Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-based File. Part IV.[1] Cyclic Fragments

By **George W. Adamson, Jeanne Cowell, Michael F. Lynch,* William G. Town,** and **A. Margaret Yapp,** Postgraduate School of Librarianship and Information Science, University of Sheffield, Sheffield S10 2TN

The frequencies of monocycles and of primary rings in 1 : 1- and 1 : 2-fused polycycles have been analysed by means of a simple and rapid computer procedure. The analysis deals with the great majority of ring systems in a sample of the Chemical Abstracts Registry System. The data are presented in terms of ring sizes and composition. In each case, the preponderance of six-membered carbocyclic rings is evident.

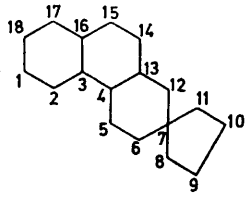
ALTHOUGH the provision of adequate topological screens for rings and ring systems is an essential component of a substructure search system, it is remarkable that little systematic analysis of cyclic characteristics to facilitate the design of appropriate screens has yet been reported. Instead, much of the work on computer manipulation of ring information has been directed toward devising algorithms for ring system analysis,[2-8] or toward formalisms for describing ring interrelations.[9] There are undoubtedly many complex ring systems which present a variety of problems, in terms both of analysis and of representation. The problem is greater within the pages of the *Ring Index* in which each system, once identified, is included, than it is with actual files of structures, where the vast majority of ring systems consists of monocycles and 1 : 2-fused types, with a small proportion of spiro-fused rings. That ring screens are required is evident from the prevalence of rings in chemical compounds; as Leiter and Leighner [10] reported, the average number of rings per structure in the Chemical Abstracts Registry System is *ca.* 2·25, with 86% of the structures containing at least one ring. In the light of these figures it is not surprising that ring bonds predominate in the file, comprising 55% of all bonds between non-hydrogen atoms.

The work reported here reflects a concern with the question of identifying and representing the predominant proportion of relatively simple ring types, rather than the small minority of complex cases. The

[1] Part III, G. W. Adamson, D. R. Lambourne, and M. F. Lynch, *J.C.S. Perkin I*, 1972, 2428.
[2] J. T. Welch, *J. Assn. Comp. Mach.*, 1966, **13**, 205.
[3] C. C. Gotlieb and D. G. Corneil, *Comm. Assn. Comp. Mach.*, 1967, **10**, 780.
[4] K. Paton, *Comm. Assn. Comp. Mach.*, 1969, **12**, 514.
[5] N. E. Gibbs, *J. Assn. Comp. Mach.*, 1969, **16**, 564.
[6] P. L. Long, R. F. Phares, J. E. Rush, and L. J. White, ' Fast Access to Rings in Chemical Structures,' Abstract CHLT 15, 160th Meeting Amer. Chem. Soc., 1970.
[7] E. J. Corey and W. T. Wipke, *Science*, 1969, **166**, 178.
[8] E. J. Corey and G. A. Petersson, *J. Amer. Chem. Soc.*, 1972, **94**, 460.
[9] R. Fugmann, U. Dölling, and H. Nickelsen, *Angew. Chem. Internat. Edn.*, 1967, **6**, 723.
[10] D. P. Leiter and L. H. Leighner, ' A Statistical Analysis of the Structure Registry at Chemical Abstracts Service,' Amer. Chem. Soc. Meeting, Chicago, 1967.

work is an aspect of continuing research on a methodology for the design and implementation of screening systems for substructure search, the general strategy for which has already been described, as well as data from earlier analyses.[1,11,12] A sample file of 28,963 structures, a random sample from the Registry System, is employed.

A procedure to isolate and identify the smallest set of primary rings of monocycles and 1 : 1- and 1 : 2-fused systems was developed. While not general, in the sense that it excludes other fusion types (including *peri*-fusions, bridged rings and rings of rings), it accounts rapidly and accurately for a very high proportion of the structures. Individual ring systems in the connection table are isolated by considering only cyclic bonds. The number of connections to other atoms in the ring system is then determined for each cyclic atom. Systems containing only atoms with connectivities equal to two are thus identified as monocycles. In systems containing atoms with connectivities greater than two, paths are traced from fusion atom to fusion atom; those with a path length of two are considered as fusion pairs. By combining fusion pairs and fusion paths it is possible to identify the primary rings, spiro-rings being characterised by the fusion path beginning and ending with a single fusion atom. In successive stages, as illustrated in the Figure, rings fused to one other ring only are reconstituted; as each ring is formed, the corresponding fusion pairs and fusion paths are removed from the lists. Thus, the formation of the ring consisting of atoms 3-16-17-18-1-2 results in removal of the pair 3-16 and the path 3-2-1-18-17-16; it also results in reassignment of the pair 3-4 as a path of length two, which is ultimately recombined with paths 4-13 and 13-14-15-16 to give the ring 4-13-14-15-16-3. Ring systems with other types of fusions are rejected.



Fusion pairs	Fusion paths	Stages of ring identification
3-16	3-2-1-18-17-16	3-16-17-18-1-2
	4-5-6-7	7-8-9-10-11
3-4	7-8-9-10-11-7	4-13-14-15-16-3
4-13	7-12-13	4-13-12-7-6-5
	13-14-15-16	

Fusion pair and fusion path analysis of ring systems

In the sample of 28,963 structures, 85·5% (24,779) contained at least one ring system. The total number of ring systems was 60,162, an average of 2·08 ring systems per compound, or of 2·43 ring systems for each ring-containing compound. The proportions of monocycles (49·8%) and polycycles (50·2%) were similar. Among the polycycles identified, only 1996, or 4·75%, contained fusion types which were not analysed by the simple procedure.

Again, the number of six-membered monocycles was found to be 25,139, or 84% of all monocycles, while the number of six-membered rings containing only carbon (20,862) was 70% of all monocycles, but 83% of six-membered monocycles.

Simultaneously with identification of ring size, the ring formula was also determined. For monocycles, the data are given in Table 1, for polycycles, the primary rings in ring systems successfully analysed by the program are presented in Table 2. The processing time for operation of the procedure for the full file of 28,963 compounds was 1080 c.p.u. seconds on an ICL 1907 computer. The programming language used was PLAN.

The pattern of ring sizes and compositions evidenced in the Tables illustrates the great preponderance of the six-membered carbocyclic ring, which is even greater than might have been surmised. In general, the patterns shown in each case are essentially similar, the exception being the difference between the incidence of five-membered rings in fused and unfused systems. Thus the five-membered fused carbocyclic ring is nine times more frequent than the corresponding unfused ring.

TABLE 1

Frequencies of ring size and compositions in monocyclic ring systems

Ring composition	Ring size 3	4	5	6	7	>7
C	181	108	274	20,862	80	74
1N	108	32	688	1599	46	12
>1N	1	7	780	1442	12	20
1O	79	19	886	596	5	7
>1O	1	1	135	89	5	7
X	5	12	318	51	2	11
N + O	1	4	299	373	0	6
N + X	0	3	464	74	4	24
O + X	0	7	42	49	5	25
NOX	0	0	3	4	0	8

TABLE 2

Frequencies of primary rings in terms of size and composition, in polycycles containing 1 : 1- and 1 : 2-fusions

Ring composition	Ring size 3	4	5	6	7	>7
C	112	77	2480	17,631	207	54
1N	14	109	1154	1739	76	22
>1N	2	3	969	1143	72	26
1O	138	7	557	680	21	8
>1O	0	1	410	125	7	11
X	37	27	308	268	43	15
N + O	1	0	230	131	12	10
N + X	2	3	532	331	26	5
O + X	10	7	156	140	6	10
NOX	1	1	36	6	2	1

[11] J. E. Crowe, M. F. Lynch, and W. G. Town, *J. Chem. Soc.* (*C*), 1970, 990.

[12] G. W. Adamson, M. F. Lynch, and W. G. Town, *J. Chem. Soc.* (*C*), 1971, 3702.

This difference, not surprisingly, is smaller in the corresponding heterocycles.

The data presented underline the necessity for the analysis and representation at a more detailed level of six-membered rings, especially carbocycles; the results of a deeper analysis will be described in a future paper. Because of the preponderance of monocycles and simple fused systems, an efficient strategy for the generation of ring screens is to design the programs so that they deal with these simple ring systems rapidly and easily.

We thank the Office for Scientific and Technical Information for financial support for this work, and Chemical Abstracts Service for providing the file of chemical structures in machine-readable form.

[2/2617 *Received*, 20*th November*, 1972]