tape, and the computer skips to the next one. If there is a match, the machine searches only the desired groups from the update tape. When an input document satisfies the strategy, it is saved on the output tape for later printing. When the end of the update tape is reached, it is rewound and the process is repeated for the next strategy.

SUMMARY

If an information center is to be successful, it must be responsive to the demands of its users and clients. If a center has its own computer system, it can schedule batched runs, special runs, or evening runs to satisfy client demands, to meet higher priorities, or to overcome equipment failures. When a center has its own machine, it is paying a flat rental fee or fixed monthly amortization charge. Thus, additional computer use results in a lower per unit cost. Until recently, computer systems with good input/output were too expensive for most centers. Now, small, low-cost machines are available that permit a center to consider acquiring its own dedicated computer system. This paper has described a chemical information retrieval system currently being used on such a machine at the New England Research Application Center. Future work will report on the operational characteristics of this system.

ACKNOWLEDGMENT

We thank Stuart Harris for his contributions to programming of the system.

LITERATURE CITED

(1) Swid, R. E., "Linear *vs.* Inverted File Searching on Serial Access Machines," 26th Annual Meeting of the American Documentation Institute, Chicago, Ill. October 1963.
(2) Prentice, D., deGraw, G., Smith, A., and Warheit, I., "1401 Information Storage and Retrieval System (The Combined File System) 1401 IBM General Program Library 10.3.047."
(3) Starke, A. C., Whaley, F. R., Carson, E. C., and Thompson, W. B., "GAF Document Storage and Retrieval System," *Amer. Doc.* **19** (2), 173–80 (1968).
(4) Williams, Martha E., and Schipma, Peter B., "Design and Operation of a Computer Search Center for Chemical Information," *J. Chem. Doc.*, **10**, 158–62 (1970).
(5) Roberts, Anita B., Hartwell, Ieva O., Counts, Richard W., and Davila, Roberta A., "Development of a Computerized Current Awareness Service Using Chemical Abstracts Condensates," *Ibid.*, **12**, 221–3 (1972).
(6) Wilde, D. U., "Iterative Strategy Design," *Amer. Doc.* **20**, 90–91 (1969).
(7) Wilde, D. U., "Using a Small/Low Cost Computer in an Information Center," Proc. A.S.I.S. Mid-Year Regional Conference, Dayton, Ohio, May 1972.
(8) IBM Corporation, "TEXT-PAC, S/360 Normal Text Information Processing, Retrieval, and Current Information Selection System," (360D-06.7.020).
(9) Williams, M. E., *et al.* "Educational and Commercial Utilization of a Chemical Information Center," IITRI Rep. No. **C6156-18**, July 30, 1972, Chicago, Ill.
(10) Onderisin, E. M., "The Least Common Bigram: A Dictionary Arrangement Technique for Computerized Natural-Language Text Searching," IITRI, Chicago, Ill.

An Evaluation of a Substructure Search Screen System Based on Bond-centered Fragments

GEORGE W. ADAMSON, JUDITH A. BUSH, ALICE H. W. McLURE, and MICHAEL F. LYNCH
Postgraduate School of Librarianship and Information Science, University of Sheffield, Sheffield S10 2TN, England

Received October 8, 1973

A substructure search screening system based on bond-centered fragments has been evaluated using 108 queries derived from user SDI profiles. The average screenout value obtained was 98.42%. Simple, augmented, and bonded pairs are used as a hierarchy of structural descriptors giving easy coding and good performance for both general and specific queries.

This paper reports on an evaluation of the Sheffield screen search system at a stage in its development and forms one of a series reporting on the Sheffield substructure search system. The basic philosophy of the system and a description of the screen search system have already been presented.[1] The evaluation has been carried out using queries obtained from user profiles supplied by the Experimental Information Unit at Oxford. A quantitative investigation of the gross structural characteristics of queries has also been made. The evaluation so far has only involved measurements of screenout. Precision will be determined at a later stage when iterative search programs become available locally. The queries were run against a bit screen file generated from the Chemical Abstract Service sample file of 28,963 compounds. The bit screen layout has already been described[1] and uses simple, augmented, and bonded pairs[2] as a hierarchy of structural descriptors. Thus, the screen generation program is designed so that at whatever level a fragment is initially defined in a structure from the file, the lower levels of description are also automatically included. In the description of queries on the other hand, fragments are described in the query bit string at the level specified in the search question. If a description is not available at this level in the screen set, then the search program automatically describes the fragment at the next less specific level. With

this facility and the availability of pair descriptions at three different levels in the search file bit string, it is not necessary to know in advance the levels of description available for encoding, and pairs may be described in the search question at the highest level thought to be appropriate. This leads to cheaper and easier coding of generic queries without sacrificing high screenout for more specific queries. It also leads to faster search times in the case of generic queries by reducing the need for 'OR' logic.

A total of 108 queries were extracted from user SDI profiles supplied by the Oxford Experimental Information Unit. The user profiles were analyzed, and the search requirements were assessed and translated into a form suitable for input to the screen search program. Some problems were encountered in interpreting the original profiles. For example, in some queries, the structural requirements were not clearly defined or a large number of options were specified. In such cases, the query would be coded so as to satisfy certain minimum requirements. In the cases where a limited number of alternatives was specified in a given profile, these were sometimes handled as separate queries, and sometimes as a single query using 'OR' logic or general fragments. Where considerable similarity occurred in the coded queries, individual handling was of questionable value, particularly, if an infrequently occurring structural component was also present. Related to the problem of similarity of queries was that caused by the availability of the various pair descriptions in the system, and occasionally, it was convenient to group together related substructure queries because limitations in the type of fragment description available prevented their resolution. However, the results show that this did not seriously depress the screenout levels obtained. There was a small number of queries which asked for groups of compounds which were better defined wih respect to origin or properties than by their structures—e.g., carbohydrates, alkaloids, terpenes, etc. These queries do not seem well suited for a system which deals only with structure searches but could be answered better by a text or structure/text searching system.

CODING OF SEARCH QUERIES

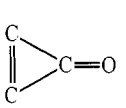
The system requires the coder to analyze each query and decide on its minimum requirements. In the version of the system evaluated, counts of various structural features, ring descriptions, and bond-centered fragment descriptions at the bonded, augmented, and simple pair levels are then chosen accordingly. The system is able to deal with combinations of fragments linked by logic in the form:

A AND B AND ... AND (D OR E OR ...) AND (G OR H OR...)
.... NOT (J OR K OR ....)

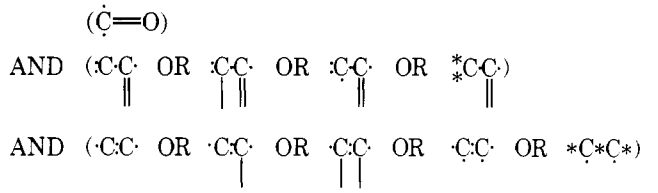
where A, B, C, etc., represent structural fragments.

The user does not need to know in detail which fragments constitute the screen set and chooses the levels of description which best fit the demands of the query. The program then matches the fragments coded for the query with the fragments in the screen set. In the following example, all fragments introduced at the bonded pair level had been included in the screen set, and none was reduced by the program to a lower level of description for use in the query bit strings.

Query Example.[1] All structures containing



including all fused and substituted derivatives. This gave a screenout of 99.47%. The minimum atom requirements were expressed in a series of counts. In this case, a minimum of one oxygen atom is indicated, but if, for example, thioketone derivatives had been equally acceptable, neither the sulfur atom nor the oxygen atom could be specified here. Minimum bond requirements were expressed similarly. The query does not exclude the possibility that the double bond is part of an alternating or delocalized fused ring system and so the possibility that the data base contains such a structure has to be allowed for in the query. Thus, the ring double bond becomes a variable part of the substructure and cannot be included in the minimum bond requirements. The two single bonds on the other hand are definite requirements as is the double bond of the carbonyl group and each can be identified in this section. Atom descriptions which are used to specify atoms other than C, O, N, S, F, and Cl were not required in this example. A number of bond-centered fragment descriptions were included. They are shown below. First, all invariant parts of the substructure were described in one logical 'AND' group, and following this all variable parts of the substructure by a combination of fragments in a series of logical 'OR' groups. Thus the carbonyl function was described at the bonded pair level to establish its occurrence on a ring system, and since only one type of external bond pattern is possible in this case, the fragment was included as part of the logical 'AND' group. The ring single bonds were also described at the bonded pair level to establish the relationship between the ring double bond and the carbonyl group. In this case, various external bond patterns are permissible, and to include all possibilities, it was necessary to describe a series of bonded pairs in the form of a logical 'OR' group. As fusions are only possible at the ring double bond, it was again necessary to describe this bond at the bonded pair level to leave no doubt as to the nature of possible substituents. Thus, if only one of the atoms of the pair is substituted, then the substituent must be acyclic. If both atoms are substituted, the substituents must be exclusively acyclic or exclusively cyclic. The required bonded pair descriptions were included in a second logical 'OR' group. Finally, a ring formula description was included for the three-membered ring, and as fusions were optional, the ring was described as a 3-membered carbocycle occurring either as a monocycle or in a 1:2-fused ring system. These two possibilities were included in a third logical 'OR' group which is not illustrated. The bond-centered fragments used in the query are shown below. Bonds indicated by lines occur in chains, those by dots in rings, and asterisks indicate delocalized ring bonds.



QUERY CHARACTERISTICS AND SCREENOUT PERFORMANCE

To save space the results reported here will not refer individually to all 108 queries, but they will be treated in groups. The list of the individual queries and of the screenout value obtained for each is available from the authors.

The average percentage screenout for all 108 queries was 98.42%, and the average search time per query was 28 seconds of CPU time for a file size of 28,963 structures.

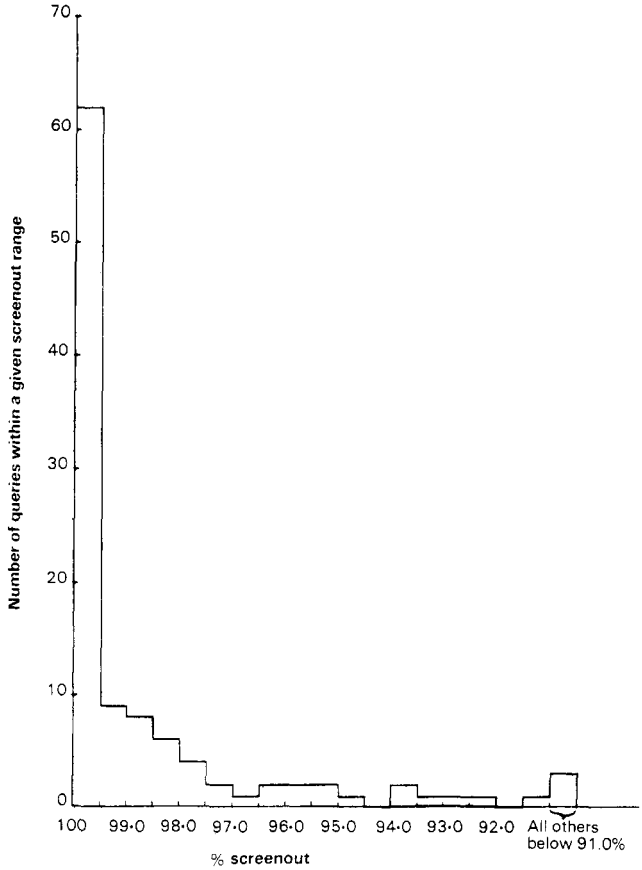


Figure 1. Distribution of the screenout values obtained in the evaluation

The number and percentage of queries falling within a given screenout range over the range 100–91% was examined and is shown in Figure 1. The results show a highly skewed distribution over this range with 57.4% of queries screening out in the range 100 to 99.5%. Only 3% of queries gave screenout values lower than 91%. The screenout values in the range 100 to 99.5% also exhibited a skewed distribution with 12.0% of queries giving 100% screenout, 9.3% of queries giving 99.99%, 6.4% of queries giving 99.98%, and 5.6% of queries giving 99.97% screenout.

Another analysis of the results was carried out in which the queries were classified into seven groups. The results are summarized in Table I, and a detailed analysis is shown in Figure 2. An analysis based in part on similar gross features of substructure search queries at NCI has been reported.[4] The results of an analysis of queries based on the occurrence of simple, augmented, and bonded pairs will be reported elsewhere.[2]

Group (i) contains queries where the part of the substructure defined by the inquirer was exclusively acyclic and group (ii) those where the substructure was exclusively cyclic. Group (ii) excludes queries which specified ring systems with acyclic substituents and these are included in group (iii). Both groups (i) and (ii) gave an average screenout below the over-all average. Group (iii), where both cyclic and acyclic components are specified in the substructure query, formed the largest group and gave the highest average screenout.

The remaining groups define queries where the substructure or part of the substructure is unspecified in that it may be either acyclic or part of a ring system. Very few queries fell into these categories. Group (iv) contains queries where the specified part of the substructure is exclusively acyclic, and this group gave an average screenout well below the over-all average. Group (v) contains queries where the specified component is exclusively cyclic. Only one query fell into this category, and the relatively good screenout in this case was due to the well defined cyclic component. In group (vi), the whole substructure may be either part of a chain or part of a ring system. Very few queries fell into this category, and as expected, the average screenout for the group was well below the over-all average.

Table I. Analysis of Query Types

Query Type	Number of Queries	Percentage of Queries	Average Screenout %
(1) Acyclic	20	18.5	98.94
(2) Cyclic	38	35.2	98.16
(3) Cyclic + acyclic	44	40.8	99.11
(4) Acyclic + unspecified component	2	1.9	95.62
(5) Cyclic + unspecified component	1	0.9	98.51
(6) Unspecified	3	2.8	90.27
(7) Cyclic + acyclic + unspecified	0	0	—
Total	108	100	98.42

As shown in Figure 2, the groups including specified ring systems [groups (ii), (iii), and (v)] were subdivided according to ring type, and each subdivision was in turn broken down into two groups identifying those queries which allowed fused derivatives and those which did not. As only one query occurred in group (v), it is possible to make comparisons only between groups (ii) and (iii). In group (ii), monocyclic and bicyclic ring systems were present in almost equal proportions, and for both of these subgroups, screenout was below the over-all average. In group (iii), on the other hand, very few bicyclic queries occurred and monocyclic queries formed the largest subgroup. Both subgroups in this case gave an average screenout above the over-all average. Polycyclic ring systems occurred in groups (ii) and (iii) in roughly equal proportions, but those occurring in group (iii) gave a higher average screenout which was well above the over-all average. Of the polycyclic ring systems occurring in group (ii) most consisted of 1:2 fusions only, and this was also the case in group (iii), although in the latter group a lower proportion of the queries fell into this category. One query, where only part of the ring system is defined in the substructure, occurred in group (ii), and the high screenout in this case was due to the requirement for an oxygen-oxygen ring bond. A larger percentage of such queries occurred in group (iii). None of these included an infrequently occurring component, and so the average screenout for the group was very much below the over-all average. Finally, in both groups (ii) and (iii), within each subgroup a lower screenout was obtained where fusions were allowed, indicating the effectiveness of bonded pairs in minimizing the retrieval of fused derivatives when they are not requested. No queries arose in the groups in question where a fully defined ring system occurred in conjunction with a partially defined ring system.

In all queries, unless the inquirer stated otherwise, it was assumed that all fusions, substitutions, and branching patterns in parts of the structures unspecified by the user were possible answers, and the system was designed, and queries coded, such that no possible answer was excluded.

DISCUSSION

The high screenout obtained for most of the queries indicates the value of the use of bond-centered fragments in screening systems. The results show that, for the present

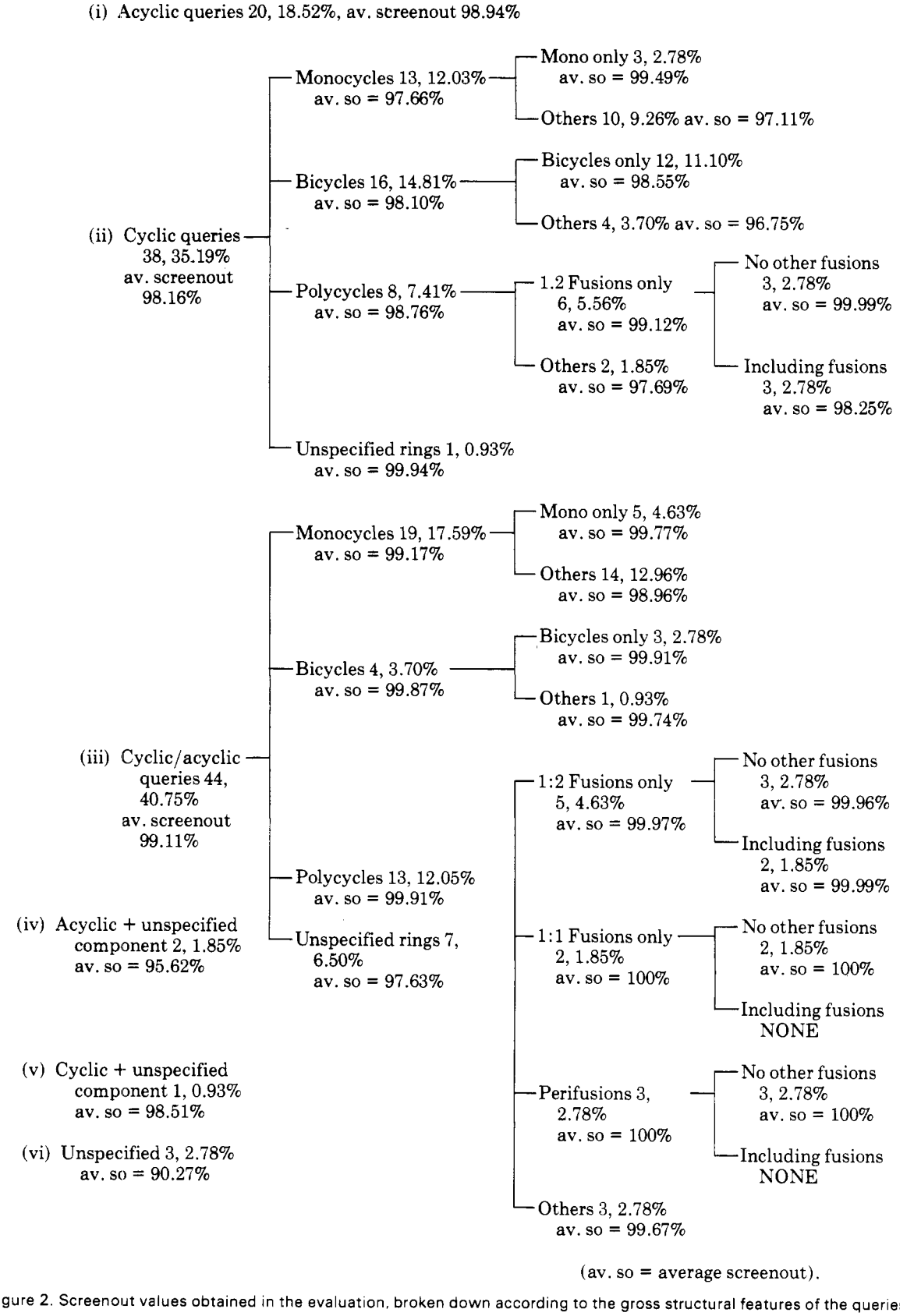


Figure 2. Screenout values obtained in the evaluation, broken down according to the gross structural features of the queries.

levels of fragment description, the actual levels permitted for each pair type are very useful, and the suitability of the existing pair set is well demonstrated by the observation that only in very few queries has the resolution of related substructures been prevented by the absence of appropriate fragments in the screen set. Only in very few cases did such restrictions prevent relationships between adjacent bonds in the substructure from being established. One example, that of the adjacent double bonds in allenic compounds, is illustrated below.

Substructure	Screenout
C—N=N—C	97.56%
—C=N=N	99.98%
C=N—N=C	97.46%
C=C=C	76.10%

This case could easily be catered for by a slight modification to the screen set. The highest level of description, the bonded pair, was particularly useful in minimizing the number of fused derivatives retrieved for cyclic queries, where fusions other than the ones indicated were not permitted. Bonded pair descriptions were also useful in the identification of ortho substituents. However, in the coding of certain cyclic queries where peripheral bond arrangements were of greater significance than the type of ring bond present, or where the degree of saturation was not precisely specified or was not specified at all, it was found that a generalized ring bond description would have been useful in reducing the need for 'OR' logic. Other fragment types have been introduced to fill the few and minor shortcomings of the present system as indicated by this evaluation. The determination of precision is an important step in the evaluation of a screening system, and we plan to report precision measures in the near future.

EXPERIMENTAL

The evaluation was carried out on an ICL 1907 computer with a core store of 24-bit words and a cycle time of approximately 2 μ s. The search program was written in PLAN, which is the ICL 1900 series assembly language, and runs in 16K of core store.

Queries were run in batch mode, and during processing, descriptions of the batch of queries were built up for display purposes. The screen search program used operates in three distinct phases. In the first, the query bit strings are set up, in the second the bit screen records of structures to be searched are read in from magnetic tape file and matched against the query bit strings. A sample of the answers from the screen search are retained in core for each query. In the final phase, the performance of the search for each query is summarized, and the registry numbers of a sample of answers for each query are displayed after sorting into numerical order.

CONCLUSIONS

The high average screenout value obtained for these SDI queries indicates the value of bond-centered fragments as screens in substructure searching. Screen generation[1] and screen search are very economical.

Bonded pairs span up to three bonds and two atoms, and as they are implemented may contain up to seven bonds and two atoms (bonded pairs are not generated for atoms which are connected to more than four non-hydrogen atoms). The results in Figure 2 show that the bonded pairs are effective in distinguishing 1:2 fusions from unfused or other fused systems. The bonded pair is also expected to be effective in identifying 1,2-disubstituted derivatives, although it will probably be less effective for 1,3- and 1,4-disubstituted derivatives. The performance of other fragment types is being tested and will be reported in the future.

Important characteristics of the present system are that the fragments are generated algorithmically without iterative search and transformed automatically for identification and handling into records in canonical form. The choice of which fragments are included in the screen set is made on the basis of frequency and can be made completely automatically without any reference to "chemical significance" and without any human participation in this process. Similarly, when a query is coded, the coder needs to know only the broad fragment types used in the system (e.g., simple pairs, etc.) and does not need to know the identity of individual species of a particular fragment which are present in the screen set. This eliminates the need for the query coder to refer to a fragment dictionary.

The high screenout recorded has been obtained using very small bond-centered structural fragments and crude ring descriptors which are cheap to generate and easy to use. Except in the case of a small number of very common structural features, only the presence or absence of a fragment in a structure is noted in the bit string, and multiple occurrences of the same fragment in a structure are not indicated. The results indicate that despite this simplification, very high screenout is obtained.

ACKNOWLEDGMENT

We thank the Office for Scientific and Technical Information (London) for financial support, Chemical Abstracts Service for providing the data base, and the Experimental Information Unit (Oxford) for supplying the user profiles.

LITERATURE CITED

(1) Adamson, G. W., Cowell, Jeannie, 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).

(2) 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. Acyclic Fragments," *J. Chem. Soc. (C)* **1970,** p. 990.

(3) Adamson, G. W., Clinch, V. A., and Lynch, M. F., "Relationship between Query and Data-base Microstructure in General Substructure Search Systems," in preparation, University of Sheffield, Sheffield, England.

(4) Milne, M., Plotkin, M., et al., "N.C.I. Screen Search Analysis and Recommendations," Contract NIH **71-2187,** University of Pennsylvania, Philadelphia, Pa., 1972.