Analysis of Structural Characteristics of Chemical Compounds in the Common Data Base

GEORGE W. ADAMSON, SUSAN E. CREASEY, and MICHAEL F. LYNCH*
Postgraduate School of Librarianship and Information Science
University of Sheffield, Western Bank, Sheffield, S10 2TN, England

Received March 5, 1973

The distributions of a variety of structural characteristics, including bond-centered, atom-centered, and ring fragments, have been investigated for a specialized file of chemical structures. The results are compared with those obtained for a random sample of the Chemical Abstracts Registry System and are found to reflect the biomedical bias of the file, while conforming to the over-all pattern of the random sample analysis. The implications of these findings for the development of screening systems for substructure search are set out.

As an extension of work on the development of screening systems for substructure searching,[1-4] which has been carried out on a random sample of structures drawn from the Chemical Abstracts Service Registry System, attention has now been directed to the Common Data Base (CDB). This file contains the unique connection tables for some 21,000 substances recorded in the CAS Registry System, the compounds represented being well-known substances drawn from reference works such as the "Merck Index," "Pesticide Index," USAN, and the "Colour Index." Methods developed during work on the random sample file have been applied to this specialized, organically-biased file, which is more representative of actual files likely to be encountered in chemical information work. This paper describes the results of file analysis of structural characteristics for the CDB.

This type of analysis is a necessary preliminary to the design of an efficient screening system, since a knowledge of the distribution of structural features is one factor which guides the choice of appropriate screens for any particular file.[1,4] Analyses for various types of bond-centered, atom-centered, and ring fragments, as detailed in Figure 1, were carried out on the CDB, and the results obtained compared with those obtained for the random sample file using the same fragments.[1,3] The number of structures in each file was: CDB 20,349, random sample file 28,963.

A major obstacle to be overcome, before a satisfactory comparison of results between the CDB and the random sample file could be made, concerned the fact that in the CDB, 12 different bond types were recognized, whereas for the random sample file, there were only eight possible bond codes, one of which was not used. (The random sample file was chosen from the earlier implementation of the CAS Registry System on an IBM 7010 computer). In all analyses carried out on the CDB, the eight bond codes of the random sample file were substituted for the original 12, as shown in Figure 2. This was possible because, on counting the occurrence of each bond code in the CDB, it was found that, of the five which had no equivalent in the random sample file, four occurred either not at all or only once or twice; these were substituted by the most suitable random sample file codes. The remaining CDB code, 3, for acyclic tautomeric bonds (generally occurring in pairs containing one or no carbon atoms), occurred 47,700 times and was substituted by the unused random sample file code 3. These substitutions meant that programs used for analysis of the random sample file could be applied to the CDB with no further change.

In the programs used for file analysis, which were similar to those described previously,[1,3] the compound structures were first transformed from the multilevel compact lists in which they were stored on magnetic tape into redundant connection tables. Bond order was indicated, and bonds were identified as being cyclic or acyclic as shown by the eight-bond codes used. The frequency and incidence (number of compounds in which a fragment occurs) were determined as follows for each fragment. As each

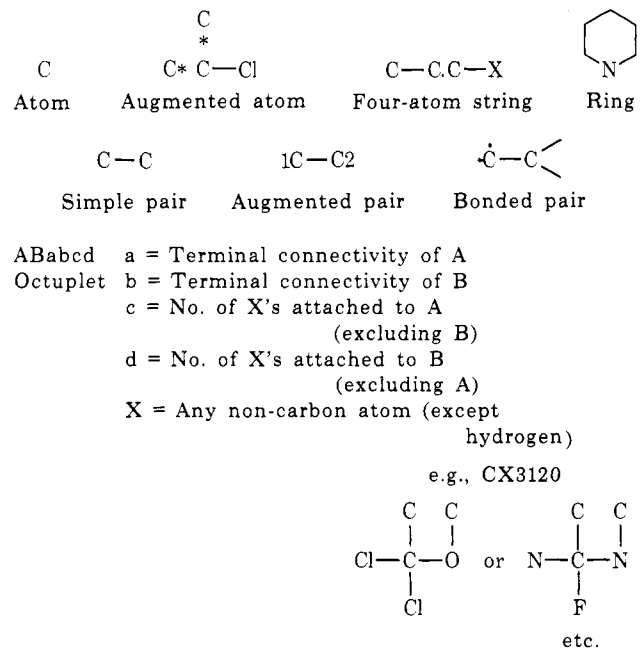

Figure 1. Fragments included in the analyses

*To whom correspondence should be addressed.

Random Sample File Bond Codes	Bond Symbol	Equivalent CDB Bond Codes	Frequency of CDB Bond Codes
1 acyclic single	—	00 dot disconnection	1
		21 acyclic single	298,464
2 acyclic double	=	22 acyclic double	65,232
3 not used	...	23 acyclic tautomer	47,700
		24 acyclic delocalized	0
4 acyclic triple	≡	28 acyclic triple	1,060
5 cyclic single	·	11 cyclic single	148,794
6 cyclic double	:	12 cyclic double	11,742
7 aromatic	*	17 cyclic alternating	324,258
8 cyclic triple	⋮	13 cyclic tautomer	1,070
		14 cyclic delocalized	0
		18 cyclic triple	2

Figure 2. Bond codes for the two files

fragment was produced, it was transformed into a canonical representation by application of the appropriate set of simple ordering rules, which have been described previously.[1,2] It was then compared against the list of fragments already generated for the file; if already present in the list from the same compound, only the frequency count for that fragment was incremented. If the fragment was present in the list, but from a different compound, both frequency and incidence counts were incremented. Fragments not already in the list were added at the end and incidence and frequency counters were allocated to them. This method of counting differs slightly from that previously employed, which had to deal with augmented and bonded pairs containing only carbon, oxygen, and nitrogen, which were counted in matrices.[1] The large number of comparisons necessary with the present technique meant that with fragments where a very large number of types was found—e.g., bonded pairs—the method was too slow for the longest computer run available. In these cases, the more frequently occurring fragments were first determined and preset at the head of the list, thus reducing the number of comparisons necessary. All lists were sorted by incidence count before being output as incidence-ranked lists. A different procedure was used for ring analysis.

RESULTS

Elements. Counts of elements are shown in Table I. As with the random sample file, carbon, oxygen, and nitrogen together accounted for over 90% of the atoms in the CDB, although the incidence of carbon was lower and that of oxygen higher than for the sample file. There were 102 different atoms in the CDB (68 for the sample file). Of these, 22 occurred only once, and as several of these such as rare earths and rare gases did not occur in other analyses as pairs, etc., they were deduced to be present in the file as single atoms. In the random sample file there were eight atoms which occurred once only. Other features of CDB in comparison with the sample file include higher incidences of sulfur, arsenic, gold, and mercury, and lower incidences of fluorine, silicon, and deuterium. These trends may reflect the bias of the CDB towards compounds of medical interest. The fluorine discrepancy is explained by the fact that at the time of the compilation of the random sample file, the CAS registry file contained all fluorine-containing compounds indexed in *Chemical Abstracts* or *Beilstein,* and so would be expected to have a high incidence of fluorine-containing compounds.

Pairs. Analysis in terms of pairs showed the same general features in the CDB as were observed with the random sample file. Most of the fragments consisted of various combinations of carbon, oxygen, and nitrogen. The attenuation of the distributions, measured by the ratio of the frequency of the first-ranking species to that of the tenth, decreased markedly on increasing the size of the fragment, from element via simple pair to augmented pair, since increasing fragment size also increases resolving power of the fragments (Figure 3). The increase on progressing to bonded pairs occurs because the highest-ranking augmented pairs are those that are least effectively split into different species when they are described as bonded pairs.

Table I. Elements: Incidence-Ranked Lists

	CDB			Random Sample File		
Rank	Element	Frequency	% Incidence	Element	Frequency	% Incidence
1	C	300075	93.6	C	456529	99.6
2	O	75636	86.4	O	83395	82.6
3	N	33373	62.6	N	44776	64.0
4	S	8301	25.1	S	8034	19.9
5	Cl	4652	11.7	Cl	7455	14.1
6	P	870	3.5	F	10423	10.0
7	Br	861	2.4	P	1742	4.6
8	F	861	1.7	Br	1614	4.1
9	I	658	1.5	Si	704	1.4
10	As	180	0.81	I	472	1.1
11	Hg	145	0.66	B	389	1.0
12	Si	264	0.63	Sn	158	0.46
13	Au	77	0.38	D	345	0.44
14	B	74		Se	126	
15	Se	66		As	124	
16	Cr	73		Hg	82	
17	Sb	59		Ge	65	
18	Pt	48		Sb	52	
19	Fe	44		T	54	
20	Bi	40		Al	38	

	Random Sample File	CDB
Elements	967	1667
Simple pairs	21.9	6.9
Augmented pairs	5.2	5.3
Bonded pairs	5.7	11.1
Augmented atoms	12.5	5.5

Figure 3. Ratio of frequency of first-ranking element to that of the tenth, for incidence-ranked analyses of various fragment types

The results of analysis of the two files in terms of pairs are shown in Tables II to IV. The numbers of different species for simple, augmented, and bonded pairs in the CDB were 424, 789, and 1451, respectively, the corresponding figures for the random sample file being 362, 841, and 1733. The augmented and bonded pair analyses include all such pairs for both files, and not, as previously,[1] only those containing carbon bonded to carbon, oxygen, or nitrogen.

The larger number of types of simple pairs in the CDB reflects the greater diversity of elements found in this file. This is also shown by the lower figures for augmented and bonded pairs, since pairs consisting of elements other than carbon, oxygen, and nitrogen would not be resolved into different species on going to higher levels of description as much as would be pairs containing these three atoms. This is because the variety of bonds allowed in pairs containing the less common elements is smaller.

Other features of the element distribution in the files were seen in the pair analyses; carbon-carbon pairs of all types had lower incidence in the CDB, as had carbon-halogen pairs, except C-I. The incidence of sulfur-containing pairs, especially those containing oxygen and sulfur, was generally higher in the CDB than in the random sam-

Table II. Simple Pairs, Ranked by Incidence

	CDB		Random Sample File	
Rank	Pair	% Incidence	Pair	% Incidence
1	C—C	75	C—C	85
2	C * C	61	C * C	63
3	C—O	56	C—O	56
4	C—N	52	C=O	53
5	C=O	40	C · C	51
6	C · C	37	C—N	51
7	C · N	22	C · N	26
8	S=O	18	C : C	22
9	C—S	18	C · O	16
10	C⋯O	17	C—Cl	12
11	S⋯O	16	C—S	12
12	C : C	15	C=C	11
13	C · O	11	C : N	10
14	N=N	10	C * N	9.2
15	C=C	10	C—F	9.2
16	C—Cl	9.7	N=O	8.1
17	C * N	7.4	S=O	7.6
18	N=O	6.8	C=N	7.2
19	C : N	4.8	C · S	6.2
20	C · S	4.3	N—N	4.7

Table III. Augmented Pairs, Ranked by Incidence

	CDB		Random Sample File	
Rank	Pair	% Incidence	Pair	% Incidence
1	1C*C2	60	1C*C2	62
2	1C*C1	55	1C*C1	57
3	2C*C2	43	2C=O0	52
4	2C=O0	39	1C—C2	41
5	1C—C2	34	2C*C2	39
6	2C—C2	34	2C—C2	35
7	2C—O0	31	1C·C2	32
8	0C—C2	28	0C—C2	32
9	2C—O1	27	2C·C2	31
10	2C·C2	25	2C—O1	31
11	1C—C1	23	1C·C1	29
12	2C—N1	23	1C—C1	27
13	1C·C2	23	2C—O0	25
14	0C—C1	19	0C—C1	23
15	1C·C1	19	2C—N1	21
16	0O=S3	18	2C—N2	18
17	2C⋯O0	17	1C—O1	17
18	2C—N0	16	0C—C3	15
19	0O⋯S3	15	2C·C3	14
20	2C—S3	14	2C·N2	14

ple file. In the CDB these oxygen-sulfur pairs often contained tautomeric central bonds (bond code 3), which were not found in the sample file. This was also apparent in the case of carbon-oxygen pairs and tended to resolve the pairs, resulting in lower incidence and rank for the pair with the nontautomeric bond in the CDB than in the sample file. The simple carbonyl group C=O had 56% incidence in the random sample file and only 40% in the CDB, but the pair C⋯O also occurred in the CDB, with an incidence of 17%. A similar diversity was seen on progressing to bonded pairs, where some of the external bonds were also described by tautomeric bond code 3, in the CDB—e.g., —C(=O)— (incidence 24%) and —C⋮O (incidence 15%) both conflated in the random sample file to bonded pair —C(=O)— (incidence 41%). The sulfur-containing pairs O⋯S ≤ and O=S≪ in the CDB showed similar behavior.

Table IV. Bonded Pairs, Ranked by Incidence

	CDB		Random Sample File	
Rank	Pair	% Incidence	Pair	% Incidence
1	*C*C*	53	*C*C*	57
2	*C*C*	50	*C*C*	55
3	*C*C*	26	—C=O	41
4	—C=O	24	·C·C·	26
5	—C—C—	19	—C—C—	24
6	*C—O	18	C—C—	22
7	C—C—	17	*C*C*	20
8	·C·C·	16	*C*C*	19
9	—C⋯O	15	·C=O	18
10	·C=O	15	*C*C*	18
11	*C*C*	15	=C—O—	17
12	*C*C*	14	—C—O—	17
13	*C*C*	13	·C·C·	16
14	*C*C*	13	—C—C=	14
15	=C—O—	13	*C—O—	13
16	O⋯S—	12	=C—N—	13
17	O=S—	12	C—O—	13
18	—C—O—	12	·C·C·	12
19	*C—O—	12	—C—C—	12
20	—C*C*	11	=C—C*	11

With regard to carbon-carbon pairs, the two highest-ranked bonded pairs, *C*C* and *C*C*, exchanged ranks in the two files, although the two highest-ranked augmented pairs were the same in each file (see Tables III and IV). The fact that the highest-ranking simple pair in both cases was not aromatic, but the carbon-carbon single bond, C—C, shows the smaller resolution possible with the aromatic pair on investigating the number and type of its external connections. The random sample file tended to have higher incidence and rank for carbon-carbon, singly-bonded pairs (either cyclic or acyclic) than did the CDB. This was also shown generally in simple pairs (incidence of C·C for CDB 37%, for random sample file 51%) and augmented pairs.

Augmented Atoms. The results of these analyses are shown in Table 5. There were 1655 species for the CDB and 2331 for the random sample file, reflecting in part the greater size of the latter file. Of these, 528 occurred once in the CDB and 729 once in the random sample file. The same pattern emerged as was seen in the pair analyses with regard to element distribution and resolving power of the tautomeric bond in oxygen- and sulfur-containing fragments. The aromatic fragment C*C*C had rank 1 in both files and the aromatic species C*C*C where X = N,O or S generally had higher incidences in the CDB than in the random sample file. Where X = C the incidence of

Table V. Augmented Atoms, Ranked by Incidence

In singly connected augmented atoms, the atom which has one connection is given first

	CDB		Random Sample File	
Rank	Atom	% Incidence	Atom	% Incidence
1	C*C*C	60	C*C*C	61
2	C—C	46	C—C	55
3	O=C	40	O=C	53
			C	
4	O—C	36	C*C*C	38
	C			
5	C*C*C	36	O—C	30
	O			
6	C*C*C	29	C—O—C	30
	N			
7	C*C*C	27	C—C—C	30
8	C—O—C	26	C·C·C	28
			C	
9	C—C—C	24	O=C—O	24
10	C·C·C	19	O—C—C	21
			N	
11	O—C—C	18	C*C*C	20
12	O=S	18	N—C—C	19
			O	
13	O=C	17	C*C*C	19
14	N—C	17	C—N—C	15
	C			
15	O=C=O	16	C·O·C	15
			C	
16	C—C—N	16	C*C*C	15
17	O=S	16	C·C:C	14
	S		C	
18	C*C*C	14	C·N·C	13
19	C—N	14	C—O	13
	C			
	*			
20	C*C*C	14	Cl—C	12

the fragment in the CDB was lower than in the sample

C
*

file, but the fragment C*C*C had a higher incidence, indicating a greater proportion of fused aromatic rings in the CDB [see section on Rings]. Species involving cyclic, nonaromatic bonds, either to carbon or other atoms, were generally more prevalent in the random sample file.

Octuplets and Four-Atom Strings. These are examples of larger, bond-centered fragments. Octuplets were described in the totally-generalized form (see Figure 1), and four-atom strings had bond codes specified and atoms generalized to C or X. Inspection of the analysis results in Tables VI and VII shows that in the CDB fragments containing at least one noncarbon atom, such as XX0302, CX2101, CX2303, C*C—X=X, were more common than in the random sample file. The use of generalized fragments made the numbers of different species generated more comparable for the two files; there were 512 four-atom fragments and 182 octuplets for the CDB, and 519 and 204, respectively, for the random sample file. The highest-ranking, four-atom string containing a tautomeric bond in the CDB was C*C—X=X (incidence 12%), and once again fragments containing cyclic bonds appeared to be more common in the random sample file.

Table VI. Octuplets, Ranked by Incidence

	CDB		Random Sample File	
Rank	Octuplet	% Incidence	Octuplet	% Incidence
1	CC1100	71	CC1100	79
2	CC1201	62	CC1201	65
3	CC1200	59	CC1200	64
4	CX2000	49	CX2100	46
5	CX2010	43	CX2010	43
6	CC2201	39	CX2000	40
7	CX2100	39	CC2201	38
8	CC2211	34	CX2110	34
9	CC2200	33	CX1100	33
10	CX2110	27	CC2200	32
11	CX1100	24	CC2211	26
12	CC1101	20	CC1101	26
13	CC2202	17	CC0200	18
14	CC0200	16	CX1200	18
15	CX1200	15	CC2202	17
16	XX0302	15	CC0100	16
17	CX2101	14	CC1210	16
18	CC0100	14	CC1202	14
19	CX2303	14	CX0100	14
20	CX2200	13	CX1000	13

Table VII. Four-Atom Fragments, Ranked by Incidence

	CDB		Random Sample File	
Rank	Fragment	% Incidence	Fragment	% Incidence
1	C*C*C*C	60	C*C*C*C	62
2	C*C*C—X	47	C*C*C—X	42
3	C—C*C*C	36	C—C—X—C	40
4	C—C—X—C	32	C—C*C*C	39
5	C—C—C—X	25	C—C—C—X	33
6	C*C—X—C	22	C·C·C·C	30
7	C—C—C—C	22	X=C—X—C	27
8	X=C—X—C	21	C·C·X·C	27
9	C·C·C·C	20	C—C—C—C	25
10	X—C*C—X	20	C*C—X—C	23
11	C·C·X·C	20	C—C·C·C	22
12	C*C—X=X	19	X—C—C—X	19
13	C·C·C·X	17	C·C·C·X	19
14	C—C·C·C	16	C—C—C=X	18
15	X—C—C—X	14	C*C—C—X	16
16	C—C—C*C	13	C—C—C*C	16
17	C·C*C*C	13	C·C·C—X	16
18	C*C—C—X	13	C·C*C*C	15
19	C—C*C—X	13	C—C·X·C	13
20	C*C—X=X	12	C·C:C·C	13

Rings. The program used for counting rings has been described previously.[3] It can deal only with monocycles and primary rings in 1:1- and 1:2-fused systems. The main features to emerge from a comparison of the ring analyses for the two files are the smaller proportion of monocycles in the CDB, especially those containing heteroatoms and the larger proportion of six-membered, fused carbocyclic rings. In view of the augmented atom analysis, these fused rings are largely aromatic. The proportions of other fused rings appear to be generally smaller for the CDB than for the random sample file.

CONCLUSIONS

The general characteristics quantified by the previous analysis of the random sample file—i.e., that the variety of species, especially of those with small incidence, in-

creases rapidly as the fragment types increase in size—have been shown to apply to the Common Date Base. The analysis results reflect the differing natures of the two files with respect to less common elements, especially sulfur, which might be expected to be more prominent in a collection containing substances of biomedical interest. The preponderance of carbon, oxygen, and nitrogen in the CDB means that, as for the random sample file, fragments to serve as screens for substructure search containing these elements must be treated in more detail than other fragments. In this respect, the files are similar.

The differences in the results show that a more flexible screen generation program than that previously employed for screen generation with the random sample file[4] may be advantageous. This has now been developed for augmented atoms, octuplets, and four-atom fragments, where the screens are selected on the basis of fragment incidences in the files being used. Thus, the screens set may be different for files of different characteristics, and the changes can be made automatically without the systems operator being aware which fragments are being used. Details of this work will be reported shortly.

ACKNOWLEDGMENT

We thank the Office for Scientific and Technical Information, London, for financial support for this work, and Chemical Abstracts Service for the provision of the two files in machine-readable form. Conversion of the CDB file into a form which could be handled on the ICL 1907 computer at Sheffield was carried out by J. M. Harrison in cooperation with UKCIS Research Unit, University of Nottingham. This work is gratefully acknowledged.

LITERATURE CITED

(1) Crowe, J. E., Lynch, M. F., and Town, W. G., "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part I. Non-cyclic Fragments," *J. Chem. Soc. (C)*, **1970**, 990.
(2) Adamson, G. W., Lynch, M. F., and Town, W. G., "Part II. Atom-Centered Fragments," *J. Chem. Soc. (C)*, **1971**, 3702.
(3) Adamson, G. W., Cowell, J., Lynch, M. F., Town, W. G., and Yapp, A. M., "Part IV. Cyclic Fragments," *J. Chem. Soc. Perkin I*, in press.
(4) Adamson, G. W., Cowell, J., Lynch, M. F., McLure, A. H. W., Town, W. G., and Yapp, A. M., "Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files, *J. Chem. Doc.* **13**, 153-7 (1973).

Technological Forecasting—an Experiment Relating to the Pulp and Paper Industry*

D. J. FLOTO, R. A. HAGSTROM, T. L. HEYING, R. E. MAIZELL,** and W. J. MAYER
Olin Corp., New Haven, Conn. 06504

Received April 30, 1973

The methodology of a technological forecast, integrating consultant, industrial, and academic input from three successive Delphi rounds, is presented. Questions gleaned from information sources and from panel suggestions made during the progress of the study were answered anonymously. Each respondent indicated his personal familiarity with the development or event, his estimate of the stated event's impact, and his estimate of the year in which there is a 50% chance of occurrence. Part of the study's success is determined by convergence and degree of helpfulness. The point is made that a TIS Department can play a leadership role in studies of this kind.

The chemical industry executive has at his disposal a large tool kit for helping him arrive at decisions. These include market research, econometrics, committee meetings, educated intuition, laboratory and pilot plant work, and consultants, to name a few. One of the newer additions to this tool kit is the technique of technological forecasting.

This paper will show how a technical information services department can make use of the Delphi technique as a technological forecasting tool and thereby provide a basis for chemical industry executive decision making.

Our experiment dealt with the future of the pulp and paper industry, and the probable impact of that future on the chemical industry, in particular on the Chemicals Group at Olin.

TECHNOLOGICAL FORECASTING—GENERAL REMARKS

Technological forecasting is 10 to 15 years old, but is still regarded as one of the newer techniques to help in the executive decision-making process. Interest in the field has developed sufficiently for formation of at least one major professional society—the World Future Society—and for publication of a number of professional journals,

* Presented before the Division of Chemical Literature, Middle Atlantic Regional Meeting, ACS, Washington, D. C., January 16, 1973.
** To whom correspondence should be addressed.