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

Table I. Median Numbers of Structures Retrieved Averaged over the Set of Queries

screen set size	atom-centered	bond-centered
120	643	697
240	243	281
480	117	145
720	109	76
960	96	60

Table II. Numbers of Unique Representations When the 28 790 Structures Were Sorted into Ascending Order of Their Fragment Bitstrings

screen set size	atom-centered	bond-centered
120	25 676	24 732
240	26 862	26 245
480	27 300	26 863
720	27 477	27 104
960	27 530	27 193

between many specific screens when carrying out a generic search using a large screen set.[4] The queries were input to the search program as redundant connection tables and these were used to produce integer strings for screen assignment as with the compounds in the structure file; however, only the most specific screen matched with each string needed to be represented in the query bitstring. Unspecified connections in the query structure were filled by dummy atoms.[7]

The effect of fragment variety on substructure search performance was investigated using a set of 45 queries drawn from user profiles supplied by the Experimental Information Unit at Oxford. The chosen queries did not include generic searches for specific rings or ring systems since a systematic evaluation of such queries would necessitate the use of ring-centered screen sets in addition to the atom- and bond-centered fragments used here. In addition, since the study considers only the effect of controlled variations in screen set size on screenout, detailed atom, bond, ring, and connectivity counts, such as are described by Graf et al.,[2] were not included in the fragment bitstrings. In a previous investigation using these queries, Adamson et al. found that the distribution of retrieval set sizes was extremely skewed;[4] also a poorly screened query may yield a very large output which leads to distortions in the average performance if the mean is used as an overall measure of retrieval effectiveness. Accordingly, the median number of structures retrieved, averaged over all of the queries, was used to evaluate the substructure searches, and the appropriate figures for each screen set size are shown in Table I. For both types of fragment considered, there is an initial rapid decrease in the number of structures retrieved followed by a levelling off as the set size is increased; i.e., the screenout tends toward a maximum value. Comparable results are obtained if registration searches are considered; in this case the bitstrings of each size were sorted so as to determine the number of unique representations present in the file and these are shown in Table II. Thus, above some screen set size, an increase in the number of screens available for assignment, based, at least, on the controlled approach invoked here, can add little to the discriminatory power of the screens already used for characterizing the structures in a file.

DISCUSSION

As the fragments produced by an iterative fragmentation procedure grow in size, there is a large increase in the variety, i.e., in the number of fragment types identified, and a decrease in the average incidence of the fragments. The addition of large fragments to a screen set will hence lead to an increase in the selectivity of the set. Use of a large set, which will contain many big fragments, means that a wide range of specific screens will be available for assignment to the compounds in a file. Accordingly, when several screens are used to characterize a query, large variations in the numbers of structures retrieved may be expected if screen sets of varying size are used. However, the tendency to increased screenout with increased screen set size will be lessened by at least three factors. Firstly, by the variation in incidence of the screens. If infrequently occurring fragments are added to a screen set, the increase in discrimination will be marginal since such screens are most unlikely to be required for characterizing a query. Related to this is the fact that although screens are chosen so as to reflect frequencies of occurrence in the structure file, this is done so as to predict frequencies in the queries that will be put to the system. If many of the queries are highly generic in character, the large, highly discriminating screens in a set may be assigned much less frequently than would be expected from their incidence in the file; in such a case, the wider range of fragments available in a large screen set would again add little to the resolving power of the set as a whole. Adamson et al.[8] found a reasonable level of agreement between the frequencies of assignment to structures and to queries but not only were the fragments studied quite small, consisting of simple, augmented and bonded pairs, but also the agreement was noticeably closer for the more specific queries tested. Thirdly, screenout performance cannot be directly calculated from incidence data since it has been shown that screen set assignment frequencies are not independent of each other. One type of association common to iterative fragmentation procedures, that between a fragment and its parent, has been studied by Hodes,[5] but associations also exist between apparently unrelated screens. Thus an analysis of the coassignment frequencies of pairs of screens showed that the associations increased with fragment size and that certain of them might prove to have a considerable effect on screenout.[1]

Many of these restrictions may be alleviated if the appropriate fragments are chosen for inclusion in a set. Thus the variant incidences may be in large part eliminated by the exclusion of infrequent fragments so that the relative entropy of the set, i.e., the degree of equifrequency of the screens, is high;[7] in an alternative approach, variations in the number of bits assigned to each screen in a superimposed coding system were used to compensate for the differing frequencies.[6] The hierarchical associations between fragments may be removed by the exclusion of filial screens which occur only slightly less frequently than their parents since the presence of the more specific screens will add little to the selectivity of the more generic fragment. Again, the selection procedure used here is slightly biased toward the selection of smaller, more generic substructures, so that at least some of the excessively specific screens implied by the second point above are not present. Accordingly, attention was focussed on the effect of interfragment associations. While negative associations have been observed,[1] the majority are found to be positive so that, in general, the number of structures retrieved by a query will tend to be greater than that predicted upon the basis of the incidences of the query screens.

To test the effect of fragment associations, experiments were carried out using the 17 substructures which required only a single connection table as the input query representative. Whole file incidences were obtained for each of the screens in each of the screen sets. Then, for each of the queries, the incidences of the assigned screens were multiplied together to give an expected value for the number of structures retrieved by the use of each screen set. For both the atom- and the bond-centered sets, a comparison between the observed and the expected numbers of structures shows that the agreement between the two figures does indeed decrease as the screen set size becomes greater. Thus while large screens are essential

for high screenout in many types of query,[4] their general utility may be somewhat less than expected. While the actual screenout figures obtained here are specific to the particular file, queries and screen set selection procedure used, analogous results would seem to be applicable to any procedure which assigns screens to each and every bond in a structure. Accordingly, taking the other points above into consideration, significant increases in screenout performance above some point are unlikely to be gained unless the number of screens available for assignment is considerably enlarged, or alternative types of descriptor are used.

CONCLUSIONS

For the set of substructural queries and the screen selection procedure used here, there is a noticeable tradeoff between the number of screens available for assignment to a structure file and the resolving power of the screen sets. Although discrimination increases with increasing screen set size, improvements in retrieval effectiveness above a certain point 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. The exact point at which the marginal increase in discrimination is outweighed by increased storage and search times will depend on, inter alia, the nature of the fragments, the size of the file, and the efficiency of the iterative search algorithm used for exact matching as well as computer hardware limitations.

ACKNOWLEDGMENT

Thanks are due to David Cooper and Michael Lynch for valuable discussions, to Chemical Abstracts Service for the provision of the structure file and to the Department of Education and Science for the award of a British Library Postdoctoral Research Fellowship.

REFERENCES AND NOTES

(1) M. F. Lynch, "Screening Large Chemical Files" in J. E. Ash and E. Hyde, Eds., "Chemical Information Systems", Ellis Horwood, Chichester, 1975.
(2) 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).
(3) J. F. B. Rowland and M. A. Veal, "Structure-Text and Nomenclature Text Searching for Chemical Information: an Experiment with the Chemical Abstracts Integrated Subject File and Registry System", *J. Chem. Inf. Comput. Sci.*, **17**, 81–89 (1977).
(4) G. W. Adamson, J. A. Bush, A. H. W. McLure, and M. F. Lynch, "An Evaluation of a Substructure Search Screen System Based on Bond-Centered Fragments", *J. Chem. Doc.*, **14**, 44–48 (1974).
(5) L. Hodes, "Selection of Descriptors According to Discrimination and Redundancy. Application to Chemical Structure Searching", *J. Chem. Inf. Comput. Sci.*, **16**, 88–93 (1976).
(6) A. Feldman and L. Hodes, "An Efficient Design for Chemical Structure Searching. I. The Screens", *J. Chem. Inf. Comput. Sci.*, **15**, 147–152 (1975).
(7) P. Willett, "A Screen Set Generation Algorithm", *J. Chem. Inf. Comput. Sci.*, **19**, 159–162 (1979).
(8) G. W. Adamson, V. A. Clinch, and M. F. Lynch, "Relationship between Query and Data-Base Microstructure in General Substructure Search Systems", *J. Chem. Doc.*, **13**, 133–136 (1973).

Fisher Discriminant Functions for a Multilevel Mass Spectral Filter Network

G. T. RASMUSSEN, G. L. RITTER, S. R. LOWRY, and T. L. ISENHOUR*
Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27514

Received February 20, 1979

Fisher linear discriminants are described and applied to classification problems using mass spectral data. A two-level network of discriminants is used to improve classification. These discriminants provide a useful basis for two-dimensional projections of multidimensional patterns.

In many experimental problems the results are displayed in two-dimensional graphs, which are frequently plots of two measured physical or chemical properties. In the context of classification problems, some obscure property is to be determined on the basis of the information in the two-dimensional graphs. Often more than two properties are measured and all information cannot be contained in a two-dimensional form. This paper presents a novel and useful method of projecting multidimensional data onto two dimensions in a way that maintains discriminating information. The method relies on Fisher linear discriminant functions used in a two-level filter network. It is particularly amenable to use in interactive pattern recognition systems having graphic displays, such as those systems described by Sammon and by Koskinen and Kowalski.[1-3] The availability of such low-dimensional graphs of high-dimensional data allows the chemist to assume a greater role in the interpretation of the data.

The classification of organic compounds by using mass spectral data is the example selected to illustrate the Fisher ratio method. The general problem is to determine the presence or absence of specific molecular substructures in organic compounds from mathematical analysis of the low resolution mass spectra of the compounds. Each mass position represents a separate dimension and the measured property is the intensity of the peak at each mass position. One hopes to discriminate between compounds which do or do not contain a specific structural feature by applying a pattern recognition method. In the past, a variety of pattern recognition techniques have been applied to classification problems using mass spectral data. Linear learning machines were used in early studies.[4-7] Adaptive digital learning networks, simplex methods, and progressive filter networks have also been used.[8-10] Recently some of these methods and others, including *k*-nearest neighbor methods and Bayesian discriminant analysis, have been reviewed and compared.[11,12] Linear discriminants which maximize the Fisher ratio offer a useful complement to these methods.

THEORY

In applying a Fisher linear discriminant, the data set is first divided into two discrete categories or classes. One then attempts to find the direction in the multidimensional space defined by the measured properties such that data points projected onto a line in this direction will be maximally discriminated according to the selected classes. The criterion for discrimination is the Fisher ratio. For a single measured property, the Fisher ratio is a number equal to the square of

0095-2338/79/1619-0255$01.00/0