Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files

GEORGE W. ADAMSON, JEANNE COWELL, MICHAEL F. LYNCH,*
ALICE H. W. McLURE, WILLIAM G. TOWN, and A. MARGARET YAPP
Postgraduate School of Librarianship and Information Science,
University of Sheffield, Western Bank, Sheffield S10 2TN, England

Received January 23, 1973

A major problem in the design of screening systems for substructure searches of chemical structure files is the development of a methodology for selection of an optimal set of structural characteristics to act as screens. The set chosen for a particular application will depend on the characteristics of the collection, as well as on its size and growth rate. A strategy which takes account of the disparate frequencies of the various species of fragments in a data-base by use of differential, and, in part, hierarchical levels of description is detailed. The distributions of a variety of structural characteristics, including bond-centered, atom-centered, and ring fragments in a 30,000-compound sample of the Chemical Abstracts Service Registry System are summarized. Implementation of the approach, using primarily bond-centered fragments, by means of simple and highly efficient computer programs, is detailed.

The need to provide flexible and economic searches of chemical structure files to fulfil chemists' requirements for substructure searching within more general chemical information systems poses complex problems with interesting implications both practical and theoretical in nature. Many approaches have been advocated,[1] embodying a variety of viewpoints. In no respect has opinion been more varied than in the design of screening systems. These entail the selection of structural characteristics on the basis of which an approximate match between queries and potential answers is made. This stage may be followed by a more detailed search involving atom-bond-atom path tracing. The adequacy of the selection of characteristics on the basis of which the collection is indexed is critical both to the extent to which the system can fulfil the variety of queries addressed to it and to the over-all costs of searching.

The work reported in this paper arose from the conviction that it was essential to develop a general methodology for the design of screening systems, which could then be applied with equal validity to collections differing widely both in size and composition. (The need for such a methodology is borne out by even a cursory examination of the diversity of conventional fragmentation codes,[2] which generally reflect both of these factors. Thus a system devised for an alkaloid file will place heavy emphasis on ring-system skeletons and on the environments of nitrogen atoms, whereas a code devised for a large collection will, of necessity, be more specific and contain a greater number of characteristics than that for a small file.) In terms of size, therefore, the assumption was made that a greater level of selectivity is required in searches of larger files than in smaller ones; if a constant proportion of structures were retrieved, searches of large files might result in impractical numbers of structures being retrieved. In terms of composition, it was assumed that the queries addressed to a collection would roughly mirror the characteristics of the file; this is again borne out by experience with fragmentation codes,[3] and by recent work in Sheffield.[4]

The principal technical problems to be faced are the total number of structures known, and their great variety which, as Lederberg[5] has demonstrated, is far from exhausted. The variety of structural queries that chemists may express may well be no less varied. At the same time, the distribution of structural characteristics, however these may be defined, is extremely disparate in terms of the frequencies of their occurrence. This is at once both disadvantageous and advantageous. It represents a disadvantage in that a great variety of characteristics may need to be taken into account to allow for the variety of possible enquiries, yet it is also an advantage in that the disparate distribution enables us to describe the characteristics at differing levels of description. While it is possible to envisage a screening system so comprehensive that a search at this level would attain a precision ratio of 100%, it would be enormously costly both in the generation of the fragments and in their storage and search. What must be sought, therefore, is a balance between the variety of characteristics to be identified at the outset, and the costs of generating and searching them. A balance must be sought too, between the initial costs of screen generation and the costs at run time, although the former is a one-time cost. The balance finally decided on will be a system implementer's decision; it may involve either relatively low initial screen generation costs, with higher running costs, or a higher initial investment, to be amortized by more frequent and less expensive individual searches.

The general philosophy underlying the present work is that account must be taken of the disparate frequency of characteristics in chemical structures by employing different levels of description. This entails the description of frequent characteristics at a substantial level of detail, while those that are less common are described in more general terms. In no sense is this notion novel. It has been applied in the design of manual fragmentation systems, nonmenclature, and indeed, in indexing too, but its application in the context of computer-generated screening systems is the exception rather than the rule. However it may also be necessary to describe frequent characteristics at a general level as well, to provide for easy query encoding. Where possible, this should be performed within a hierarchical screen structure.

The validity of this approach is borne out by even intui-

*To whom correspondence should be addressed.

tive applications of information theory; Mooers[6] for example, articulated certain ideal requirements for indexing systems as long ago as 1951, stating that the index terms should be independent of one another, and should be present in 50% of the items and absent in the others. Given these circumstances, the combination of ten terms should discriminate a single structure among a thousand, since $2^{10} = 1024$. This will apply to a search for structural identity carried out at the screen search level. However, this simple reasoning must be modified in the case of substructure searches, where the number of screening fragments in queries will generally be less than the number in the structures on file fulfilling the requirements, and so the performance of screens will not reach this ideal. Also it is found in practice that structural characteristics are not independent of one another,[7] nor do they occur with equal frequencies, and one must be content with less than ideal solutions.

The implications of the theory are that the screens should occur with as even a distribution as possible. That this is not the case with a variety of characteristics of a large general collection (based on a sample of almost 30,000 structures from the Chemical Abstracts Registry System) has been amply demonstrated in preceding studies.[8-10] An extreme case is that of the distribution of elements in the sample file. In terms of the total number of atoms, carbon accounts for 79%, and oxygen and nitrogen for 14% and 7%, respectively. These three elements thus account for almost 95% of the total number of atoms. After these, the approximate rank order is fluorine, sulfur, chlorine, phosphorus, bromine, silicon, and iodine, with the last accounting for less than 0.1% of the total. This is a Zipfian distribution, with a frequency ratio of almost 1000:1 between the first and tenth ranking elements. Similar, if less extreme, distributions have been observed for other characteristics.

These figures imply that the values of elements used alone as screens are disparate. A search for compounds containing iodine will be highly selective if iodine is selected as a screen, but the possibility of such a search being requested is low compared with those directed towards substructures containing carbon, oxygen, or nitrogen. Similar considerations apply to fragments consisting of groups of atoms and bonds. If these are chosen on the basis of the occurrence of functional groups such as carboxylic acids, esters, amines, aromatic nitro-groups, etc., then these too will show very disparate distributions, and are likely to vary equally in utility. These problems have been discussed by Meyer[11] in relation to the GREMAS code, and by Craig[3] with reference to the SK&F fragmentation code.

The basic dilemma in screening system design is that selection of fragment types must be made in such a way that neither too great a variety of screens—i.e., their total number—nor too extreme a frequency variation results from the selection. In addition, the fragment types should be easily amenable to automatic generation.

This problem is not confined to chemical structure retrieval systems; it is also encountered in subject description for indexing, as well as in text-search systems. In the context of subject indexing within a particular discipline, the indexer describes the most frequent topics in greater detail than topics which are peripheral to the subject area. This factor is especially evident in the case of articulated subject indexes, in which the indexer records statements of subject content in a number of descriptive phrases.[12] Similar considerations apply with use of structured thesauri. This modulation of the description of subject content is still very much an intellectual process. Again, chemical substructure searches and natural-language text searches are basically searches for substrings. In the case of text-searches, the substrings searched for are words, word sequences, or word fragments; in the case of structures, the substrings may be either one or two-dimensional, and additionally, varying degrees of relaxation of substring definition may be allowed, as when a bond or atom may be chosen from several alternatives.

A useful generalization, which is applicable to file-organization of both structural and textual data-bases for substring search, and is amenable to computer use, can be stated at this point. Given records in which a variety of individual symbols occur with variant frequency (atoms and bonds in the structure case, alphanumeric symbols in the case of text), symbol aggregates of different sizes can be produced by concatenation of symbol strings. If these aggregates are chosen so as to reflect the relative frequencies of the constituent symbols, sets of characteristics which are larger in number than the original symbol set may be produced, with, however, mucn less disparate frequency distributions. These characteristics are also more highly discriminant in search than the original set, or uniformly-produced aggregates thereof. That this generalization applies in the context of searches of textual databases has recently been demonstrated;[13] it also forms the basis of the present work.

FRAGMENT SELECTION

There are two primary centers in molecules which can form the foci for fragment generation—atoms and bonds. (The ring may be considered as a third and higher-level focus.) The foci differ considerably in the characteristics shown by fragments centered on each of them, particularly in regard to their variety and distribution as successive stages of their immediate environment are taken into account. The atom-centered fragment shows the progression: atom, coordinated atom, bonded atom, and augmented atom. Figure 1 shows examples of these species. In terms of variety, the sample studied (28,963 structures from the Chemical Abstracts Service Registry System) showed 68 atom types, 132 coordinated atom types, 313 bonded atoms, and 2331 augmented atom types.[9] In terms of distribution, 960 of the 2331 augmented atoms each occurred in a single structure. (The figures relate to fragment types in which the bonds are differentiated as cyclic or acyclic.) Thus, whereas the increases in variety from atom to coordinated atom and bonded atom types are relatively slight, the increase to the next stage of detail is very substantial.

The bond-centered fragments show the progression instanced in Figure 2. (The terminology used here does not match that of the atom-centered fragments, but each usage is already entrenched in the literature). Here the progression from the bond, through the simple and aug-

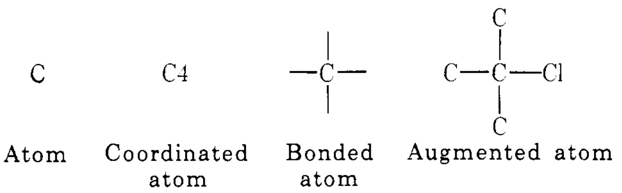

Figure 1. Atom-centered fragments

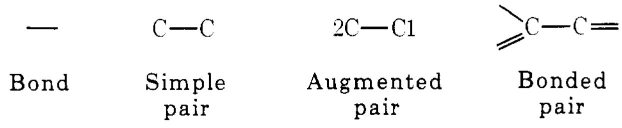

Figure 2. Bond-centered fragments

mented pairs to the bonded pair is made in more gradual, less abrupt stages. The variety of simple pairs, for example, was found to be 393 (again with differentiated bonds). Increasing the extent of the immediate environment described to include the number of bonds external to the simple pair resulted in identification of 841 differentiated augmented pairs; spelling out the nature (bond order, and cyclic/acyclic nature) of these bonds resulted in production of 1733 differentiated bonded pairs. These figures, when compared with those for the atom-centered fragments, illustrate that the increase in fragment variety for bond-centered fragments is a more even process.

The bond-centered fragments also form a natural hierarchy. This was seen to offer a substantial advantage over the augmented atoms, not only by their providing the possibility of selecting a series of levels of detail at which characteristics of differing frequency could be described, but also in terms of generic search capability, simplifying the process of query encoding. Moreover, the ease of incorporating the decisions as to the level of detail appropriate to a given bond-centered fragment into simple programs operating at high speeds made it imperative to compare a screening system based primarily on the bond-centered fragments with the most widely used alternative, the augmented atom, as used in the CAS system, the GREMAS system, and at the University of Georgia.

Thus far, the rationale of the approach to a file-sensitive substructure search screening system has been described. The paper now details the screen types included at the present time, with timings for their generation. Subsequent papers in this series describe the evaluation of their performance, a comparison of the characteristics of queries with those of the file, and a detailed analysis of the structure of the bit screens utilized in the search system, as well as the inclusion of further fragment types.

Although the implementation of the system thus far is tape-based, the arguments adduced apply equally to inverted file systems.

SCREEN CHARACTERISTICS

The contents of the screen set are shown in Table I. The screens are held on magnetic tape, and as the computer used is an ICL 1907 with 24-bit words the addresses in the screen records are given in terms of words and bits.

In words 4, 5, and 6 counts of common and important structural features are stored. The maximum values of the counts are shown in the Table I. The counts are used as binary numbers in the system, and several counts are packed in each of these three words. The maximum number which can be represented in a count field is related approximately to the density of occurrence of the fragment.[8] If a structure contains more fragments of a particular type than can be stored in the appropriate field, then the count in that field is set to its maximum value. The use of count fields rather than bit strings in this part of the screen record saves considerable space.

The remainder of the screen set is in the form of a bit string and contains indications of the presence or absence of certain atoms, pairs, and ring types.

In the atom bit string, there is one bit for each of the 68 atom types which occur in the file studied and one bit for all atom types which do not occur in this file. The remaining three bits in the atom bit string are used to indicate whether or not bits are set further along the string. If further bits are not set then a search would be terminated at that point.

The next 31 words are used to indicate the presence or absence of pair types. The most common pairs lead to a bit being set in the bonded, augmented, and simple pair fields of the bit string; less commonly occurring pairs are indicated in the augmented and simple pair fields; and rarely occurring pairs have their presence or absence indicated only in the simple pair field.

Table I. General Description of the Screen Record

Words	Bits	Contents	
1	0–23	Record length (= 46 words)	
2–3		CAS Registry number of compound structure	
		COUNTS	
4	0– 5	Total atoms	max. value 63
	6–11	Carbon atoms	max. value 63
	12–17	Oxygen atoms	max. value 63
	18–23	Nitrogen atoms	max. value 63
5	0– 3	Ring	max. value 15
	4– 7	Degree of connectivity = 3	max. value 15
	8–11	Fluorine atoms	max. value 15
	12–14	Sulphur atoms	max. value 7
	15–17	Chlorine atoms	max. value 7
	18–20	Degree of connectivity = 4	max. value 7
	21–23	Degree of connectivity = 5	max. value 7
6	0– 4	Single bonds (chain)	max. value 31
	5– 9	Single bonds (ring)	max. value 31
	10–14	Aromatic bonds	max. value 31
	15–17	Double bond (chain)	max. value 7
	18–20	Double bond (ring)	max. value 7
	21–23	Triple bonds	max. value 7
		BIT STRINGS	
7– 9	0–23	Atom bit string	
10–27 28	0–23 0– 1	Bonded pair bit string	
28 29–31	2–23 0–23	Simple pair bit string	
32–40	0–23	Augmented pair bit string	
41–45	0–23	Ring bit string	
46	0–23	Unused	

Table II. Simple Pair Types Included in the Bit String

C	C	
C	O	
C	N	
C	S	
C	F	
N	O	
C	Cl	
O	S	
O	P	
N	N	
C	Si	
C	Br	
C	P	
N	S	
C	I	
N	P	
O	Si	
O	O	
X	X	(for all pairs other than those given above—i.e., exception pair bits)

SIMPLE PAIRS

Each of 18 of the most common pairs of atoms (see Table II) have fivebit subfields in the simple pair field, each bit corresponding to a bond type between the atoms of the pair. The bond types allowed are: single bond (chain), double bond, delocalized bond (ring), triple bond, N

	C—C	0,	1,	2,	3
	0,				
M	1,			X	
	2,				
	3				

Figure 3. Implicit table for setting C—C augmented pair bits M = Number of external non-H atoms bonded to the first C atom N = Number of external non-H atoms bonded to the second C atom

and single bond (ring). Four bits are available for all the pair types which do not have individual subfields.

AUGMENTED PAIRS

The process for setting the appropriate bit for pairs whose presence is indicated at the augmented pair level involves the use of implicit two-dimensional tables. The tables do not exist as such in the program. The process is illustrated in Figure 3 for the case of augmented pairs containing the C—C simple pair.

The position X in the table corresponds to the augmented pair 1C—C2. A subfield in the augmented pair field is for augmented pairs which contain the C—C simple pair. The places in the table are counted across the rows; 1C—C2 is in the 7th position. The position of the bit corresponding to 1C—C2 is then the 7th bit of the subfield for C—C. Table III shows the simple pair types which define the inclusion of augmented pair types.

BONDED PAIRS

The setting of bits showing the presence or absence of bonded pairs is also based on the use of implicit two-dimensional tables. The pairs which are described at the level of bonded pairs are those CC, CN, and CO pairs which contain the simple pairs shown in Table IVa. The permitted bond patterns are shown in Table IVb. For example, if the bonded pair ·C·N· occurs in a structure, then the bonded pair contains the simple pair C·N. The permitted bond patterns for C· and N· are listed in Table IVb. This is a table with 11 rows and 4 columns and ·C·N· will be found in the second column of the first row. The bit for ·C·N· will therefore be the 2nd bit of the C·N bonded pair subfield.

The use of implicit tables results in a very fast screen generation and a small program, but some of the bit positions are never used as they correspond to pairs with noncanonical records. The bit string length could be reduced by about 40% without any reduction in the screenout, recall, or precision, and with only a very slight increase in screen generation time and program size. However, in this work the bits in the pairs screen which are not used by pairs will be used for further fragment types.

Figure 4 illustrates the general procedure for generating screens records. The first stage in the process, at pair level, is the determination of the canonical form of each pair in the structure. If the pair is other than CC, CN, or CO, it is not included in the bit string at augmented or bonded pair levels. It is therefore checked against the simple pair list and, if found, the appropriate bit in the bit string is set. If the pair is not found in the simple pair list, then a bit for the exception pair XX is set. If the pair is a CC, CN, or CO pair, then an attempt is made to set the pair in the bonded pair bit string. If the pair is of high incidence then this will succeed. Whether or not the presence of the pair is indicated as a bonded pair an attempt is next made to indicate its presence as an augmented pair by comparing it with the appropriate augmented pair table. The pair is then compared with the simple pair list and its presence indicated as a simple pair.

Table III. Augmented Pair Types Included in Bit String (Bond symbols are explained in ref. 9)

C—C
C═C & C:C
C*C
C≡C & C⋮C
C · C
C—O
C═O
C≡O
C · O
C—N
C═N & C:N
C * N
C≡N & C⋮N
C · N

Table IVa. Pairs Represented as Bonded Pairs Contain Only the Above Simple Pair Types

C—C	C—O	C—N
C═C	C═O	
C · C	C · O	C · N
C : C		
C * C		C * N

Table IVb. Atom and Bond Types Included in the Bonded Pairs

Atom Type and Central Bond Type	Bond Patterns Included in Bonded Pairs
C—	$-C-$ $-C$ $-\overset{\vert}{C}=$ $-\overset{*}{C}*$ $-\overset{\vert}{C}-$ $-\dot{C}\cdot$ $-\dot{C}:$ $-\underset{\cdot}{\dot{C}}-$ $-\overset{\vert}{\underset{\vert}{C}}-$ $-\underset{\cdot}{\dot{C}}\cdot$ $-C=$
C═	$=C-$ $=\overset{\vert}{C}-$ $=C$ $=\dot{C}\cdot$
C·	$\cdot C\cdot$ $\cdot\overset{\vert}{C}\cdot$ $\cdot\dot{C}\cdot$ $\cdot\underset{\cdot}{\overset{\vert}{C}}\cdot$ $\cdot\dot{C}=$ $\cdot\dot{C}*$ $\cdot\ddot{C}\cdot$ $\cdot\underset{\vert}{\dot{C}}-$ $\cdot\underset{\cdot}{\dot{C}}\cdot$ $\cdot\overset{\vert}{C}:$ $\cdot C:$
C:	$:C\cdot$ $:\overset{\vert}{C}\cdot$ $:\dot{C}\cdot$
C*	$*C*$ $*\overset{\vert}{C}*$ $*\dot{C}*$ $*\overset{*}{C}*$
O—	—O— —O
O═	═O
O·	·O·
N—	$-N-$ $-\overset{\vert}{N}-$ $-N$ $-\dot{N}\cdot$ $-\overset{\Vert}{N}=$ $-N=$
N·	$\cdot\overset{\vert}{N}\cdot$ $\cdot N\cdot$ $\cdot\dot{N}\cdot$ $\cdot N:$
N*	*N*

RING DESCRIPTION

The structures which contain ring systems are divided into two classes, those containing only monocycles or fused systems with 1:1 and 1:2 fusions, and those which contain more complex fusions.[10] The structures contain-

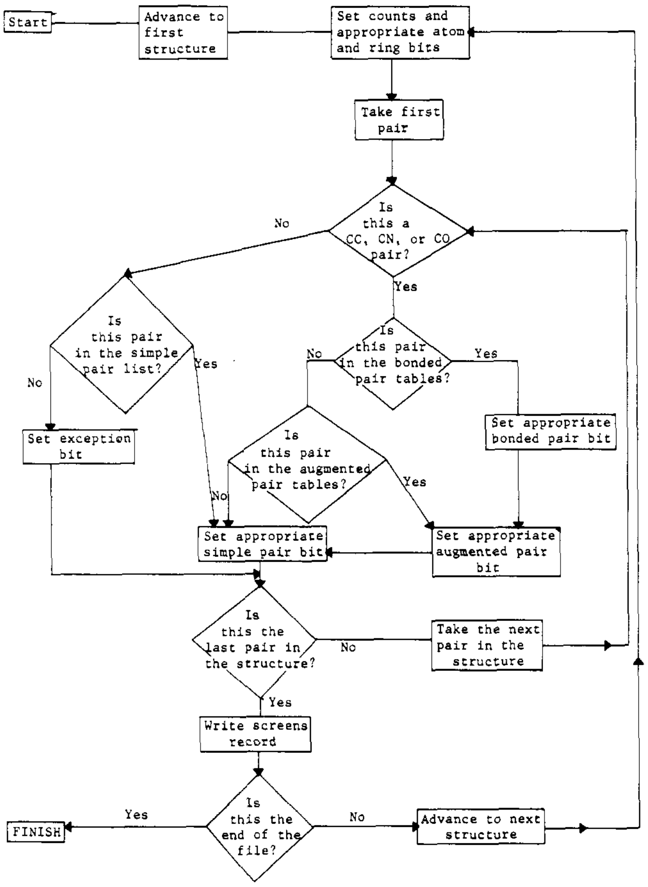
Start
Advance to first structure
Set counts and appropriate atom and ring bits
Take first pair
Is this a CC, CN, or CO pair?
No
Yes
Is this pair in the simple pair list?
Yes
No
Is this pair in the bonded pair tables?
No
Yes
Set appropriate bonded pair bit
Set exception bit
Is this pair in the augmented pair tables?
No
Yes
Set appropriate simple pair bit
Set appropriate augmented pair bit
Is this the last pair in the structure?
No
Take the next pair in the structure
Yes
Write screens record
Is this the end of the file?
Yes
FINISH
No
Advance to next structure

Figure 4. Generalized flow design of screen-generation procedures, showing details of the hierarchical bond-centered screen production

ing only monocyclic and simple fused ring systems are analyzed to give the primary rings included in them. The bit corresponding to each ring is again found by using implicit tables. The position of the bit is defined by three characteristics of the ring: whether it is a monocycle or part of a fused system, the ring size, and the atomic constitution of the ring, as shown in Table V.

If a structure contains a complex ring system which cannot be analyzed by the algorithm, then all bits in the ring screens field are set on. This removes the possibility of recall failures at a cost of a small decrease in precision and is an interim measure pending implementation of more comprehensive treatment.

PROGRAM CHARACTERISTICS AND PERFORMANCE

The screen generation programs were written in PLAN, the ICL assembly language. The screens for a sample of 28963 structures[8] (average size 21 atoms) from the Chemical Abstracts Service Registry System were generated on an ICL 1907 computer. This computer has a 24-bit word length and performs simple machine operations—e.g., add into accumulator in 2.86 microseconds. The screen generation took 2100 seconds of C.P.U. time and needed 9.3 K words of core store and two magnetic tape decks. One tape deck was used for the tape which held the structures in the form of nested, nonredundant connection tables, and the other was for the magnetic tape on which the screen records were written. No additional storage was required.

Table V. Ring Characteristics Used in Screen Generation

I. Fusion type
(a) monocycle
(b) 1:1 or 1:2, fused system
(c) more complex fused system
II. Ring sizes (n) (fusion types a and b only)
3, 4, 5, 6, 7, ⩾8
III. Atomic constitution (fusion types a and b only)
C_n, $C_{n-1}N$, $C_{n-x}N_x (x \geqslant 2)$, $C_{n-1}O$, $C_{n-x}O_x (x \geqslant 2)$,

$C_{n-1}X$, $C_{n-2}NO$, $C_{n-2}NX$, $C_{n-2}OX$,

$C_{n-(a+b+c)}N_aO_bX_c$ (excluding rings included in the atomic constitution classes given above)

ACKNOWLEDGMENT

We thank OSTI for financial support for this work, and the Chemical Abstracts Service for provision of the sample of structures.

LITERATURE CITED

(1) Lynch, M. F., Harrison, J. M., Town, W. G., and Ash, J. E., "Computer Handling of Chemical Structure Information," London, Macdonald, and New York, Elsevier, 1971.

(2) "Survey of Chemical Notation Systems," NAS/NRC Publ. No. 1150, Washington, D. C., 1964.

(3) 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).

(4) Adamson, G. W., Clinch, V. A., and Lynch, M. F., University of Sheffield, Sheffield, England, unpublished data.

(5) Lederberg, J., "Application of Artificial Intelligence for Chemical Inference, I. The Number of Possible Organic Compounds. Acyclic Structures Containing C,H,O and N," *J. Amer. Chem. Soc.* **91**, 2973-6 (1969).

(6) Mooers, C. N., "Zatocoding Applied to Mechanical Organization of Knowledge," *Amer. Doc.* **2**, 20-32 (1951).

(7) Adamson, G. W., Lambourne, D. L., and Lynch, M. F., "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File, Part III, Statistical Association of Fragment Incidence," *J. Chem. Soc. (Perkin I)*, **1972**, 2428-33.

(8) Crowe, J. E., Lynch, M. F., and Town, W. G., "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File, I, Non-cyclic Fragments," *J. Chem. Soc. C*, **1970**, 990-6.

(9) Adamson, G. W., Lynch, M. F., and Town, W. G., "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File, II, Atom-centered Fragments," *J. Chem. Soc. C*, **1971**, 3702-6.

(10) Adamson, G. W., Cowell, J. Lynch, M. F., Town, W. G., and Yapp, A. M., "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File, Part IV, Cyclic Fragments," University of Sheffield, Sheffield, England, unpublished data.

(11) Meyer, E., "Superimposed Screens for the GREMAS System," *Proc. FID-IFIP Conference*, K. Samuelson, Ed., Rome, **1967**, North-Holland, **1968**, 280-8.

(12) Armitage, J. E., and Lynch, M. F., "Some Structural Characteristics of Articulated Subject Indexes," *Inform. Stor. Retr.* **4**, 101-11, (1968).

(13) Clare, A. C., Cooke, E. M., and Lynch, M. F., "The Identification of Variable-length, Equifrequent Character Strings in Natural-Language Data-Bases," *Computer J.* **15**, 259-62, (1972).