Graphical Representation for Automated Retrieval of a Class of Fused Six-Rings

SAMUEL D. BEDROSIAN* and MARGARET B. MILNE

The Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pennsylvania 19174

Received October 18, 1976

The graph-theoretic approach presented here provides an effective means of distinguishing a class of fused six-rings as part of substructure search routine of a computer based chemical data file. This is an extension of the "characteristic graph" concept of Balaban denoted simply as a *compressed graph.* Quantitative results are given for a sampling of six-ring systems from the Ring Index to show effectiveness of partitioning in response to a query.

INTRODUCTION

There exists a century-long tie between graph-theoretic techniques and the structural properties of chemical compounds.[1,2] A chemical structure can be viewed as a graph in which the nodes correspond to atoms and the edges to bonds between the atoms. Such applications have been encouraged further by the increased use of computer-based processing of chemical information in the past decade. Of interest here is the emphasis on so-called substructure searches. Considerable attention has been focused on organic polycyclic structures because of their practical importance and the apparent theoretical difficulties in manipulation.[3-5]

Herein *substructure search* refers to retrieval from a data file of chemical structure representations of the subset containing some prescribed substructure of interest to the research chemist. We note also that the graphs representing chemical structures follow the common practice of including only the nonhydrogen atoms of the pertinent compounds. Furthermore, we observe that ring structures occur in more than half of the compounds. A recent study[6] indicates that of all prime rings about 82.5% are six-rings, 14.9% are five-rings, and 2.6% are rings of other sizes. Following the overwhelming predominance of six-rings, one finds as expected that substructure search requests often include six-rings in a variety of fused arrangements. Since files of chemical information with 10^4 to 10^6 structures have become relatively common, it is important to note that a direct mapping approach for substructure searches is, in general, prohibitively expensive.

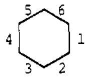
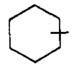
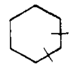
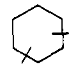
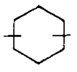
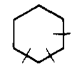
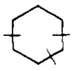
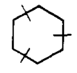
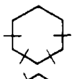
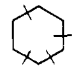
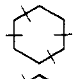
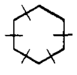
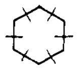
The alternative to direct mapping involves analysis of the compounds in the file to characterize them prior to search with likely *keys* or *screens.* These keys or screens include characteristics such as number and size of rings, presence of commonly encountered structural fragments, as well as the number and species of atoms. By a number of different file organization techniques, the subset of compounds in the file exhibiting at least the same set of screens as the substructure can be rapidly retrieved.[7-10] Thus one is usually left with a very small subset of the data file that must be searched by direct mapping to find any that satisfy exactly the search request.

We present here a method for representing six-graphs using an extension of the "characteristic graph" concept of Balaban.[11] While not unique, the method does effectively partition the set of six-graphs, especially the subset encountered in chemical structures. The method for using this representation to search for arbitrary six-graphs either isolated or embedded in a large six-graph is only outlined here since it will be presented in detail in a later paper.

DEFINITION OF THE CLASS OF STRUCTURES

A class of graphs can be defined for the set of all fused six-membered rings such that no two six-rings share more than one such fused edge in common. The allowable types of fusion in six-ring graphs are shown in Figure 1. Even with the restrictions indicated, the number of possible configurations with any number of six-rings is quite large. We note in passing that the type-a configuration in Figure 1 was treated within the class of linear iterative cellular arrays in ref 12. Also it is important that the class of graphs treated here is not limited to the configurations of types a and b, i.e., the tree-like polyhexes of Harary and Read.[13]

Table I. Distinct Fusion Patterns in Six-Rings

Pattern ('/' denotes fused edge)	Edges Fused	Symbolic Representation
	∅	N No fusion
	1	U Unit fusion
	1,2	O Ortho fusion
	1,3	M Meta fusion
	1,4	P Para fusion
	1,2,3	MO_2
	1,2,4	MOP
	1,3,5	M_3
	1,2,3,4	M_2O_3P
	1,2,3,5	M_3O_2P
	1,2,4,5	$M_2O_2P_2$
	1,2,3,4,5	$M_4O_4P_2$
	1,2,3,4,5,6	$M_6O_6P_3$

The class of graphs formed from the fused six-ring structures as described above is referred to simply as *6-graphs.* This paper describes a method for symbolic representation of 6-graphs that enables rapid search and retrieval of any specific 6-graph whether found isolated or embedded within a larger 6-graph when querying a suitably stored file of chemical information.

The resulting representation is not unique; that is, two different 6-graphs may receive the same representation. Likewise the retrieval methods based on this representation

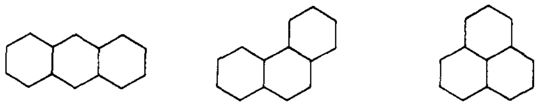
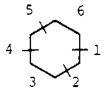


Figure 1. Class of fused six-ring graphs; examples of subclasses.

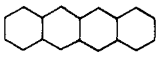
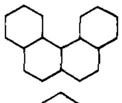
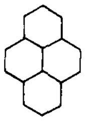
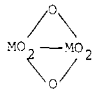
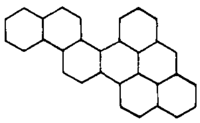
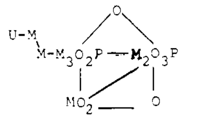
Table II

Edges	Relationship
1,2	O
1,4	P
1,5	M
2,4	M
2,5	P
4,5	O



Fusion Pattern

Table III. Graph Representations of Fused Six-Rings

6-Graph	Compressed Graph
	U-P-P-U
	U-M-M-U
	
	

may retrieve some graphs that do not satisfy the requirement of the query. We have a system that operates in a fail-safe mode in that no correct graphs will be missed. The number of such "false drops" is small and can be readily weeded out by visual or automated comparison of the retrieved graphs with the requested one. Thus the retrieval methods utilize this "screen" for rapidly eliminating the majority of the file from consideration as possible responses.

CONCISE CODING OF FUSION PATTERNS

Any of the six edges of the six-ring can be fused to other rings. Using an arbitrarily labeled six-ring and taking symmetries into account, we find that the 2^6 possible fusion patterns yield the 12 distinct fused patterns as shown in Table I. In addition to seeking distinct representations for each of these, we also desire the flexibility of retrieval by subsets of the fused rings. Thus the representation for a ring fused on edges 1, 2, and 3 should be capable of retrieval by screens seeking the 1,2 or the 1,3 fused configuration as well.

To accomplish this, we adopt the following basic single symbol designations (see first five entries in Table I): N, no fusion; U, exactly one fusion; and O (ortho), M (meta), or P (para) for the corresponding two fused edges. A symbolic representation of six-rings with more than two fused edges is developed as follows.

Consider all pairwise combinations of the fused edges and denote their relative positions by O, M, or P. Then sum the number of occurrences of each pair and cite them in alphabetical order to obtain a symbolic coding of the fusion pattern. We illustrate this for the fusion pattern in Table II at the right where the edges 1, 2, 4, and 5 are fused. Thus the pairwise relationships of the edges are as shown in the table. The symbolic representation for this configuration can be cited as $M_2O_2P_2$.

The complete set of fusion patterns is given in Table I. Note that each of the fusion patterns is distinct and yields a unique symbolic representation.

A NEW COMPRESSED GRAPH REPRESENTATION

Given the symbolic representation of individual fused rings from Table I, a complete system of fused six-rings, i.e., a 6-graph, can be replaced by a new *compressed graph* wherein each labeled node denotes a (fused) six-ring and each edge denotes a fusion bond in the original structure. In particular, each node is labeled with the symbolic representation corresponding to its fusion pattern. Similarly the edges joining the nodes of the new graph indicate a common fused member in the rings denoted by the nodes. Rather than refer to this new graph as a "6-graph compression", we simply use "compressed graph". Note that this result can be viewed as an extension of the "characteristic graph" concept of Balaban.[11] Table III gives examples of fused six-ring structures represented by 6-graphs and their corresponding compressed graphs.

PARTITIONING CAPABILITY OF THE NEW COMPRESSED GRAPH

To illustrate the partitioning capability of the new compressed graph described above, the set of six-ring systems from the Ring Index[14] having four to eight rings were given their corresponding compressed graph representation. These represent a substantial majority of the ring systems that have been encountered in known chemical structures. Table IV shows the rather effective partitioning of this sample subset of the 6-graphs resulting from compressed graph generation. For example, the second row tells us that for the 28 6-graphs containing 5 rings, 20 unique compressed graphs resulted: of these 20, in 15 cases each 6-graph yielded a unique compressed graph; in 3 cases two 6-graphs gave the same compressed graph; in 1 case three 6-graphs gave the same compressed graph; and in 1 case four 6-graphs gave the same compressed graph.

EXTENSION TO OTHER CLASSES OF RING SYSTEMS

While the technique described above applies only to simple fusions of six-rings, one can envision extension of the method to include other classes of nongauche graphs encountered in chemistry. In particular, consideration should be given to

Table IV. Results of Partitioning Selected Compounds from Ring Index

No. of 6-rings in system	No. of unique 6-graphs	No. of unique compressed graphs	No. of compressed graphs yielded by r redundant 6-graphs						% unique compressed graphs for
			$r = 1$	2	3	4	5	>5	$r \leq 2$
4	7	6	5	1	-	-	-	-	100
5	28	20	15	3	1	1	-	-	90
6	52	34	21	10	2	-	1	-	79
7	65	49	36	11	1	1	-	-	95
8	41	37	33	4	-	-	-	-	100

bridged and spiro systems as well as systems containing other ring sizes.

With respect to *bridged* and *spiro* systems, note that in our compression technique each edge in the resulting compression corresponds exactly to one side shared between the connected rings. In spiro structures, no sides (and only one atom) are shared, while in bridged structures multiple sides are shared. Thus bridged and spiro systems could be handled by defining distinct edge types in the compression corresponding to 0 (→ spiro), 1, 2, . . . sides shared between the connected rings. For *bridged* structures there exists another problem in that more than one set of prime rings can be defined. Computer programs for finding all such sets of rings already exist. Since the number of such sets per graph is low, generation of all of the corresponding compressions is feasible if this is required by the application.

With respect to other ring sizes, at least in principle, a similar coding scheme to that descibed for distinguishing fusion patterns of six-rings could be defined for other ring sizes.

Previously cited results indicate that 14.9% of prime rings are five-rings, so that distinguishing their fusion patterns by different node types may indeed be worthwhile. As only the remaining 2.6% are rings of other sizes, it may often be sufficient simply to record the ring size at that node in the compression, without giving the exact substituent pattern.

SEARCH

From the discussion of "compressed graphs" given above, it should be clear that they can be used for selective retrieval of six-graphs whether isolated or embedded in a larger six-graph. The search method is summarized here and will be detailed in a subsequent paper. The method involves application of the augmented atom (AA) concept that has been used in chemical structure retrieval for some time.[15] For our purposes, we take each atom in turn as a central atom (CA) and generate one augmented atom key for each combination of the CA plus one attachment, CA plus two attachments, and CA plus *n* attachments, where $n \leq 6$ is empirically determined depending on file characteristics and search needs. An inverted list is created for each AA key citing the compressed graphs containing the key. The nodes of the keys are represented as bit strings where, e.g., bit 1 = N, bit 2 = U, bit 3 = O_1, bit 4 = O_2, ... A requested isolated compressed graph is retrieved by generating and intersecting the lists for AA keys. Embedded compressed graphs are similarly retrieved except that the bit string, representing the nodes of its generated AA keys, must now be logically included in the bit strings of the corresponding nodes in the file AA key. This is in contrast to retrieval of isolated graphs where the nodes must match exactly.

CONCLUSIONS AND RECOMMENDATIONS

These results suggest a practical working screen for a retrieval system. This conclusion is justified by the observation that this graph representation problem is related to a particularly difficult (if not intractable) combinatorial question: that of determining exactly how many such distinct (six) graphs there are for a given six-ring structure.[16] Further extensions including mixed ring systems particularly systems containing five-rings should be investigated since the ring index (with supplements) covers more than 15 000 entries.

LITERATURE CITED

(1) A. Cayley, "On the Mathematical Theory of Isomers", *Phil. Mag.*, **47**, 444–446 (1874).

(2) A. T. Balaban, "Chemical Applications of Graph Theory", Academic Press, New York, N.Y., 1976.

(3) M. Plotkin, "Mathematical Basis of Ring-Finding Algorithms in CIDS", *J. Chem. Doc.*, **11**, 60 (1971).

(4) A. Zamora, "An Algorithm for Finding the Smallest Set of Smallest Rings", *J. Chem. Inf. Comput. Sci.*, **16**, 40 (1976).

(5) D. Lefkovitz and C. T. Van Meter, "An Experimental Real Time Chemical Information System", *J. Chem. Doc.*, **6**, 173–183 (1966).

(6) M. Milne, "Design of a User Interface to a Chemical Substructure Search System", M.S.E. Thesis, Moore School of Electrical Engineering, University of Pennsylvania, Aug 1976.

(7) M. Milne, D. Lefkovitz, H. Hill, and R. Powers, "Search of CA Registry (1.25 million compounds) with the Topological Screen System", *J. Chem. Doc.*, **12**, 183–189 (1972).

(8) R. J. Feldman in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldman, and E. Hyde, Ed., Wiley-Interscience, New York, N.Y., 1974, Chapter 3, p 55.

(9) E. Meyer, ref 8 Chapter 5, p 105.

(10) A. Feldman and L. Hodes, "The Efficient Design for Chemical Structure Searching. I. The Screens", *J. Chem. Inf. Comput. Sci.*, **15**, 147–152 (1975).

(11) A. T. Balaban and F. Harary, "Chemical Graphs. V. Enumeration and Proposed Nomenclature of Benzenoid Cata-Condensed Polycyclic Aromatic Hydrocarbons", *Tetrahedron*, **24**, 2505–2516 (1968).

(12) S. D. Bedrosian, "Properties and Applications of a Class of Polynomials", *J. Franklin Inst.*, **296**, 469–474 (1973).

(13) F. Harary and R. C. Read, "The Enumeration of Tree-Like Polyhexes", *Proc. Edinburgh Math. Soc.*, **17**, 1–13 (1970).

(14) A. M. Patterson, L. T. Capell, and D. F. Walker, "The Ring Index", 2nd ed, American Chemical Society, Washington, D.C., 1960.

(15) M. F. Lynch in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldman, and E. Hyde, Ed., Wiley-Interscience, New York, N.Y., 1974, Chapter 2, p 31.

(16) F. Harary, E. M. Palmer, and R. C. Read, "On the Cell-Growth Problem for Arbitrary Polygons", *Discrete Math.*, **11**, 371–389 (1975).