teractive Versatile, Three-Dimensional Display, Manipulation and Plotting System for Biomedical Research," *Ibid.*, **12**, 234–7 (1972).
(5) Heller, S. R., and Feldmann, R. J., "An Interactive NMR Chemical Shift Search System," *J. Chem. Ed.* **49**, 291 (1972).
(6) Heller, S. R., and Jacobson, A. E., "GINA—A Graphical Interactive Nuclear Magnetic Resonance Analysis Program," *Anal. Chem.* **44**, 2219–22 (1972).
(7) Heller, S. R., "Conversational Mass Spectral Retrieval System and Its Use as an Aid in Structure Determination," *Ibid.*, **44**, 1951–61 (1972).
(8) Heller, S. R., Fales, H. M., and Milne, G. W. A., "An Interactive Mass Spectral Search System," *J. Chem. Ed.* **49**, 725 (1972).
(9) Heller, S. R., Fales, H. M., and Milne, G. W. A., "A Conversational Mass Spectral Search and Retrieval System. II. Combined Search Options," *Org. Mass Spectrom.* **7**, 107–14 (1972).
(10) Heller, S. R., Fales, H. M., and Milne, G. W. A., "A Conversational Mass Spectral Search System. III., Display and Plotting of Spectra and Dissimilarity Comparisons," unpublished data.
(11) Heller, S. R., "DCRT/CIS Mass Spectral Search System User's Manual," DCRT, NIH, Bethesda, Md., November 1972.
(12) *Chem. Eng. News*, pages 16–17, Oct. 16, 1972.
(13) *Mass Spectrom. Bull.*, page iii, August–December 1972.
(14) *Ind. Res.*, page 34, December 1972.
(15) *Res./Develop.*, page 44, February 1973.
(16) *J. Chem. Doc.*, **13**, 47 (1973).
(17) Scott, W., Maxwell, D. C., and Ridley, R. G., "Classification of Mass Spectral Data Files," Proc. 20th ASTM Meeting, pages 372–374, Dallas, Tex., June 1972.

Relationship between Query and Data-Base Microstructure in General Substructure Search Systems

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

Received May 2, 1973

The distributions of bond-centered fragments which form a simple hierarchy have been investigated for a sample of substructure search queries. They are found to conform with the general pattern of fragment incidences in a file which might be used in a substructure search system. This result has been assumed in previous work on the design of screening systems for substructure search.

In previous work carried out at Sheffield on the design of screening systems for substructure search of chemical structure files,[1] it was assumed that characteristics of search requests would roughly mirror those of the file to be searched. Thus, a search for compounds containing a relatively infrequent atom, although easy to carry out, would not be requested often, whereas searches for structures containing various combinations of carbon, oxygen, and nitrogen atoms would be required much more frequently.

This assumption has been borne out by experience with fragmentation codes,[2] and we now describe our own recent work on this topic. A sample of queries was analyzed to find the incidences of fragments used as screens in the requests, and these were then compared with the incidences of such screens in the file to which the queries were to be addressed. The sample consisted of 50 user queries, supplied by the Oxford Experimental Research Unit. These were "real" queries which might be addressed to the search system. A sample of 50 queries derived from the titles of articles covered by *Current Abstracts of Chemistry* (CAC) for May 26, 1971, was also examined. These queries, as might be expected from their rather artificial derivation, were much more specific than the real queries which were usually stated in fairly general terms. This was shown by an examination of the incidences of cyclic, acyclic, aromatic, and nonaromatic bonds for the two sets of queries. The CAC queries were detailed as having alkyl and/or aromatic substituents, whereas the Oxford queries were mainly concerned with derivatives of a certain broad class of compounds, regardless of the nature of the substituent. The file used for substructure search was a random sample of 28,963 structures taken from the Chemical Abstracts Service Registry System.

Each query was analyzed in terms of differentiated simple, augmented and bonded pairs, certain species of which are used in the screen set.[1] This series of pairs forms a simple hierarchy. Some queries requested potentially varied substitution patterns, which would be covered by Boolean OR groups in a real query put to the substructure search system.[3] In these cases, all possible fragments were listed, and a value of one was assigned to each OR group, and also to each fragment which must be present (these latter are covered by AND logic). Within each OR group, equal values were assigned to equally acceptable pairs. In some cases, there were several distinct possibilities which affected likely pair fragments. For example, in $R\overset{a}{-}CH_2\overset{b}{-}X$, where R can be alkyl or aryl and X is not hydrogen, there are four possibilities for bond (a) at the augmented pair level. These are OC—C1, 1C—C1, 1C—C2, and 1C—C3. If R is aryl, 1C—C2 must be present; it may also be present if R is alkyl. Each major possibility is assigned an equal value, and the derivatives from each are then further divided. Thus, R may be aryl, in which case 1C—C2 would have the value $^1/_2$, or it may be alkyl, and the value $^1/_8$ would be assigned to each of the four possibilities. Hence, in the OR group for bond (a), 1C—C2 would have the value $^5/_8$, with OC—C1, 1C—C1, and 1C—C3 being assigned $^1/_8$ each.

* To whom correspondence should be sent.

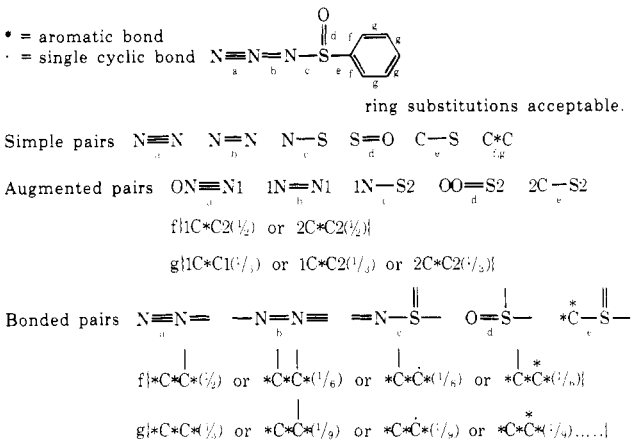

Figure 1. Calculation of incidence values for pairs in a query

At bonded pair level, a similar, though more complex, procedure was adopted. However, here there was often a very large number of possibilities, which should each have been assigned a very small, and perhaps therefore insignificant, value. This could have resulted in inaccuracy if one were to try to take all fragments into account, and so the arbitrary value of 1/12 was taken as the minimum that would be considered for any pair in a query. An example of the assignment of values for incidence calculations is given in Figure 1.

Where the same fragment was included more than once with different values, the highest value was taken as that to be used in calculating fragment incidence, thus ensuring that the full value of one was assigned to a fragment that must be present, and that no less important fragment took a higher value.

The main interest concerning the fragment analysis of the queries centered around the most common fragments and the effect of the hierarchy simple-augmented-bonded pair on the fragment distribution. This is in line with investigations carried out on the random sample file.[4-7] Thus, pair types occurring less frequently in queries were not analyzed in detail. The most common simple pairs were noted, and from these were derived the related augmented pairs and their rankings. From these in turn, the related bonded pairs were derived. At this highest level of description, there was a considerable spread of fragments, since even the bonded pairs derived from the most com-

Table I. Incidence of Simple Pairs in Queries and Random Sample File

	CAC Queries		Oxford Queries		All Queries		Random Sample File[a]	
Rank	Fragment	% Incidence	Fragment	% Incidence	Fragment	% Incidence	Fragment	% Incidence
1	C—C	76	C·C	70	C·C	58	C—C	85
2	C*C	63	C·N	38	C—C	56	C*C	63
3	C·C	46	C—C	36	C*C	46	C—O	56
4	C—N	40	C=O	34	C·N	33	C=O	53
5	C—O	38	C:C	30	C=O	32	C—N	51
6	C=O	30	C*C	28	C—N	31	C·C	51
7	C·N	28	C·O	22	C—O	29	C·N	27
8	C:C	24	C—N	22	C:C	27	C:C	22
9	C:N	20	C—O	19	C·O	21	C·O	16
10	C·O	19	C·S	13	C:N	15	C—S	12
11	C=C	14	C:N	10	C=C	10	C—Cl	12
12	C*N	12	C=C	6	C·S	8	C=C	11
13	N·N	10	O=S	6	C*N	8	C:N	10

Table II. Incidence of Augmented Pairs

	CAC Queries		Oxford Queries		All Queries		Random Sample File[a]	
Rank	Fragment	% Incidence	Fragment	% Incidence	Fragment	% Incidence	Fragment	% Incidence
1	1C*C2	57	1C·C2	35	1C*C2	38	1C*C2	62
2	1C*C1	49	2C=O0	34	2C·C2	34	1C*C1	57
3	2C*C2	36	2C·C2	34	2C=O0	32	2C=O0	52
4	2C·C2	33	2C*C2	24	1C*C1	31	1C—C2	41
5	2C—C2	31	2C·C3	19	2C*C2	30	2C*C2	39
6	2C=O0	30	1C*C2	19	1C·C2	29	2C—C2	35
7	1C—C2	25	1C·C1	17	2C—C2	23	1C·C2	32
8	1C·C2	24	2C—C2	16	2C·N2	20	0C—C2	32
9	2C—O1	23	2C·N2	14	1C—C2	18	2C·C2	31
10	0C—C2	21	1C*C1	13	2C—O1	15	2C—O1	31
11	2C·N1	17	1C·C3	13	2C·O1	15	1C·C1	29
12	2C:N1	17	2C·O1	12	2C·N1	14	1C—C1	27
13	2C·O1	17	2C·N1	12	1C·C1	14	2C—O0	25
14	2C—N2	17	1C:C2	12	2C·C3	14	0C—C1	23
15	2C·N2	16	2C—O0	11	2C:C2	12	2C—N1	21
16	2C—N1	16	1C:C1	10	2C—O0	12	2C—N2	18
17	2C:C2	15	1C—C2	10	0C—C2	12	1C—O1	17
18	0C—C1	15	1C·O1	10	2C—N2	12	0C—C3	15
19	0C—N2	14	2C:C2	9	2C:N1	12	2C·C3	14
20	2C—O0	13	2C·S1	9	2C—N1	10	2C·N2	14

Table III. Incidence of Bonded Pairs

	CAC Queries		Oxford Queries		All Queries		Random Sample File[8]	
Rank	Fragment	% Incidence	Fragment	% Incidence	Fragment	% Incidence	Fragment	% Incidence
1		51		32		31		57
2		49		18		30		55
3		24		13		28		41
4		20		12		19		26
5		17		12		15		24
6		15		11		11		22
7		15		10		9		20
8		14		10		9		19
9		14		9		9		18
10		13		9		9		18
11		13		9		8		17
12		11		8		8		17
13		11		8		8		16
14		11		8		8		14
15		10		7		8		13
16		10		7		7		13
17		10		7		7		13

mon augmented pairs were present in only relatively few compounds; the results were made more diffuse by the presence of varying substitution patterns.

Fragment incidences for simple, augmented, and bonded pairs are presented in Tables I to III, for the two separate sets of queries, for all queries together, and for the random sample file. All exhibit a Zipfian distribution, with incidence falling off rapidly with decreasing rank. The cut-off incidence level for simple pairs was taken as 7% for the total set of queries, and above this level there were 13 simple pairs. These were analyzed further to the augmented pair level. The cut-off level for augmented pairs which were analyzed to the bonded pair level was also 7%.

It was not possible to perform rank correlation tests on the results, since the same fragments did not necessarily occur in the separate top n rankings. However, inspection of the results shows that the characteristics of the Oxford queries are reasonably close to those of the compounds in the random sample file. The generic nature of the queries accounts for the differences, as is seen by the even closer correspondence of the file characteristics with those of the more specific queries taken from article titles.

All three sets of results for queries show some similarities. Thus, the first 10 augmented pairs in each of the three groups are derived from the first six or seven simple pairs for that group. The diversity on progressing to bonded pairs is shown by the fact that the top 10 are derived from the augmented pairs ranked 1, 2, 4, 6, 7, 10, 14, 16, and 18 for Oxford queries; 1, 2, 3, 6, 9, 12, and 18 for CAC queries; and 1, 3, 4, 6, 10, 13, 15, and 19 for all queries combined. However, five simple pairs (C·C, C*C, C:C, C=O, and C·O) give rise to the first 17 ranked bonded pairs from the Oxford queries, and 6 simple pairs (C—C, C*C, C:C, C:N, C—O, and C=O) account for the first 17 ranked bonded pairs in the CAC queries. Only three pairs (C*C, C:C, and C=O) are common to both groups. In the random sample file, the first 6 simple pairs (C—C, C*C, C—O, C=O, C·C, and C—N) give rise to the first 19 augmented pairs and at least the first 17 bonded pairs, the first 10 of which are derived from the augmented pairs ranked 1, 2, 3, 5, 11, 12, and 14.

CONCLUSIONS

The closeness of the results obtained on analysis of pair fragment incidence characteristics indicates that it is reasonable to assume that the characteristics of a file to be used in a substructure search system will be similar to the requests addressed to such a file. Thus, efficient searches are achieved if such fragments are used as screens in the substructure search system.

ACKNOWLEDGMENT

We thank P. Leggate and J. F. B. Rowland of the Oxford E.I.U. for providing the sample of user queries, and Chemical Abstracts Service for the data base. V.A.C. acknowledges the award of an OSTI information science 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,** 133 (1973).

(2) Craig, P. N., and Ebert, H. M., "Eleven Years of Structure Searching Using the SK&F Fragment Codes, *J. Chem. Doc.* **9,** 141-6 (1969).

(3) Adamson, G. W., Bush, J. A., Lynch, M. F., and McLure, A. H. W., unpublished data.

(4) 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,** p. 990.

(5) Adamson, G. W., Lynch, M. F., and Town, W. G., "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part II. Atom-Centered Fragments," *J. Chem. Soc. (C),* **1971,** p. 3702.

(6) Adamson, G. W., Cowell, J., Lynch, M. F., McLure, A. H. W., Town, W. G., and Yapp, A. M., "Analysis of Structural Characteristics of Chemical compounds in a Large Computer-Based File. Part IV. Cyclic Fragments," *J. Chem. Soc., Perkin I,* **1973,** p. 863.

(7) Adamson, G. W., Creasey, S. E., Eakins, J. P., and Lynch, M. F., "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part V. More Detailed Cyclic Fragments," *J. Chem. Soc., Perkin I,* in press.

(8) 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).

Computerized Monitoring of the Inventory and Distribution of Research Chemicals*

STEPHEN J. FRYCKI,** FREDERICK F. GIARRUSSO, PAUL A. ROSKOS, DOLORES E. DANCSECZ, SALVATORE J. LUCANIA, and DENNIS M. O'BRIAN***

E. R. Squibb & Sons, Inc., Worldwide Headquarters, Princeton, N. J. 08540

Received May 29, 1973

A one-time data entry system, coupled with an efficient use of the computer, which provides inventory management, distribution, and audit reporting, the ability to answer special queries, and to produce customized reports is described. Results achieved by the system have been: paperwork has been drastically reduced, streamlined, or both; clerical labor has been significantly reduced; the preparation of reports to management has been automated; and the Research Chemicals Distribution Center can devote more time and resources to its prime function, the distribution of chemicals for biological testing.

Current research efforts by the pharmaceutical industry are creating large numbers of potential drugs and great volumes of data associated with their storage, distribution, and use for biological testing.

In the past, when research teams of the pharmaceutical industry were much smaller, the bench scientist would keep track of his own inventory of chemicals and the biological data resulting from their testing. This is still an excellent practice. However, the volume of data generated by current multidisciplinary research teams and the need for information by the professionals who comprise such teams has required a greater expenditure of time and effort than is optimal for any individual whose major effort is in research.

More recently, a central depository of data, manually controlled, was in use. This method, although accurate, was hindered by the inordinate amount of manual labor necessary to search through huge volumes of data for retrieving information requested by the research staff. Additionally, the manual processing of more than 30,000 transactions of chemical compounds per year was approaching unmanageability. The manipulation of data proved to be particularly difficult when a search of the historical records or of complete inventories was needed.

The clearest way to alleviate the problem appeared to be the use of a computerized system that would make the entering of data easy and simple and would be versatile in the retrieval of data. To this end, the Chemical Audit and Distribution (CHAUD) System was designed.

SUMMARY OF MANUAL SYSTEM PROBLEMS

The difficulties associated with the manual operation of the chemical distribution function were divided into six problem areas:

1. Extraordinary volume of paper and paperwork
2. Multiple entries of the same data
3. Long search times required for any queries other than one variable lookup—e.g., difficulty in finding all chemical compounds that had been tested in a number of biological assays
4. Difficulty in maintaining adequate inventories of important compounds
5. Much manual time expended in producing regularly scheduled reports
6. No simple way of maintaining "history files"

Associated with the third problem area were the typical types of queries that were handled by the Research Chemicals Distribution Center:

1. Find all compounds sent for examination in test number NNNN

* Presented at the 8th Middle Atlantic Regional Meeting, Washington, D. C., Jan. 16, 1973.

** To whom correspondence should be addressed.

*** Present address: Johnson and Johnson, Inc., New Brunswick, N. J.