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

By **George W. Adamson, Susan E. Creasey, John P. Eakins,** and **Michael F. Lynch,*** Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield S10 2TN

The immediate structural environments of six-membered rings in a sample of chemical compounds taken from the Chemical Abstracts Registry System were studied. Two levels of description were used, the first indicating bonds attached immediately to six-membered rings, the second also including connected atoms. Disparate distributions similar to those identified for rings of different sizes and compositions were noted. The progression to more detailed levels of description did not lead to even resolution of the most frequent ring types.

We have described previously[1] the results of a ring analysis of a random sample file of structures taken from the Chemical Abstracts Service (CAS) Registry System. The analysis dealt only with the smallest of primary rings from monocycles and 1 : 1- and 1 : 2-fused systems, which together represent the great majority of ring systems in the file. The results were presented in terms of ring size and ring formula, and illustrated the great preponderance of the six-membered carbocyclic ring in the file. On the basis of these findings, a simple ring screen for size and formula was included in the screen set used in an experimental substructure search system.[2] The case for including ring screens is supported by the fact that of the sample of 28,963 structures, 85·5% contained at least one ring system, the average being 2·08 ring systems per compound, or 2·43 for each ring-containing compound. Also, in an evaluation carried out on the Sheffield substructure search system, ring systems formed some part of over half the queries tested.[3]

These simple ring screens are not likely to prove efficient for more detailed queries which focus on substituted six-membered carbocyclic rings, and this underlines the necessity for a deeper analysis and

[1] Part IV, G. W. Adamson, J. Cowell, M. F. Lynch, W. G. Town, and A. M. Yapp, *J.C.S. Perkin I*, 1973, 863.

[2] G. W. Adamson, J. Cowell, M. F. Lynch, A. H. W. McLure, W. G. Town, and A. M. Yapp, *J. Chem. Docum.*, 1973, in the press.

[3] G. W. Adamson, J. A. Bush, M. F. Lynch, and A. H. W. McLure, in preparation.

representation of the immediate environments of six-membered rings. Detailed descriptions of rings do not seem to have found much favour as screens; even in the screen generation system of Lefkovitz, no details other than ring size and molecular formula are included.[4] This apparent lack of interest may stem from the difficulties involved in formulating an efficient program for generating ring screens from structures in a file and from queries. In this paper, approaches to providing this more detailed information are described.

Extension of Simple Ring Analysis Program.—The first approach involved the use of an extended version of the simple ring analysis program.[1] It identifies any six-membered rings present, generates a description of each ring with any external bonds, and with or without its immediate substituent atoms, converts the description into canonical form, and makes a count of the frequency of each ring fragment type in the file.

The two levels of description which this program can generate will be called the complete ring and the bonded ring. The former includes the immediate substituent atoms, the latter does not. These are only two of many levels of description possible for ring systems, some of which are shown in Figure 1. Others are arrived at by

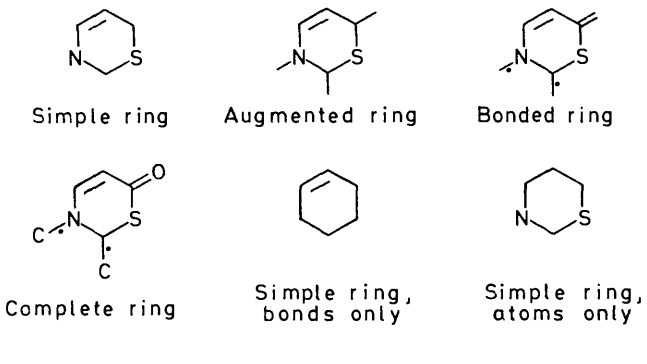


FIGURE 1 Some levels of ring description

considering atoms only, or bonds only. Only one other type, the augmented ring, will be considered here.

The extended ring analysis program consists of four segments, the first of which converts a compact list connection table into a redundant connection table. The second segment is a modification of the ring trace program described previously.[1] It traces paths through all ring systems encountered and identifies all six-membered rings which form part of monocyclic systems, *ortho*-fused systems, or those with a spiro linkage. Up to fifteen six-membered rings can be processed for each compound. In the third segment, a complete topological description of each ring at the appropriate descriptive level (bonded or complete ring) is built up. If all ring atoms and bonds are not identical it is necessary to manipulate the record to obtain a canonical representation. This is done by considering the twelve alternative descriptions of the ring which are arrived at by entering a six-membered ring at any one of six positions and tracing an atom-by-atom path in either of two directions, and selecting the one with the greatest value when regarded as a bit strong. In building up the canonical representation the following hierarchical ordering scheme is used: ring atom–ring bond order; external bond value; external atom type. This involves working outwards in stages from the ring in building up its description. An example of a ring record is given in Figure 2.

S 05C 05N 05C 06C 05C 05	ring atom–ring bond order, words 1—6
000005000500000000000200	external bond value, words 7—12 (for codes, see ref. 7)
C C O	external atom type, words 13—18

FIGURE 2 The canonical record for the ring shown in Figure 1, at the complete ring level

The fourth segment of the program takes the canonical description of each ring and compresses it to a twelve-word (48-character) record. Each record is then compared against those already found for the file. If the ring type has already been detected in a compound, only the frequency count for that type is incremented; if, however, this is the first occurrence of a type in a particular compound, both frequency and incidence counts for that type are incremented. Records of the eight commonest types are preset at the head of the list, the majority of occurrences thus being matched at an early stage, resulting in time saving. If a ring type fails to match any already in the list, it is added to the end of the list and incidence and frequency counters are allocated. At the end of the file analysis, the records are sorted and output as a ranked list.

The program was run on 1 in 100 and 1 in 30 samples of the CAS random sample file, representing 289 and 965 compounds, respectively. It was immediately apparent from the results that ring fragments, in common with other fragments investigated during studies on the design of screening systems,[5-7] exhibit a Zipfian distribution, a small number of fragment types accounting for the majority of ring occurrences in a file sample, with most other fragment types occurring once or twice only. The variety of ring types increased in proportion to sample size, as is shown in Table 1. This presents

TABLE 1

Number of fragment types occurring in different random sample file samples; the figure in parentheses shows the number occurring only once

	Sample size	
	1% (289 compounds)	3·3% (965 compounds)
Bonded rings	156 (102)	355 (240)
Complete rings	252 (191)	661 (479)

[4] D. Lefkovitz and A. R. Gennaro, *J. Chem. Docum.*, 1970, **10**, 86.

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

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

[7] G. W. Adamson, S. E. Creasey, and M. F. Lynch, in preparation.

practical problems, since for large samples the space and time needed to count the fragments would be substantial.

Table 1 shows that the potential variety of complete ring types in a file is enormous, since so many different substituent atoms may occur; it shows too that the number of bonded ring types may also rise rapidly with file size, although ultimately levelling off at a much lower value than for complete ring types.

The analysis results for the 1% file sample will now be described in more detail. They are listed in Tables 2a and 2b.

Both incidence and frequency counts were found to fall off rapidly with decreasing rank, in the case of both bonded and complete rings, although the ratio of frequency of the first-ranked type to that of the tenth, at about 8 : 1, was comparable with that found for bonded pairs.[5] The first few most commonly occurring rings, for both levels of description, are shown in Figure 3. All but one of the eight highest-ranked bonded rings were based on the benzene nucleus, the most common being the monosubstituted benzene ring, with *o*-substituted benzenes (disregarding bond type) second, and those with *p*-substitution third. The saturated ring in seventh place was seen to be the C-ring of a steroid; other fragments characteristic of steroids also figured prominently in the sample.

The increase in the level of ring description on going to the complete ring reduced the frequency of some high-ranking bonded rings. This was especially true of the 1,2,4-trisubstituted benzenes, probably because of the large variety of possible combinations of three substituent atoms. Other fragments, notably those from steroids and similar structures, were hardly differentiated at all.

The degree to which highly detailed descriptions can differentiate fragments which occur too frequently to be adequately handled by simple descriptions is related to their usefulness as screens. In order to investigate this property, a fresh 1% sample of the file, analysed at the complete ring level, was examined. Records were grouped manually to give frequency counts for four of the most common benzenoid fragments and two of the steroid fragments, at augmented, bonded, and complete ring levels. This confirmed that the resolving power of the more detailed levels of description is highly uneven for reasons inherent in the nature of the fragments, which are described below.

At the augmented ring level the monosubstituted benzene nucleus was most common, occurring 78 times. Since only one bond type is permissible, no differentiation at the bonded ring level occurred, and at the complete ring level only the substituent atom could vary

TABLE 2a

Bonded rings, ranked by incidence and frequency, for 1% file sample (289 compounds)

Incidence-ranked			Frequency-ranked		
Rank of ring type	No of. types at that rank	Incidence of type	Rank of ring type	No. of types at that rank	Frequency of type
1	1	48	1	1	73
2	1	34	2	1	43
3	1	31	3	1	32
4	1	21	4	1	23
5	1	16	5	1	18
6	1	15	6	1	16
7	1	13	7	1	14
8	1	10	8	1	13
9	1	7	9	1	10
10	2	6	10	3	6
12	2	5	13	2	5
14	3	4	15	2	4
17	7	3	17	12	3
24	27	2	29	26	2
51	106	1	55	102	1

TABLE 2b

Complete rings, ranked by incidence and frequency, for 1% file sample (289 compounds)

Incidence-ranked			Frequency-ranked		
Rank of ring type	No. of types at that rank	Incidence of type	Rank of ring type	No. of types at that rank	Frequency of type
1	1	43	1	1	58
2	1	13	2	1	17
3	1	12	3	1	15
4	2	10	4	2	12
6	3	7	6	3	8
9	2	6	9	2	7
11	3	4	11	2	6
14	13	3	13	2	5
27	23	2	15	3	4
50	202	1	18	13	3
			31	30	2
			61	191	1

giving a distribution of ring types as asymmetric as that for atoms in the file,[5] with the type containing a substituent carbon accounting for over 70% of all occurrences of this fragment. For the second-ranked augmented type (*o*-disubstituted benzenes), differentiation at bonded and complete ring level was much better, since both atoms and bonds could vary.

The third- and fourth-ranked augmented fragments (*p*-disubstituted benzenes and 1,2,4-trisubstituted benzenes) were differentiated only slightly or not at all at bonded ring level, since, as bridged rings were not considered, only acyclic single bonds were allowed for *p*-substitution, while the other case gave rise to only three types of bonded ring. Inclusion of bridged rings would permit cyclic single and aromatic bonds for *p*-disubstituted benzenes, and at least seven types of bonded ring would be allowed on 1,2,4-trisubstitution. Progression to the complete ring level gave rise to much differentiation through inclusion of different combinations of substituent atoms.

Steroid fragments were not differentiated at all above the bonded ring level, since there is obviously a limitation on the type of atom involved in such structures. The inclusion of bond values to differentiate augmented types into bonded rings was somewhat uneven. This suggests that detailed description of this rather specialized fragment would be unprofitable.

The optimum level of description seems to be different for each type of fragment, and further work is necessary to determine these levels precisely. However, the augmented ring presents itself as a promising candidate for further investigation, particularly as quite a high proportion of the queries tested against the experimental substructure search system were expressed in this form.[3] This level of description should also prove faster in computer generation since no substituent atoms or bond types need be specified.

A General Algorithm for Ring Analysis.—Having completed the work described above, we investigated the design of an algorithm to produce all rings from all systems likely to be encountered in chemical compounds, since any analysis of ring systems carried out as a preliminary step in screening system design should preferably include all systems in the file. Other algorithms for ring system analysis have been published,[8-15] but these all concentrate on selecting a subset of rings which provides a concise but complete description of a compound's ring system. For example, Corey's algorithm [15] perceives ' synthetically significant ' rings, and the ring set generated consists of any ring which is the smallest ring containing some bond (the fundamental ring for this bond), plus all rings containing six or fewer atoms. The decision as to whether a particular species of ring can be said to exist within a structure can be a problem—for example, does naphthalene contain one ten-membered ring, or two six-membered rings, or both?—and Corey has chosen to collect only rings which are meaningfully related to the chemistry of the system.

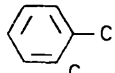
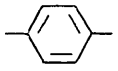
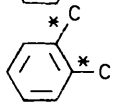
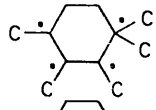
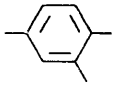
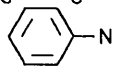
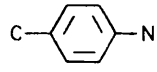
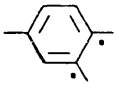
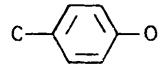
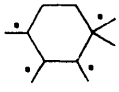
In designing the algorithm to be described, attention was primarily focused upon six-membered rings, in line with our earlier investigation. A rigorous program to analyse all six-membered rings, be they fundamental or envelope rings, in all ring systems, was required. The program developed involves generation of spanning trees,[10] and will now be explained in some detail.

First, all non-cyclic connections are eliminated from the redundant connection table. Then, all ring fusion atoms in the compound are identified. These have connectivities greater than two, in contrast to ordinary ring atoms, whose connectivities are always equal to two. A spanning tree is grown from each fusion atom in turn, being continued for six levels. The whole of the tree is stored in the computer as it is generated. Any path which returns to the start atom before the sixth level is ignored, as it described a ring of fewer than six atoms. At the sixth level, any atom which is the same as the start atom is taken and a path traced back through the tree to give a description of a six-membered ring, which is stored for later treatment. Each ring containing the fusion atom for which the tree is being grown will be described twice in that particular tree, since the ring will be traced in each of two directions. Also, any ring containing more than one fusion atom will be described in the tree of each fusion atom it contains. At the end of the tree-growing process, the unique six-membered ring descriptions are found from the collection of descriptions generated, and are stored. A spanning tree is shown in Figure 4.

After description of all fused rings, the fusion atoms are removed from the connection table and any remaining monocycles in the structure are generated by a simpler version of the above tree-growing procedure. Their unique ring descriptions are stored after the fused ring records. Each description takes the form of six words in the computer, containing the addresses of six atoms in the redundant connection table. These can be in ring order or may be sorted into ascending numerical order for comparison purposes.

The program was tested by input of individual connection tables for compounds containing bridged rings, *peri*-fused rings, monocycles only, and fused rings plus monocycles. All six-membered rings were analysed correctly. The analysis for cubane gave sixteen six-membered rings. Analysis of all **28,963** compounds in the CAS random sample file took **1820** s of c.p.u. time, on an ICL 1907 computer, which is considerably more

[8] J. T. Welch, *J. Assn. Comp. Mach.*, 1966, **13**, 205.
[9] C. C. Gotlieb and D. G. Corneil, *Comm. Assn. Comp. Mach.*, 1967, **10**, 780.
[10] K. Paton, *Comm. Assn. Comp. Mach.*, 1969, **12**, 514.
[11] N. E. Gibbs, *J. Assn. Comp. Mach.*, 1969, **16**, 564.
[12] P. L. Long, R. F. Phares, J. E. Bush, and L. J. White, ' Fast Access to Rings in Chemical Structures,' Abstract CHLT15, 160th Meeting, Amer. Chem. Soc., 1970.
[13] R. Fugmann, U. Dölling, and H. Nickelsen, *Angew. Chem. Internat. Edn.*, 1967, **6**, 723.
[14] E. J. Corey and W. T. Wipke, *Science*, 1969, **166**, 178.
[15] E. J. Corey and G. A. Petersson, *J. Amer. Chem. Soc.*, 1972, **94**, 460.

Bonded rings	Incidence	Frequency	Complete rings	Incidence	Frequency
	48	73		43	58
	34	43		13	17
	31	32		12	12
	21	23		10	15
	16	18		10	12
	15	16		7	8
	13	13			
	10	14			

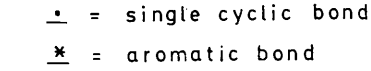
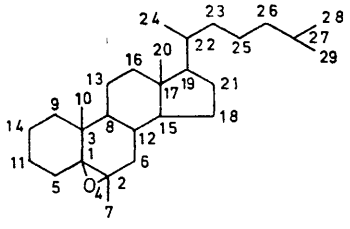


FIGURE 3 Most common bonded and complete rings, for 1% file sample (289 compounds)



Rings from fusion atom 1 (* = end of ring)

3-membered: 1-2-4-1, 1-4-2-1 } ignored by program

6-membered: 1-3-9-14-11-5-1
1-5-11-14-9-3-1
1-2-6-12-8-3-1
1-3-8-12-6-2-1

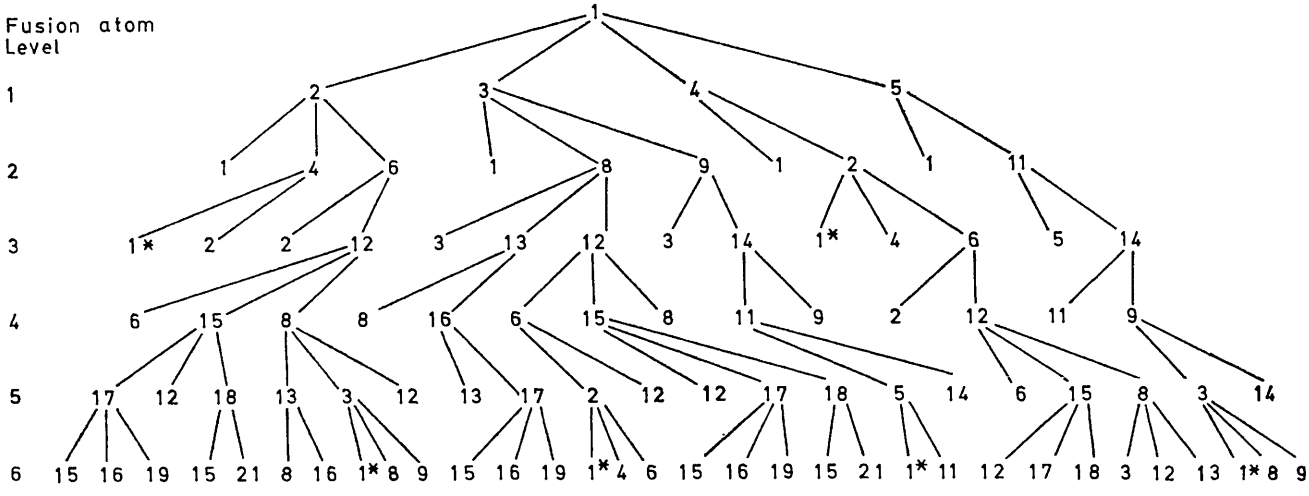


FIGURE 4 Example of a spanning tree

than the 1080 c.p.u. s taken to analyse the file for rings of any size using the earlier ring analysis program.[1]

The program could be adapted to analyse rings of all sizes. For those of fewer than six members, ring information is already in the spanning tree and could be extracted by tracing back to the start atom from the appropriate level as is done for six-membered rings. The growing of levels to find rings with more than six members would, however, increase time and space requirements.

This tree-growing segment was used in place of the ring trace segment [1] for analysis and counting of six-membered rings at the complete ring level. For a 965-compound (1 in 30) sample of the random sample file, the number of rings found using the tree-growing segment was 1745, compared with 1593 found using the ring trace segment. The extra rings were seen to arise from bridged and *peri*-fused systems, with which the ring trace segment could not deal. Several of these extra ring descriptions were not handled correctly by the existing counting segments, and so these were rewritten. The completed program was capable of correct analysis and counting of all six-membered rings at the complete ring level. However, analysis using this improved program was found to take considerably longer than did the use of the tree-growing segment with the older counting segments. The times for analysis of a 2896-compound (1 in 10) sample of the random sample file were 785 and 459 s of c.p.u. time, respectively. This sample contained 5176 complete six-membered rings, an average of 1·79 rings per compound. There were 1545 different complete ring types, of which 1029 occurred once only. The most frequent ring type was the mono-substituted benzene ring with a substituent carbon atom attached, which occurred 527 times.

Conclusions.—As for many other fragments, the distribution of detailed ring fragments in a file follows the familiar Zipfian pattern. The resolving power of different levels of ring description appears to be rather more uneven than that observed in other cases of progressive fragment description, for example the hierarchy of simple, augmented, and bonded pairs.[5] Difficulties may arise in coding some generic queries unless subfragments are used; for example, an *ortho*-substituted benzene ring may be a substructure of 1,2,3- and 1,2,4-trisubstituted benzenes, among others. For these reasons it would be difficult to implement a differential screen system for rings such as that used for pairs,[2] in an operational search system without detailed cost-benefit studies. Nonetheless, certain ring types and levels of description present themselves as likely candidates for further investigation as screens.

We thank OSTI for financial support, and the Chemical Abstracts Service for provision of the sample of structures.

[3/572 *Received*, 19*th March*, 1973]