering and representation of alternating single and double bonds, yet they do not require canonicalization of the structure. Thus, our identification of unique ring systems via hash code is rapid compared to methods depending on canonicalization, at the cost of possible occasional "collisions" which would misclassify a ring system.

Ring Systems as Structural Features. The idea of using ring systems as structural features is not new. Adamson et al.[12,13] have pointed out the importance of rings and ring systems and have carried out detailed analyses on relatively simple ring types. They have also suggested the use of ring systems as structural keys for effective substructure search. Feldman and Heller[14] have used rings and acyclic fragments as features in a tree-based substructure search system. They have also used hash codes to rapidly retrieve structures. Corey et al.[15] have reported ring analysis to determine synthetic feasibility. Recently, Randić[16] has introduced ring ID numbers that are based on weighted bond paths. The ring ID number, however, is a ring descriptor rather than a hash code. The values of ring ID numbers show certain structure-dependent trends that may be used to classify a set of substituted rings.

In our study, we have used ring hash codes solely as unique identifiers. These are not expected to show any structure-based trends. Although we have dealt exhaustively with all the different ring types, we do not carry out any detailed ring system analysis. We simply store the ring-system connection tables as well as certain gross characteristics of the rings systems like the number of bonds, the number of atoms, and a complexity heuristic. The latter can be used to carry out topological searches on the database, including Boolean op-

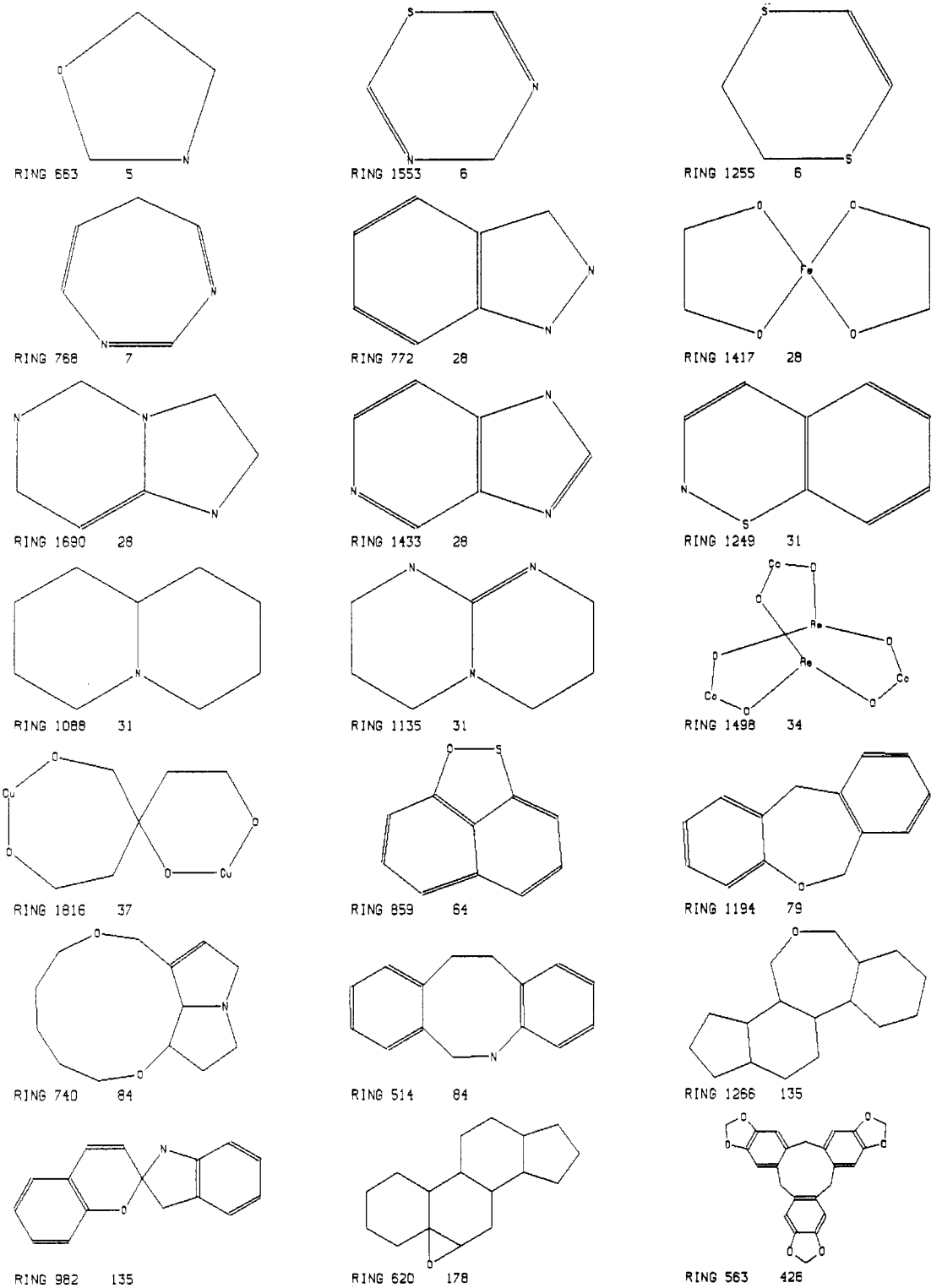

Figure 1. Some ring systems from the *Fine Chemicals Directory* arranged in order of increasing complexity (cf. eq 1). While there is no unique way of classifying ring systems, it can be seen that our complexity measure classifies ring systems in an intuitively acceptable manner.

erations between different queries.

CONCLUSION

Ring-system-oriented search capabilities when used in conjunction with traditional search techniques greatly enhance and enrich the power of the latter.

APPENDIX I

The *from-list* is a list of atoms which is so constructed that each atom is thought to arise from a lower numbered atom to which it is directly bonded. The *from-list* defines a spanning tree for the graph generated by the connection table of the compound. The ring-closure list is a list of pairs of atoms which when connected generate all the cycles in the graph.

Each member of each ring-closure pair is traced back to its parent atom by using the *from-list*. The "exclusive or" of the two bond paths plus the ring-closure bond will give all the bonds involved in the simple cycle to which the ring-closure pair belongs. In this manner all the simple cycles are identified. All the bonds that are not involved in any of these cycles are dropped. Then Floyd's algorithm[17] (a simple shortest paths algorithm) is used to identify the connected components in what remains of the graph. (Any of several other algorithms

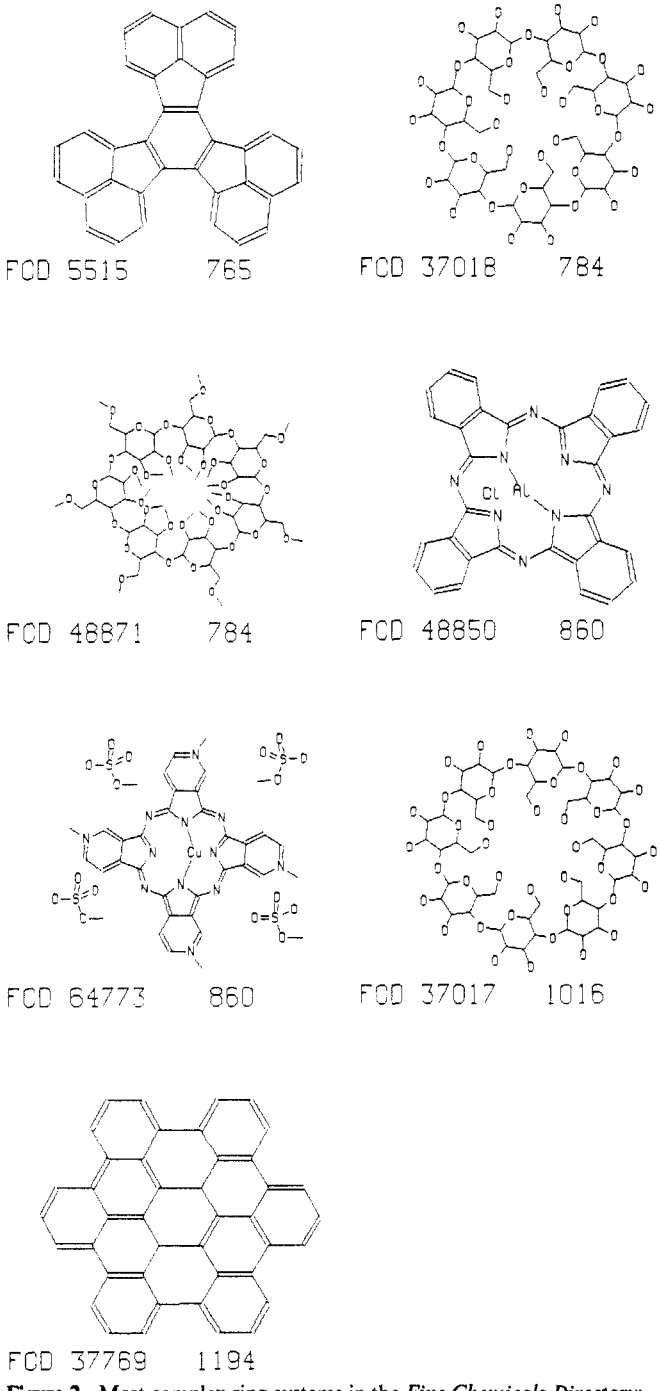

Figure 2. Most complex ring systems in the *Fine Chemicals Directory*; complexity is as defined in eq 1. The figure shows the FCD compounds that contain these rings.

could equally well be used to do this.) Each connected component now corresponds to a ring system.

APPENDIX II

Separate hash codes were constructed from the representation of the compounds as atom pairs and as topological torsions. Each descriptor is packed into 32 bits. The atom-pair hash code is the sum of the products of each atom-pair descriptor right-shifted by 2 bits and the same descriptor right-shifted by 13 bits; overflows are ignored. The torsion hash code is similar, except that the factors are the unshifted descriptor and the descriptor right-shifted by 16 bits. The shifts were chosen to minimize the effects of null bits on the low-order bits of the code; by test, they gave uniform coverage

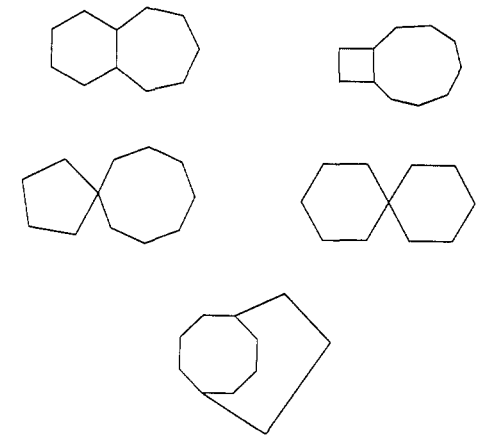

Figure 3. Ring topologies of the same complexity as the benzodiazepine 6–7 fused ring system.

in the sense that the final 32-bit hash codes averaged having 16 bits set. The important point is that the hash code does not require canonicalization (a slow process), yet it depends only upon the structure and not upon the particular numbering of its atoms.

REFERENCES AND NOTES

(1) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure–Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
(2) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82–85.
(3) Rusinko, A., III; Sheridan, R. P.; Nilakantan, R.; Bauman, N.; Venkataraghavan, R. Using CONCORD To Construct Databases of Three-Dimensional Coordinates from Large Connection Table Databases. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 251–255.
(4) Sheridan, R. P.; Nilakantan, R.; Rusinko, A., III; Bauman, N.; Haraki, K. S.; Venkataraghavan, R. 3DSEARCH: A System for Three-Dimensional Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 255–260.
(5) Sheridan, R. P.; Nilakantan, R.; Dixon, J. S.; Venkataraghavan, R. The Ensemble Approach to Distance Geometry: Application to the Nicotinic Pharmacophore. *J. Med. Chem.* **1986**, *29*, 899–906.
(6) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.
(7) Sheridan, R. P.; Venkataraghavan, R. Designing novel nicotinic agonists by searching a database of molecular shapes. *J. Comput.-Aided Drug Des.* **1987**, *1*, 243–256.
(8) MACCS is an acronym for Molecular Access System, a chemical database management system supplied by Molecular Design Limited, San Leandro, CA.
(9) System 1032 is a database management system supplied by Software House.
(10) *Fine Chemicals Directory*, copyright by Fraser Williams (Scientific Systems) Limited and Molecular Design Limited.
(11) Stobaugh, R. E. Chemical Abstracts Service Chemical Registry System. 11. Substance-Related Statistics: Update and Additions. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 180–187.
(12) Adamson, G. W.; Cowell, J.; Lynch, M. F.; Town, W. G.; Yapp, A. M. Analysis of structural characteristics of chemical compounds in a large computer-based file. Part IV. Cyclic fragments. *J. Chem. Soc., Perkin Trans. 1* **1973**, 863–865.
(13) Adamson, G. W.; Creasey, S. E.; Eakins, J. P.; Lynch, M. F. Analysis of structural characteristics of chemical compounds in a large computer-based file. Part V. More detailed cyclic fragments. *J. Chem. Soc., Perkin Trans. 1* **1973**, 2071–2076.
(14) Feldmann, R. J.; Heller, S. R. An Application of Interactive Graphics–the Nested Retrieval of Chemical Structures. *J. Chem. Doc.* **1972**, *12*, 48–54.
(15) Corey, E. J.; Wipke, W. T.; Cramer, R. D., III; Howe, W. J. Techniques for Perception by a Computer of Synthetically Significant Structural Features in Complex Molecules. *J. Am. Chem. Soc.* **1972**, *94*, 431–439.
(16) Randić, M. Ring ID Numbers. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 142–147.
(17) Floyd, R. W. Algorithm 97 Shortest Path. *Commun. ACM* **1962**, *5*, 345.