Distributions of Fragment Representations in a Chemical Substructure Search Screening System

GEORGE W. ADAMSON, VERITY A. CLINCH, 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 4, 1974

Two analyses of the distributions of representations of chemical compounds in terms of simple structural characteristics have been carried out. The compounds were sampled from the Chemical Abstracts Service Registry System; the structural characteristics consist of a simple hierarchy of bond-centered fragments—simple, augmented, and bonded pairs. The mean number of fragment types per compound increases as the size of the fragments is increased. The resolving power of the fragment representations, in terms of the proportion of compounds in the file having unique descriptions, is high, increasing as the size and number of fragment types increase.

In an earlier paper,[1] we have described the setting up of a screens file for substructure searching of files of chemical compounds. The file consists of bit screens for atoms, rings, and three different types of atom–bond–atom pairs which form a simple, non-overlapping hierarchy (see Figure 1), for a random sample of 28,963 compounds from the Chemical Abstracts Service Registry System. In addition to these screen types, a newer version of the file now also contains screens for larger bond-centered fragments such as octuplets and four-atom fragments.

It is clear that in order to obtain good screenout and precision during searches of such a file, the bit-string representation of a compound should be, if not unique, one of only a small number of strings with the same bit pattern. To determine whether this was in fact the case with our screens file, an analysis of the bit strings it contained was undertaken. Lefkovitz[2] has carried out similar work on the TSS screens.

ANALYSIS OF BIT-SCREENS

Our work has involved two types of analysis of the bit-strings in the screens file. The first involved a straightforward count on the bit-string for each compound, to obtain statistics for the number of bits set for various fragment types. The other analysis entailed comparing the bit strings for different fragment types to collect groups of identical strings, which were then further investigated.

The results obtained using a program to determine the number of bits set per compound for each fragment type are shown in Table I and Figure 2. The simple pair bit-string contained five bits (one for each of five different central bond types), for each of eighteen different common atom–bond–atom pairs, plus four general bits, set according to the nature of the central bond, for other pairs. There were 26,819 compounds in the file which had no general simple pair bits set. For the remaining compounds, the distribution of the number of specific simple pair bits set *vs.* the number of general simple pair bits set was obtained. No compound had all four general bits set. The largest group consisted of 357 compounds which had two ordinary simple pair bits and one general bit set. Of the 60 compounds which had none of the specific bits set, 42 had one, and 18 had two, of the general pair bits set.

The average and maximum number of bits set increased on progressing from simple to bonded pairs, *i.e.*, as the level of pair description and number of bits available increased. The figures for octuplets and four-atom fragments did not fall into this pattern as these fragments are not part of the same simple hierarchy and are described differently. Several compounds had no bits set for a particular fragment type. Of these, 16 compounds had no pair bits set at all. This group included single elements, simple hydrides (which would be treated in the connection table itself as simple elements, since hydrogen is ignored), and structures containing only atom pairs in which one or both of the atoms has a connectivity greater than four. (The presence of such atoms is recorded elsewhere in the bit screen.) In the augmented and bonded pair bit strings, only pairs containing a carbon atom bonded to carbon, oxygen, or nitrogen are included, and then only those pairs having certain specified external connectivities or bond patterns. Thus, the lack of carbon-containing pairs or the lack of the appropriate external connectivity or bond pattern on an otherwise acceptable pair accounts for the 222 and 254 compounds with no augmented or bonded-pair bits set, respectively. The reasons which prevent compounds having any octuplet or four-atom fragment bits set are the same as those already stated for simple pairs. Also, no four-atom fragment bits would be set for compounds containing only three or fewer atoms in any string.[4]

Prior to the second analysis, it was necessary to create files in which the strings were ordered, to bring together identical strings for each fragment type. In the original screens file, each record contained one computer word (24 bits) which included parts of both the bonded and the simple pair bit strings. Hence it was convenient first to expand the strings to give a separate section of the complete string for each of these pair types. Then, using a magnetic tape sort, separate tapes containing the screens records ordered in turn by the bit-strings for bonded pairs, simple pairs, augmented pairs, octuplets, and four-atom fragments were produced. The sort on the bonded-pair string was run with a sort key of 128 characters, to cover all the pair strings, so that the tape produced could be used for analyzing both the bonded-pair string and the whole-pair string.

The program used for group analysis of the bit strings used the sorted tapes, comparing bit strings relating to certain fragment types and obtaining the distribution of sizes of groups with identical bit pattern for the fragment string *vs.* the number of groups of each size. Thus a group of size 1 contained a unique string. The constitutions of the bit strings for some of the larger groups were also investigated. A summary of the results of the comparison is shown in Tables II and III. As would be expected, the number of unique strings increased as more of the pair bit-strings were com-

* Author to whom correspondence should be addressed.

Table I. Analysis of Bits Set for Various Fragment Types

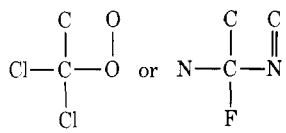
Fragment type	No. of bits available in screen set for this fragment type	Av. no. of bits set per structure	Max. no. of bits set per structure	No. of compounds with no bits set
Simple pair, including 4 general bits	94	5.67	16	16
Simple pair, excluding 4 general bits	90	5.58	16	76
Augmented pair	216	8.45	31	222
Bonded pair	434	10.07	47	254
All pair types	744	24.19	82	16
Octuplet	216	9.87	33	85
Four-atom fragment	360	11.87	61	270

C—C Simple pair 1C—C2 Augmented pair ·C—C Bonded pair C—C·C—X Four-atom string

Octuplet: ABabcd

a = terminal connectivity of A
b = terminal connectivity of B
c = no. of X's attached to A (excluding B)
d = no. of X's attached to B (excluding A)

e.g. CX3120



X = any non-carbon atom (except hydrogen)
· = single cyclic bond

Figure 1. Fragments included in the analyses.

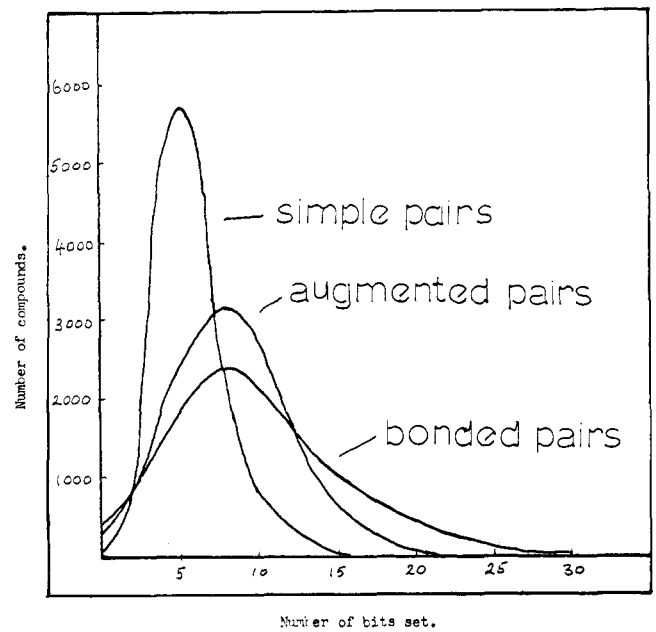


Figure 2. Distribution of number of bits set per compound.

Table II. Comparison of Screens Records by Fragment Type

Fragment type	Unique strings No.	Unique strings %	Largest group of identical strings
Simple pair, excluding 4 general bits	4,868	16.81	529
Simple pair, including 4 general bits	5,266	18.52	519
Augmented pair	18,033	62.25	222
Bonded pair	21,842	75.41	254
Simple + augmented pair	23,032	79.52	36
Simple + augmented + bonded pair	25,410	87.73	36
Octuplet	18,458	63.73	150
Four-atom fragments	17,107	59.06	270

Table III. Distribution of Groups of Identical Strings

Size of group	Total number of groups in size range: Simple pairs	Augmented pairs	Bonded pairs	All pairs	Octuplets	Four-atom fragments
1–10	7874	21,066	23,944	26,804	21,321	19,982
11–20	205	71	34	12	71	91
21–30	69	13	11	1	25	21
31–40	44	5	3	1	4	11
41–50	20	2	0	...	5	6
>50	59	6	5	...	4	9

pared, *i.e.*, as more of the more specific differential pair descriptions were considered. The highest figure, 87.73%, reached when the entire pair bit string was used, indicates that a very large proportion of the pair strings consists of unique representations. This points to the appropriateness of the pair hierarchy as a system for structure description in terms of fragments. The percentage may be even greater if the octuplet and four-atom fragments strings are considered along with the pair strings, since comparison of strings for each of these fragments alone indicates over 50% unique bit string representations.

It is interesting to note that the number of unique bit patterns obtained using the octuplet bit string alone was approximately the same as that obtained using only the augmented pair bit string. Both these strings have 216 bits available. The figure for octuplets is surprisingly high, since this fragment type involves only a generalized representation of atoms and bonds. Table III also shows a similarity between octuplets and augmented pairs, with respect to group size distribution.

The inclusion or exclusion of the four general simple pair bits in the bit string comparison makes little difference to the number of unique simple pair bit strings. This indicates that the resolving power of the specific simple pairs chosen for inclusion in the simple pair bit string is high.

As the number of unique strings rose with the length of bit string considered, the number of large groups of identical strings decreased, as did the size of such groups. The largest groups of identical strings found on comparing strings for augmented pairs, bonded pairs, and four-atom fragments comprised those compounds with no bits set at that level. The largest group of identical complete pair strings contained those strings for which only the general simple pair bit for a single chain bond was set. The size of

this group (36 compounds) could be reduced by including a greater number of specified atom pairs in the simple pair bit string, or by dividing the group into two for the general pairs C–X and X–X. The other groups of identical strings for all pair types taken together contained only bits for very common pairs, accounting for the lack of differentiation of bit string representations. For example, there were 26 compounds, the bit strings of which contained only bits for simple pairs C–C and general single bond augmented pairs OC–C1 and 1C–C1, and bonded pairs C–C– and –C–C–. This finding was also borne out at the separate pair levels and for octuplets and four-atom fragments, and is in line with the general principle that structures containing rarely occurring features are easy to search for, whereas common structural features must be described in greater detail to be easily retrievable. This group analysis is similar to earlier work involving the distribution of molecular formula group sizes[5] and leads to similar conclusions.

CONCLUSIONS

Good differentiation of the majority of structural representations in the file is obtained using the set of pair screens at present available in the Sheffield substructure search system. The inclusion of other fragment types as screens promises further improvements in the direction of unique screen representations for structures, leading to improvements in systems performance.

ACKNOWLEDGMENTS

We thank Chemical Abstracts Service for provision of the data-base, and the Office for Scientific and Technical Information, London, for financial support for this work. V. A. C. acknowledges the award of an OSTI Research Studentship.

LITERATURE CITED

(1) 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 (1973).

(2) Milne, M., Lefkovitz, D., Hill, H., and Powers, R., "Search of CA Registry (1.25 Million Compounds) with the TSS," *J. Chem. Doc.*, **12,** 183 (1972).

(3) Crowe, J. E., Lynch, M. F., and Town, W. G., "Analysis of Structural Characteristic of Chemical Compounds in a Large Computer-Based File. Part 1. Non-cyclic Fragments," *J. Chem. Soc. C,* 990 (1970).

(4) Adamson, G. W., Creasey, S. E., and Lynch, M. F., "Analysis of Structural Characteristics of Chemical Compounds in the Common Data Base," *J. Chem. Doc.*, **13,** 158 (1973).

(5) Bragg, J. H. R., Lynch, M. F., and Town, W. G., "The Use of Molecular Formula Distribution Statistics in the Design of Chemical Structure Registry Systems," *J. Chem. Doc.*, **10,** 125 (1970).

Semiautomatic Coding of Steroid Markush Formulas

J. FITTING, H. LEHNA, G. RIEGE, and K. SPECHT*

Research Laboratories, Schering AG, Berlin/Bergkamen, Germany

Received December 27, 1973

Manual coding of complicated Markush formulas is very time consuming. A semiautomatic method for the encoding of steroid Markush formulas is described. Manual encoding is necessary only for the basic structure and the variables. The permutation is done automatically by computer.

The Pharma-Dokumentationsring e.V. is encoding the FARMDOC and AGDOC patents of the CPI service (sections B and C) by using the Ringdoc- and Pestdoc-Codes. This work is done in cooperation among the Ring members of the corresponding working group. Years ago (in pre-CPI time), the Ring members worked out special rules for patent encoding because they were and are convinced that the expense of time and money is justified.

The patent coding rules of the RING allow the overcoding of chemical and biological information to a certain extent depending on RINGDOC and PESTDOC coding rules. According to these rules the encoding of Markush formulas will lead very often to a remarkably great amount of punch cards, and, therefore, it is time-consuming. As far as organic compounds other than steroids are concerned, a publication about a semiautomatic coding method has been issued from Roussel-Uclaf (H. Deforeit, A. Caric, H. Combe, S. Leveque, A. Malka, and J. Valls, "CORA—A Semiautomatic Coding System Application to the Coding of Markush Formulas," *J. Chem. Doc.*, **12,** 230 (1970)).

The method described here for semiautomatic coding of steroid Markush formulas has been developed by Schering in cooperation with the research oriented data processing section (Datenverarbeitung Forschung) and the central documentation department.

With regard to the overcoding rules, a computer program has been developed for the semiautomatic coding. The "Basic-Structure" and the variable structures are coded once only. The permutation will be done by the program. By indicating additional conditions, some of the resulting cards can be changed following permutation.

Sometimes, the encoding of steroid Markush formulas requires 100, 1000, or more punch cards. This is due to the fact that Markush formulas sometimes cover 100,000 or even 1 million mathematically possible combinations, and that—for retrieval purposes with a minimum of false

* Author to whom correspondence should be addressed.