Sampling Considerations in the Selection of Fragment Screens for Chemical Substructure Search Systems

MARIE T. GANNON and PETER WILLETT*

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield, S10 2TN, United Kingdom

Received January 17, 1979

Characteristics of the screen sets produced from subsets of a file of connection tables have been related to the size of the sample file used. The assignment of the sample derived screen sets to the file shows only a small decrease in coding efficiency when compared to the sets generated from the complete file.

INTRODUCTION

In recent years several studies have been made of systems for partitioning machine-readable structure files prior to atom-by-atom search by means of a screen set, i.e., a collection of file-dependent structural attributes. These investigations have considered the frequencies of occurrence and association of substructures and have resulted in the establishment of strict criteria by which a fragment may be judged for inclusion in a screen set.[1-4] While it is relatively simple to obtain the requisite fragment occurrence data from files containing moderate numbers of compounds, the screen set generation techniques cannot be applied directly to very large files owing to the vast numbers of fragment occurrences that would need to be cumulated. If the techniques are to be of general applicability, it is clear that sample files must be used for the generation of screen sets and that the effectiveness of such screen sets should not differ materially from that of one produced from the entire collection of compounds. To obtain an effective screen set one should ideally use a large sample, but processing constraints may make this impracticable; this paper describes an investigation of screen sets generated from samples of a file of 9643 connection tables derived from the CAS Registry System and attempts to identify the degradation in screen set effectiveness that might be expected when sets obtained from subsets of the file are used to characterize the whole of it.

COMPARISON AND ASSIGNMENT OF SCREEN SETS

The screen sets were generated from the file of connection tables using the algorithmic procedure described in an earlier paper.[4] The desired number of screens is selected from the many thousands of fragment types encountered in the structure file so that the final set of screens all occur with approximately equal probabilities and the information content of the set should be high.[2] The disparate frequencies of occurrence which are encountered are compensated for by a variable level of description with frequent fragments being described in some detail while infrequent features are given a generalized representation. In this study, atom-centered fragments were chosen for investigation, and the bonded atom, an atom together with the number and types of its adjacent bonds, was selected as the minimal level of description. Screen sets of sizes 60, 120, 180, and 240 members were generated for each (sub) file investigated; in each set, a single conflated screen was made available for assignment if no match could be obtained with any of the other members of the set. One in n (n = 5, 10, 25, 50, 75, and 100) interval samples of the file were used for screen set generation; in the subsequent discussion, a screen set will be referred to as, e.g., S25-60 if it contained 60 members and was generated from a 1 in 25 subfile. All five distinct S5 subfiles were used while for the S10 to S100 sets, eight samples were obtained for each using random starting points; however, only six S100-240 sets were created since in two subfiles there were insufficient fragments satisfying the frequency criteria for inclusion in the set.

The best way to evaluate a screen set is to assign it to a file of compounds and carry out a series of substructure searches so as to determine the average screenout: this is impracticable if many sets have to be assigned, as in this study. Instead, we have used the concept of a reference screen set, the characteristics of which may be compared with those of other sets. The obvious reference is the set produced from analysis of the complete structure file since the generation procedure will then have the maximum amount of occurrence information available upon which to base its selection of screens.

The sample-derived screen sets have been evaluated in two ways. Firstly, they have been compared with the corresponding sets obtained from analysis of the whole file to determine the numbers of screens in common. Thus, for some sample size 1 in x, each of the Sx-60 sets was matched against the S1-60 set, each of the Sx-120 sets with the S1-120 set, etc., and the number of screens common to both sets were noted. While giving a simple measure of the overlap in screen set membership, such an approach is subject to the criticism that, in the screen set generation procedure, several strings may have comparable occurrence frequencies and yet only one of them may be selected; if one of the strings is an S1 screen member, but is not chosen for inclusion in the sample set, the two sets will appear more dissimilar than would otherwise be the case even though the choice made might result in little change in the assignment characteristics of the sample set when taken as a whole. Consequently, the second evaluation considers the effectiveness of the sets in characterizing the structures in the file. A common method for assessing the degree of equifrequency of assignment of a set of symbols is the relative entropy of the assignment frequencies, H_r, which provides a rough measure of the coding efficiency of the set. Since the screens in a set have been selected so as to occur approximately equifrequently in their source file, variations in file characteristics may readily be identified by assigning the set to a different file and noting the change in the relative entropy of assignment. Accordingly, each of the sample sets was assigned to the complete file and the resultant entropy compared with that obtained from assigning the S1 set of the same size to the file; a similar approach has been used by Brack et al. in studies of the textual microstructure of bibliographic data.[5] When a set was assigned to the whole file, a count was made of the number of times each screen was assigned. If the ith screen from a set of size n was assigned a_i times, for i = 1, . . ., n, and if

$$N = \sum_{i=1}^{n} a_i$$

0095-2338/79/1619-0251$01.00/0

Table I. Mean Numbers of Screens in Common between Sx Sets and the Corresponding S1 Set of the Same Size[a]

subfile	60	120	180	240
S5	56.2 (3.6)	116.6 (1.1)	171.4 (3.1)	225.8 (3.6)
S10	54.8 (3.2)	114.1 (1.6)	167.0 (2.7)	222.4 (2.9)
S25	53.8 (3.1)	111.4 (2.4)	161.0 (2.7)	209.0 (3.0)
S50	51.3 (3.5)	106.5 (2.2)	149.6 (4.0)	194.4 (5.9)
S75	48.9 (3.5)	102.3 (4.6)	143.5 (4.0)	183.6 (6.0)
S100	49.0 (4.8)	98.0 (5.3)	139.0 (6.4)	179.8 (4.4)

[a] Standard deviations given in parentheses.

the total frequency of assignment, then the relative entropy, H_r, is given by

$$H_r = \frac{1}{\log n} \sum_{i=1}^{n} f(a_i)$$

where

$$f(a_i) = \begin{cases} 0 \text{ if } a_i = 0 \\ \frac{a_i}{N} \log \frac{a_i}{N} \quad \text{otherwise} \end{cases}$$

If the assignment is completely equifrequent, i.e., if $a_i = a_j$ for all i and j, then $H_r = 1$. In fact, the observed H_r values will be somewhat less owing to limitations in the generation algorithm and in the range of substructural types considered for inclusion in the screen set. Occasionally, a sample H_r value was found to be greater than that arising from the corresponding S1 set. Although the screen set generation procedure is both an efficient and an effective method for removing a large part of the uneven distribution of fragment frequencies, it does not produce an optimally equifrequent set of screens,[6] and thus discrepancies such as this are not unexpected; the largest difference was observed between the S1-60 set, with an H_r of 0.960, and an S5-60 and an S50-60 set, both of which had H_r values of 0.964.

RESULTS AND DISCUSSION

Details of the overlap in screen membership between sample-derived sets and the corresponding S1 set are given in Table I. For each set size, there is a steady decrease in the number of common fragments as the sample size is decreased; however, even in the case of the S100 subfiles, the sets exhibit a mean total membership of ca. 75% of the corresponding set based on the complete file. An inspection of the constituent screens in the various sets shows that the rate of decrease in commonality is much greater for the larger fragments. Crowe et al.[7] have presented figures showing that both the absolute frequencies of occurrence and the rate of decrease of frequency with rank (when the fragments of a given type, e.g., augmented atoms, are ranked in decreasing frequency order) fall off as the size of the fragment type increases; accordingly, the frequency differences between fragments of a given type are much less pronounced at the larger sizes and it is correspondingly less likely that the sample-derived sets will contain those large screens contained in the S1 sets. The file of connection tables used in this work was itself a 1 in 3 interval sample from a larger collection of 28 930 structures and the whole of this latter file was used to produce screen sets at the four sizes. These sets were then compared with each of the corresponding Sx sets to determine the degree of screen overlap. The results were very similar to those in Table I with, again, the S100 subfiles generating screen sets containing about three-quarters of the fragments that were chosen from the larger file, which in this case was 300 times as large.

Table II. Mean Relative Entropies of Assignment (H_r) for Screen Sets[a]

subfile	60	120	180	240
S1	0.960	0.965	0.954	0.950
S5	0.953 (0.019)	0.964 (0.001)	0.956 (0.001)	0.948 (0.004)
S10	0.945 (0.020)	0.963 (0.001)	0.954 (0.001)	0.948 (0.002)
S25	0.948 (0.020)	0.962 (0.002)	0.951 (0.003)	0.942 (0.003)
S50	0.938 (0.019)	0.960 (0.003)	0.945 (0.005)	0.934 (0.005)
S75	0.928 (0.019)	0.950 (0.020)	0.939 (0.004)	0.928 (0.010)
S100	0.929 (0.022)	0.936 (0.021)	0.931 (0.018)	0.922 (0.010)

[a] Standard deviations given in parentheses.

As noted above, a straight comparison of screen set membership merely counts the number of fragments in common without regard to the degree of equifrequency of assignment when a set is used to characterize the complete file of compounds. It would, however, be expected that variations in screen set composition would be fairly directly reflected in the observed relative entropies and the figures in Table II show that, for a given set size, H_r does indeed decrease, and the standard deviations increase, as the sample size is reduced. The fall-off in H_r is most pronounced for the 60-member sets and the standard deviations for this group tend to be high even when quite a large sample is being tested, e.g., the S5-60 values; a possible reason for this may be that any incorrectly selected strings, i.e., strings not included in the S1 set, will have a proportionately larger effect on the equifrequency of assignment due to the limited range of alternative screens. In the case of the 120-member sets, H_r remains almost constant until the S75 level, i.e., using a 1 in 75 sample, and then drops away; comparable results are seen for the 180- and 240-member sets at about the S50 and S25 levels, respectively.

The relative entropy of the bonded atoms identified in the complete file was 0.798; this figure is relatively high since the majority of the noncommon atoms had been allocated a common first level descriptor, this causing a considerable decrease in fragment variety.[4] Thus even the lowest H_r values indicate a considerable increase in coding efficiency when compared with that obtainable from a single level of description without the use of the selection procedure; in view of the very small numbers of compounds in the sample source files, e.g., 96 in the case of S100 sets, the improvement is quite striking.

When selecting fragments for inclusion in a screen set, care must be taken to ensure that the screens will occur with as even a distribution as possible when they are assigned to the structures in a file so as to allow adequate discrimination for a wide range of substructural queries. The selection procedure involves the cumulation of large numbers of fragment incidences and will thus make heavy demands upon computer processing facilities unless a subset of the structures is analyzed to identify the appropriate fragments. Feldman and Hodes used a 10% sample file for the development of a screening system[2] and Graf et al. have also reported the use of a sample file.[8] The work described here has shown that quite small files are sufficient to identify many of the screens that would be selected from analysis of a much larger collection of compounds. As well as possessing many such screens, the sample-derived sets have also been shown to exhibit a high degree of equifrequency when they are used to characterize the entire file of structures. The decreasing skewness of the Zipfian distribution of fragment occurrences as the fragment size is increased means that it will become more difficult to identify the correct fragments (i.e., those chosen on analysis of the whole file) for larger screen sets; however, the decreased skewness also implies that incorrectly chosen fragments are likely to have comparable frequencies to the correct ones and thus have a less marked effect on the assignment characteristics

of the set.

CONCLUSIONS

A study has been made of the screen sets produced from subsets of a file of connection tables. While there is a close relationship between the size of file used for screen set generation and the characteristics of the resultant sets, sets derived from quite small subfiles compare not unfavorably with those based upon the whole file. Accordingly, when designing a screening system for chemical structure searching, it is sufficient to base fragment selection procedures upon the frequencies obtained from quite limited subsets of the file of compounds that is to be screened.

ACKNOWLEDGMENT

We thank Robert Kay, David Cooper, Michael Lynch, and the referees for helpful advice, the Department of Education and Science for the award of a British Library Postdoctoral Research Fellowship to P.W., and the Institute for Industrial Research and Standards, Dublin, for funding M.T.G. We also thank the operating staff of the University of Sheffield Computing Services Department for their cooperation in the handling of the large number of computer runs used in this study and Chemical Abstracts Service for the provision of the structure file.

REFERENCES AND NOTES

(1) M. F. Lynch, "Screening Large Chemical Files" in J. E. Ash and E. Hyde Eds., "Chemical Information Systems", Chichester, Ellis Horwood, 1975.
(2) A. Feldman and L. Hodes, "An Efficient Design for Chemical Structure Searching. I. The Screens", *J. Chem. Inf. Comput. Sci.*, **15**, 147–152 (1975).
(3) L. Hodes, "Selection of Descriptors According to Discrimination and Redundancy. Application to Chemical Structure Searching", *J. Chem. Inf. Comput. Sci.*, **16**, 88–93 (1976).
(4) P. Willett, "A Screen Set Generation Algorithm", *J. Chem. Inf. Comput. Sci.*, **19**, 159–162 (1979).
(5) E. V. Brack, D. Cooper, and M. F. Lynch, "The Stability of Symbol Sets Produced by Variety Generation from Bibliographical Data", *Program*, **12** (2), 61–74 (1978).
(6) P. W. Williams, "Criteria for Choosing Subsets to Obtain Maximum Relative Entropy", *Comput. J.*, **21** (1), 57–62 (1978).
(7) J. E. Crowe, M. F. Lynch, and W. G. Town, "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part I. Non-cyclic Fragments", *J. Chem. Soc. C*, 990–996 (1970).
(8) W. Graf, H. K. Kaindl, H. Kries, B. Schmidt, and R. Warszawski, "Substructure Retrieval by Means of the BASIC Fragment Search Dictionary Based on the Chemical Abstracts Service Chemical Registry III System", *J. Chem. Inf. Comput. Sci.*, **19**, 51–55 (1979).

The Effect of Screen Set Size on Retrieval from Chemical Substructure Search Systems

PETER WILLETT

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield, S10 2TN, United Kingdom

Received March 19, 1979

Both atom- and bond-centered screen sets containing between 120 and 960 members have been used to characterize a file of 28 790 structures. Although the resolving power of the fragment bitstrings for substructure search increases with screen set size, improvements in retrieval performance above a certain level are likely to be gained only at the expense of a large increase in the number of screens or of alternative bases for screen selection.

Efficient searching of large files of chemical compounds is made possible by the use of screens, that is, small substructural fragments, the presence or absence of which is used to identify those few cases where full atom-by-atom search is required.[1] Many structure search systems use a sequential file organization in which queries are matched against each of the structures in the file in turn, the set of screens associated with a structure being represented by a bitstring which may be very rapidly compared with analogous strings describing the query requirements.[2-4] An important factor in the speed of operation of such systems is the number of screens which are available for assignment to the structures in the file and to the queries that are applied to it. Use of a small set of screens means that bitstring matching will be very fast but that many molecules may satisfy the query requirements, thus necessitating a large amount of iterative searching; conversely, the greater specificity of a large screen set will eliminate a greater number of nonrelevant structures at the cost of more bitstring matching and increased file creation times. Methods for the selection of fragment screens have been given by Lynch,[1] Hodes,[5] Feldman and Hodes,[6] and Willett.[7] The last procedure permits the generation of screen sets of any desired size, and this flexibility is used here to investigate the relationship between screen set size and bitstring discrimination for a set of substructural searches.

The file of structures used in this work contained 28 790 compounds drawn at random from the Chemical Abstracts Service Registry System. The connection tables of the compounds were analyzed to produce both atom- and bond-centered screen sets of sizes 120, 240, 480, 720, and 960 members, using the screen set generation procedure described earlier.[7] In this, atom- or bond-centered circular chemical substructures are characterized by strings of integers in which the first integer represents either a bonded atom or a simple pair, and subsequent integers give an increasingly detailed representation of the immediate environment of the central feature.

For each screen set an analysis was made of the connection table of each compound in the file so as to produce integer strings up to the maximum level of substructural description present in the screen set; each of the strings obtained from the table was then searched against the screen set. If a match was found for a string with one of the screens the appropriate bit was set in the bitstring; if not, the string was shortened by one integer and the set searched again. A conflated screen was available for assignment if a match could not be achieved with any of the members of the set even at the single integer level. Once a screen had been assigned, all of the smaller fragments contained within it were automatically allocated to the structure as well, thus removing the need for Boolean OR logic

0095-2338/79/1619-0253$01.00/0