Atom type is designated by displaying the atom name or drawing circles of specified radii for the different types of atoms. The former is useful for viewing an individual molecule or a portion thereof but is not very practical when the full contents of a crystallographic unit cell are displayed, due to confusion resulting from a high degree of overlap of atomic symbols and bonds. Circles show atom type equally well. The size of the circles is an input option, so that ionic, covalent, Van der Waals, or hydrogen bonding radii may be chosen to help characterize the structure of interest.

The user may choose to display an individual molecule (the crystallographic asymmetric unit) or the full contents of a crystallographic unit cell, the other molecules being related by symmetry operations to the asymmetric unit. This latter option facilitates the study of intermolecular interactions in the solid state. (A limited number of colored pictures and viewing glasses is available from the author.)

For more complicated molecules the picture sometimes becomes complicated and therefore a "zoom" feature may be invoked to display only the contents of a desired volume element, enlarged to fill the screen. With the present version of the program DISPLAY the user may ask for output of bond lengths; a future version will include extended geometric routines for extraction of bond angles, torsion angles, and best-fit planes of specific atoms. These features will be initiated by means of a track ball (X-Y) pointer that allows the operator to indicate the desired atoms individually.

As the title implies, an attempt has been made to produce three-dimensional structural and chemical information in a visual, meaningful manner with a minimum of human interaction in the steps between the library and the console. Scaling up by one or more orders of magnitude will surely present problems that will necessitate changes in strategy and program logic. At this point the essential facts are that a library exists, a real-time terminal with 3-D graphical display exists, and an automatic link has been established between them.

All programs have been written in Fortran IV. Generous support has been given by the College of Agriculture, Texas A&M University, and by the U.S. Atomic Energy Commission, through the hospitality of Brookhaven National Laboratory. The support and encouragement of Walter Hamilton and Olga Kennard is gratefully acknowledged.

LITERATURE CITED

(1) Kennard, O., and D. G. Watson, "Crystal Structure Library Manual," University Chemical Laboratory, Cambridge, U.K. (1969).

(2) Donnay, J. D. H., G. Donnay, E. G. Cox, O. Kennard, and M. V. King, "Crystal Data, Determinative Tables, 2nd Edition" (1963), Monograph **5**, American Crystallographic Association.

(3) Pearson, W. B., Ed., "Structure Reports," N. V. A. Oosthoeks Uitgevers Mij., Utrecht.

(4) Johnson, C. K., "ORTEP: A Fortran Thermal-Ellipsoid Plot Program for Crystal Structure Illustrations" (1965). ORNL, 3794.

(5) Meyer, E. F., "Three Dimensional Graphical Models of Molecules and a Time-Slicing Computer," *J. Appl. Crystallog.*, in press.

(6) Ophir, D., S. Rankowitz, B. J. Shepherd, and R. J. Spinrad, *Comm. ACM*, Vol. **11**, No. 6, p. 415 (1968).

(7) Ophir, D., B. J. Shepherd, and R. J. Spinrad, *Comm. ACM*, Vol. **12**, No. 6, p. 309 (1969).

(8) Whittingham, D. J., F. R. Wetsel, and H. L. Morgan, J. CHEM. Doc. **6**, 230 (1966).

(9) Morgan, H. L., J. CHEM. Doc. **5**, 107 (1965), cf. p. 110.

A Utility Analysis for the MCC Topological Screen System*†

DAVID LEFKOVITZ**

The Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pa. 19104

ALFONSO R. GENNARO

The Philadelphia College of Pharmacy and Science, Philadelphia, Pa.

Received August 18, 1969

This paper is the fourth in a series that has attempted to develop a systematic approach to the problem of searching chemical substructures in a large-scale automated system. The three preceding papers in the series were:

"Use of a Nonunique Notation in a Large-Scale Chemical Information System"[1]

"A Chemical Notation and Code for Computer Manipulation"[2]

"Substructure Search in the MCC System"[3]

The first of these discussed a systems concept which includes the five fundamental requisites of a chemical information storage and retrieval system—viz, structure input, registry, search file generation and storage, substructure search, and structural display. A conclusion of this paper was that a computer-oriented chemical code was a desirable component of the system in order to represent the compound as accurately as a connection table and as concisely as a notation; in addition, this code should be readily convertible to a connection table for the purposes of further computer manipulation, such as screen assignment, atom by atom search and display.

The second paper specified such a language and indicated how it met the requirements of the former paper. This language was called the Mechanical Chemical Code (MCC).

* Presented in part at the 4th Middle Atlantic Regional Meeting, ACS, Washington D. C., February 1969.

† This study was supported by the National Science Foundation under Contract No. NSF-C547.

** Present addres, COMSERV, 4025 Chestnut St., Philadelphia, Pa. 19104.

A Utility Analysis for the MCC Topological Screen System

This paper presents a three-part analysis and evaluation of a substructure search system. The system is based on the use of two indexes in order to perform the search. The first is a KWIC-style index of the machine-generated fragments. The second is an inverted index to enable random look-up in the file, based on logical combinations of the fragments. The fragments are defined by a topological algorithm, and the fragments that it produces are expressed in the MCC notation. It is therefore called the MCC Topological Screen System. A previous paper described the theory of the system. The three parts of the analysis are: the technique for ordinary substructure search, a unique technique for browsing in a compound file, and some search statistics and a measure of screening efficiency.

The third paper presented the design concept of a screen system that is amenable to machine assignment and which may be applicable to a broad range of substructure query types. The current paper is an examination of the potential utility of this screen system. Its purpose is to discuss the results of a preliminary utility analysis of the technique performed on three data bases. The first consisted of 25,000 compounds from the Chemical Biological Coordination Center (CBCC) file; the second data base consisted of 103,500 compounds selected from the CAS Registry System, and the third consisted of approximately 20,000 compounds processed and registered through the CAS system, called the Common Data Base, a particular collection of compounds assembled by the Food and Drug Administration and the National Library of Medicine. The utility analysis is presented in three parts. In the first, a straightforward example of a substructure search is presented in 3 steps. Step 1 states the question as the chemist might ask it; in step 2 the question is analyzed in terms of the system screens, and a formulation is generated for submission to the system, and in step 3 the modified question that is actually being answered by the formulation of the analysis is presented so that one can judge the effectiveness of the system.

In the second part the system's applicability to browsing is examined, and in the third a measure of screening efficiency is discussed.

The screens of the system are generated by a relatively simple set of topologically oriented rules,[3] reviewed briefly in the next section. For this reason it is called a Topological Screen System (TSS). Before proceeding to the three analyses, a unique property of the system requires some general explanation. Consider the cross-hatched region of Figure 1 to be a schematic representation of the boundary of a substructure inquiry. The set of responses may then be represented as shown, where the boundary of the inquiry is completely interior (or a subset) of the response set. The most common systems today use an approach that effectively produces the response set given a substructure inquiry. A vocabulary of screens or fragmentation codes is devised, and each file compound is assigned all applicable screens from the vocabulary. A search is conducted by accessing all structures from the filé that contain a given combination of screens correspondent to the screens of the substructure inquiry. Considering the fragments to be analogous to document index terms this technique is referred to as a coordinate index and may be implemented in automated systems by edge-notched cards, optical-coincidence cards (peek-a-boo), punched EAM cards, bit strings or integer lists in a serial magnetic tape file, and threaded or inverted lists in a random access disk file. Their common attribute is that no relationship or hierarchy is assumed to exist among the screens (descriptors in a library context) of the vocabulary; each may be used in any logical combination (AND, OR, NOT) with any other. Since, in general, an application of the screens alone is not sufficient to confine the response set to one that precisely contains the substructure inquiry, an iterative atom-by-atom search may be employed to reduce the screened set to the appropriate one implied by the response set of Figure 1. The quality of the screen system, however, may be measured by the difference between the screened and the actual response sets. The TSS offers a random access file organization technique for enabling a computer to economically access the response set of Figure 1. The first subject of this utility analysis was a test of the system's screening and random access effectiveness in producing as near to the response set as to require minimal atom by atom search. A measure of this closeness is developed in the paper and the results for a controlled test of 35 questions is presented.

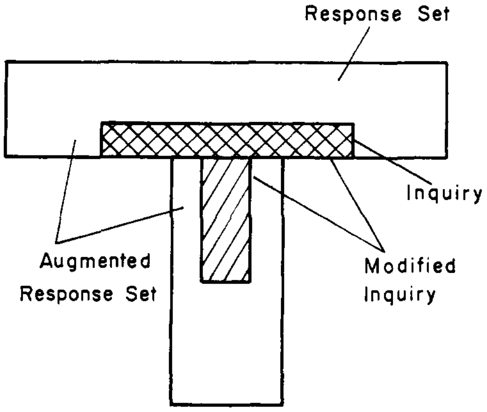


Figure 1. Logical schema for searches in an information system

However, the system, in principle, has another, and possibly more interesting, property that was the subject of a second part of the analysis. Consider another set, called the modified inquiry, in Figure 1. Assume that this represents a series of modified inquiries that partially overlap the original inquiry, where the modification may be guided and controlled in accordance with the content of the file. If the total inquiry set is then regarded as the shaded T of Figure 1, an augmented response set may be defined as the boundary of the outer T. The significance of the augmented response set is that for many searches, particularly those regarding synthesis, patents and homologs, the "near hits" contained in the upright leg of the augmented response set in Figure 1 may be as important as the original response set. In other words, the question that *should* be asked may be more important than the question that *is* asked.

In coordinate index systems, the only way to produce an augmented response set is to relax the screen conditions, which is an inefficient means of controlling the closeness of the additional responses, particularly if the screens are coarse.

The requirement to produce an augmented response set, however, is not new or unique to chemical search; it is called browsing or serendipitous search in library systems, and it is facilitated by classification of the descriptor vocabularies. That is, the screens are classified by arranging them in an index according to certain defined relationships. Thus, a look-up is first made in the classified index for the precise screen or screen set of the original inquiry, and then a systematic search may be made in this same index by screen class, whereupon all structures within the class (or within more or less specific classes in the same vicinity of the index) can be readily identified. The screens in the vicinity of the original points of entry will constitute a modified inquiry and the responses thereto, the augmented response set.

It is difficult to argue with the validity of this concept, but effectiveness of the browsing procedure is for the most part determined by the meaning and quality of the classification. Traditionally, a classification schedule or thesaurus relating the index terms or descriptors to particular classes was established on a generalized basis, without prior knowledge of the exact content of the document collection to be so classified. For example, the Dewey Decimal system was conceived in 1873 as an hierarchic (tree-like) classification with up to 10 branches per node. The root and lower branches of the tree were firmly established at the inception, with subsequent modification and expansion being limited to unused or uppermost branches. This kind of classification may be called *a priori* with respect to a given collection.

With the advent of computers it has become possible to process, in great detail and in varied ways, data associated with a given collection of documents. Thus, for example, one can generate a term concordance which enumerates and counts the various words of a text or a series of texts; or one might conceive of a way to manipulate the index terms of the document collection so as to generate a classification schedule specific to a given collection.[4, 5] Such a process could be called *a posteriori* classification, since it is performed for a particular document collection and only after having analyzed the collection. The inherent value of *a posteriori* classification is that the associations or relations implied by the classification schedule are most meaningful in the browsing process because they derive from existing relationships in the actual document collection at hand. Also, if the system is properly conceived, it can have added value by unequivocally announcing the nonexistence of certain documents in the collection by virtue of the absence of a given relationship.

The TSS is an *a posteriori* classification system, where the index terms are machine-generated fragments, and the classification schedule is an alphabetically organized, rotated (permuted) index of the fragments. Since it is founded upon two arbitrary (but systematically meaningful) procedures that are virtually unrelated to the structural chemistry—namely, the topological screen fragment and their alphabetical organization—the value of the classification produced thereby is open to question. It is this question, with specific regard to its implications for meaningful definition of the modified inquiry in Figure 1, that was also a subject of this utility analysis.

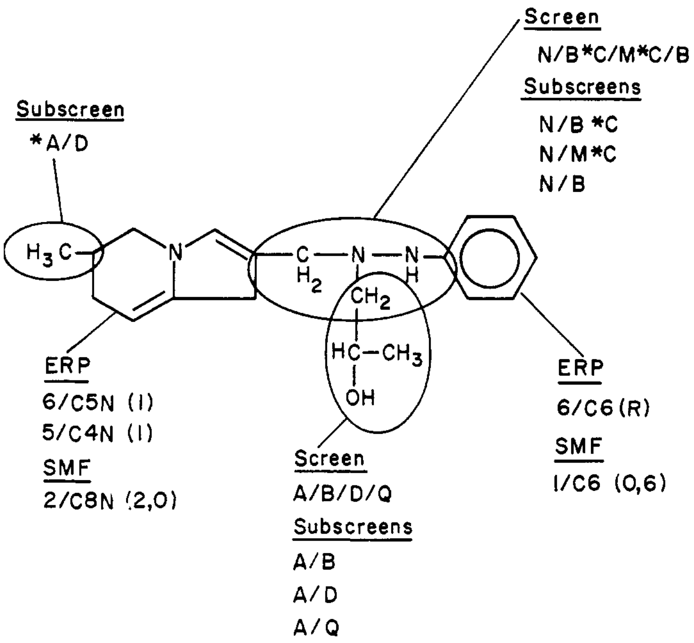


Figure 2. TSS screen assignment algorithms

ALGORITHM FOR TSS FRAGMENT GENERATION

Reference 3 presents the specific algorithms for the screen and index generation of the TSS. Similar algorithms have been developed for use with the WLN symbols by Hyde.[7] The TSS rules are briefly summarized by the example of Figure 2 and the subsequent definitions.

(1) Screen
$\chi/\alpha/\beta/\gamma$
χ = branch atom (central atom)
α, β, γ = MCC expression for radiating chains up to and including a (*i*) terminal, (*ii*) ring attachment and not including (*iii*) another central atom. If two central atoms are adjacent then they *are* included in the radiating chain of each other.

(2) Subscreen
χ/α

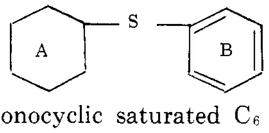
(3) ERP (Elementary Ring Population)
The set of ERP's for a nucleus is the smallest set of smallest rings.
Z/α (n)
Z = Number of atoms in ring
α = Molecular formula of ring in standard symbols
n = Number of double bonds in ring. If $n = R$, ring is resonant. If the ring is fused to one or more resonant rings, the resonant bonds in the ERP are counted as single bonds.

(4) SMF (Skeleton Molecular Formula)
Z/α (m,n)
Z = Numbers of rings in nucleus
α = Molecular formula of nucleus in standard symbols
m = Number of double bonds in nucleus
n = Number of resonant bonds in nucleus

THE BASIC METHODOLOGY OF SUBSTRUCTURE SEARCH

For the first question there is a three-part discussion. In the first, the question, as expressed by the chemist, is presented; in the second, an analysis is made in terms of the operational characteristics of the TSS screens and its file organization; in the third, the modified question is presented. This is the actual question that the system is responding to, without the use of an atom-by-atom search. In all cases the original question is either equivalent to or a subquestion of the modified question, and the value of the system is to be judged by the closeness of the modified to the original question and the size and complexity of the files and their organization.

Question 1: Find the compounds containing the cyclohexyl aryl sulfide substructure



Where A is a monocyclic saturated C_6 nucleus with no substituents except as shown, and B is an aromatic C_6 ring that may be part of a larger nucleus and may have arbitrary substituents.

Analysis: Since Ring A may have no other substituents, the attachment point must be CH, or *a in MCC notation. The attachment point in Ring B is a carbon atom with no hydrogens, or *C. The link between the nuclei is therefore *a/S*C. There are at least two nuclei, which are screened as follows:

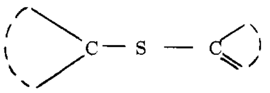
Nucleus A

SMF: 1/C6 (0,0)

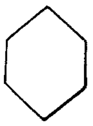
Nucleus B

ERP: 6/C6 (R)

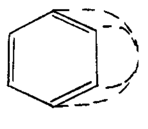
Modified Question: Find all compounds containing the substructure:



with two nuclei, one being of the form



and the other of the form



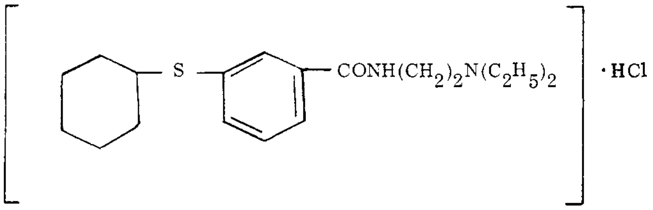
```
               1/C5N2S(3,0)                    (    1)   105
------------   1/C50(0,0)  ------------------  (  279)   106  ---
               1/C50(1,0)                      (   11)   107
------------   1/C50(2,0)  ------------------  (   38)   108  ---
               1/C6(0,0)                       (  355)   109
_____________  1/C6(0,6)   __________________  (8928)    110  ___
               1/C6(1,0)                       (  125)   111
------------   1/C6(2,0)   ------------------  (  266)   112  ---
               1/C6(3,0)                       (    1)   113
------------   1/C6N(0,0)  ------------------  (   20)   114  ---
               1/C60(0,0)                      (    3)   115
_____________  1/C60(2,0)  __________________  (    1)   116  ___
```

Figure 3. SMF Index—Question 1

where the substituents (dashed line) may or may not be cyclic.

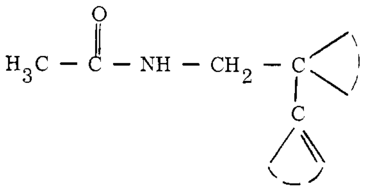
The cyclic screens for this question are highly generic, and, as identified in Figures 3 (the SMF) and 4 (the ERP), the compound populations are 355 (screen 109) and over 10,000 (Screen 212), respectively. (File size is 20,000 compounds.)

The conjunction of these three screen lists yielded a single response—namely, compound A00216970, as shown below.



BROWSING

The system appears to present opportunities for browsing by virtue of its organization of the machine-generated fragments, particularly the acyclic screens. Figure 5 presents a small part of such an index, as it is printed by the computer line printer. There are three columns in the index repeated twice per computer page. The first contains the screen; the second contains the number of inverted list postings, in parentheses; and the third contains a unique serial number for the screen. The screens are rotated (or shifted) in the index by chain around the central atom. For example, screen 19 appears in Figure 5 as *C/*C/BMLD, which represents the fragment:



It also appears elsewhere in the index (second column of Figure 5) as *C/BMLD/*C. Note, however, that it appears, in its first instance, in the following hierarchy of structural environments:

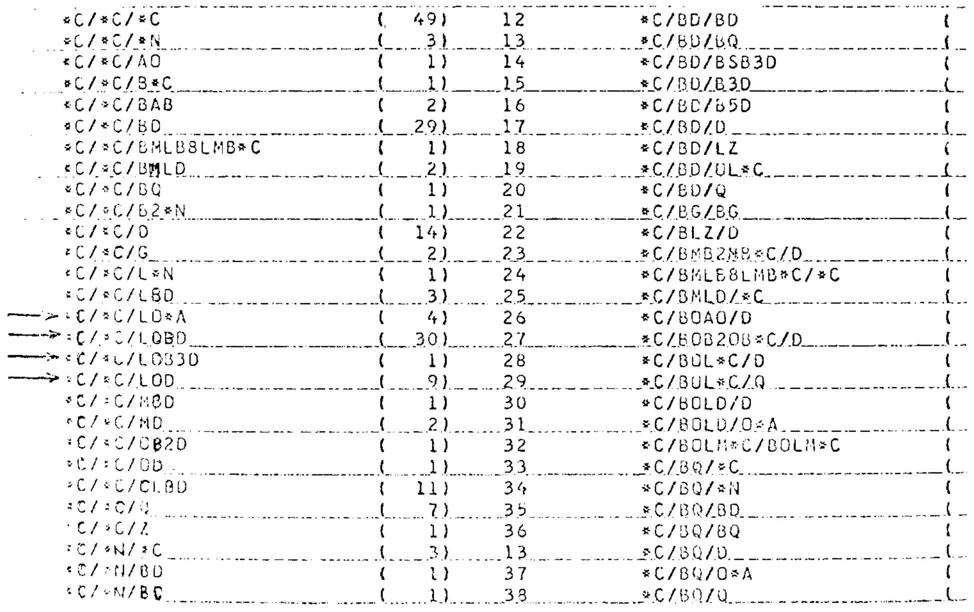
*C/ ______
*C/*C/ ______
*C/*C/B ______
*C/*C/BM ______
*C/*C/BML ______

and, in its second instance, it appears in the following hierarchy of structural environments.

*C/ ______
*C/B ______
*C/BM ______
*C/BML ______

```
          6/C50(1)                     ( 137)  206
--------  6/C50(2)  -----------------  ( 309)  207  -----
          6/C50(3)                     (   2)  208
________  6/C5S(1)  _________________  (   1)  209  _____
          6/C5S(2)                     (  28)  210
--------  6/C6      -----------------  (1487)  211  -----
          6/C6(R)                      (****)  212
--------  6/C6(1)   -----------------  (1012)  213  -----
          6/C6(2)                      (1215)  214
________  6/C6(3)   _________________  (  57)  215  _____
```

Figure 4. ERP Index—Question 1

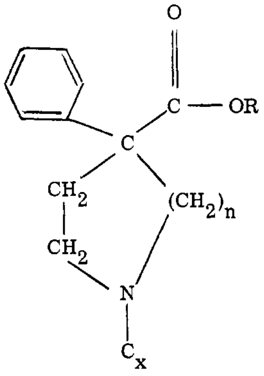


```
*C/*C/*C                   (  49)   12        *C/BD/BD                    (
*C/*C/*N                   (   3)   13        *C/BD/BQ                    (
*C/*C/AO                   (   1)   14        *C/BD/BSB3D                 (
*C/*C/B*C                  (   1)   15        *C/BD/B3D                   (
*C/*C/BAB                  (   2)   16        *C/BD/B5D                   (
*C/*C/BD                   (  29)   17        *C/BD/D                     (
*C/*C/BMLB8LMB*C           (   1)   18        *C/BD/LZ                    (
*C/*C/BMLD                 (   2)   19        *C/BD/UL*C                  (
*C/*C/BQ                   (   1)   20        *C/BD/Q                     (
*C/*C/B2*N                 (   1)   21        *C/BG/BG                    (
*C/*C/D                    (  14)   22        *C/BLZ/D                    (
*C/*C/G                    (   2)   23        *C/BMB2MB*C/D               (
*C/*C/L*N                  (   1)   24        *C/BMLB8LMB*C/*C            (
*C/*C/LBD                  (   3)   25        *C/BMLD/*C                  (
*C/*C/LD*A                 (   4)   26        *C/BOAO/D                   (
*C/*C/LOBD                 (  30)   27        *C/BOB2OB*C/D               (
*C/*C/LOB3D                (   1)   28        *C/BUL*C/D                  (
*C/*C/LOD                  (   9)   29        *C/BUL*C/Q                  (
*C/*C/MBD                  (   1)   30        *C/BOLD/D                   (
*C/*C/MD                   (   2)   31        *C/BOLD/O*A                 (
*C/*C/OB2D                 (   1)   32        *C/BOLM*C/BOLM*C            (
*C/*C/OD                   (   1)   33        *C/BQ/*C                    (
*C/*C/CLBD                 (  11)   34        *C/BQ/*N                    (
*C/*C/Q                    (   7)   35        *C/BQ/BD                    (
*C/*C/Z                    (   1)   36        *C/BQ/BQ                    (
*C/*N/*C                   (   3)   13        *C/BQ/D                     (
*C/*N/BD                   (   1)   37        *C/BQ/O*A                   (
*C/*N/BC                   (   1)   38        *C/BQ/Q                     (
```

Figure 5. Index entry for Question 2

It is in this way that the fragments are said to be classified. (Granito also noted this effect using WLN symbols,[8] and Bonnett reported on the use of classification numbers.[9]) The chemist, in each of the following examples, finds the screen or screens that respond to an original inquiry. Then he scans in the vicinity for other screens that would alter the responding structure, but which might still satisfy the general chemical activity requirement of the inquiry. In terms of the schematic of Figure 1, he constructs a modified inquiry after finding the screen (or screens) indicated by the original inquiry.

Question 2: The original question is to find all phenylpyrrolidinecarboxylic acid esters.



x = 1–5 (branched or straight chain)
n = 1–3
R = CH_3, —C_2H_5, —$CH(CH_3)_2$

The aromatic ring is isolated, with or without further substitution.

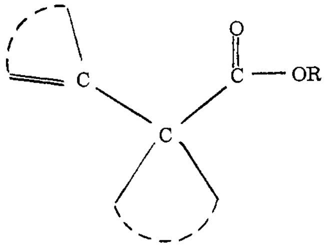
Three screens are applied to this question:

(1) SMF : 1/CnN (0,0) n = 4,6

(2) ERP : 6/C6 (R)

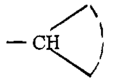
(3) Screen: *C/*C/LO

The portion of the index for the third screen, which represents the fragment



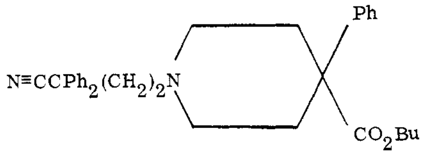
is shown (arrows) in Figure 5.

Screens 27 and 29 are responsive to the question as originally presented. Screens 26 and 28 are potentially of interest if they successfully conjoin with the other screens of the question. Both represent what in Figure 1 is called a modified inquiry. In particular, screens 26 and 28 altered the meaning of R as follows:

Screen 26: R =


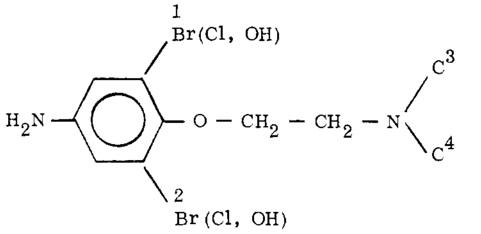
Screen 28: R = —$(CH_2)_3$—CH_3

Neither of these is indicated in the original inquiry, but screen 28 did in fact produce an analog within the file, as RN 15, 203, 053 (from the CAS registry system) which is:



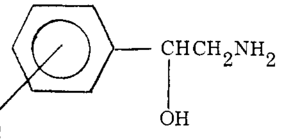
This compound is thus a member of the augmented response set but *not* of the original response set.

Question 3: Find all *p*-(dialkylaminoethoxy) anilines.

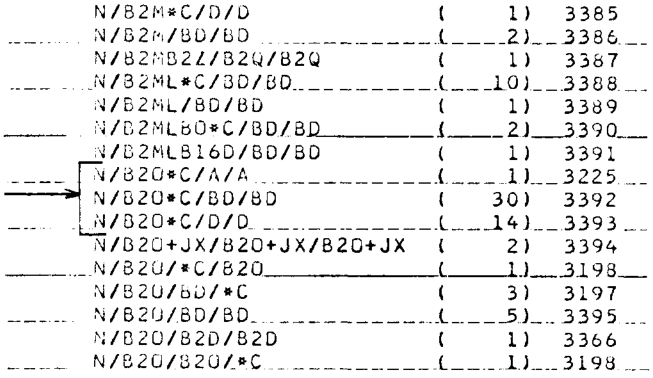


Atoms 1 and 2 may be Br, Cl, or OH. Carbon atoms 3 and 4 may be in chain or ring (length unspecified). Analogs for this question were found in a somewhat more directed way. The screen appropriate to the original query is N/B2O*C/α _____/β _____, where α and β are any carbon symbol. Figure 6 presents the three responding screens. At this point the chemist modified the question to N/BnO*C/α _____/β _____ or N/BnS*C/α _____/β _____, and these produced screens 3470 and 3474 as shown in Figure 7. On intersecting these screens with those of the remainder of the question, no responses were found, so that the search, in fact produced no actual analogs; however, any responses would have been analogs.

Question 4: Find α- and ar-hydroxybenzylamine analogs.

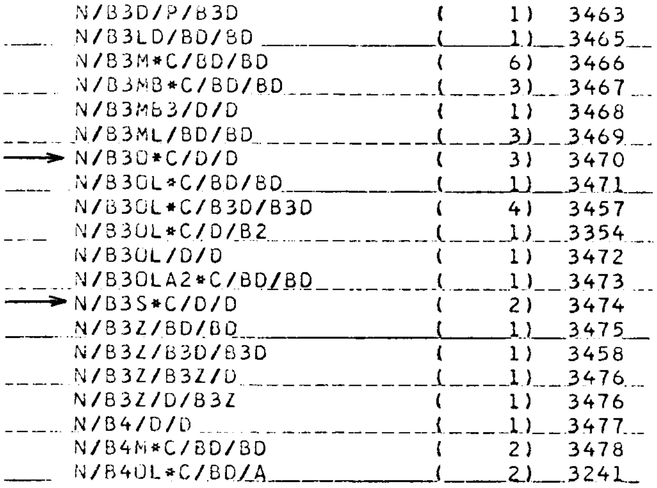


where Q is one or more OH substitutions.



```
  N/B2M*C/D/D                 (   1)  3385
  N/B2M/BD/BD                 (   2)  3386
  N/B2MB2Z/B2Q/B2Q            (   1)  3387
  N/B2ML*C/BD/BD              (  10)  3388
  N/B2ML/BD/BD                (   1)  3389
  N/B2MLBO*C/BD/BD            (   2)  3390
  N/B2MLB16D/BD/BD            (   1)  3391
  N/B2O*C/A/A                 (   1)  3225
→ N/B2O*C/BD/BD               (  30)  3392
  N/B2O*C/D/D                 (  14)  3393
  N/B2O+JX/B2O+JX/B2O+JX      (   2)  3394
  N/B2O/*C/B2O                (   1)  3198
  N/B2O/BD/*C                 (   3)  3197
  N/B2O/BD/BD                 (   5)  3395
  N/B2O/B2D/B2D               (   1)  3366
  N/B2O/B2O/*C                (   1)  3198
```

Figure 6. Index entry for Question 3



```
  N/B3D/P/B3D                 (   1)  3463
  N/B3LD/BD/BD                (   1)  3465
  N/B3M*C/BD/BD               (   6)  3466
  N/B3MB*C/BD/BD              (   3)  3467
  N/B3MB3/D/D                 (   1)  3468
  N/B3ML/BD/BD                (   3)  3469
→ N/B3O*C/D/D                 (   3)  3470
  N/B3OL*C/BD/BD              (   1)  3471
  N/B3OL*C/B3D/B3D            (   4)  3457
  N/B3OL*C/D/B2               (   1)  3354
  N/B3OL/D/D                  (   1)  3472
  N/B3OLA2*C/BD/BD            (   1)  3473
→ N/B3S*C/D/D                 (   2)  3474
  N/B3Z/BD/BD                 (   1)  3475
  N/B3Z/B3D/B3D               (   1)  3458
  N/B3Z/B3Z/D                 (   1)  3476
  N/B3Z/D/B3Z                 (   1)  3476
  N/B4/D/D                    (   1)  3477
  N/B4M*C/BD/BD               (   2)  3478
  N/B4OL*C/BD/A               (   2)  3241
```

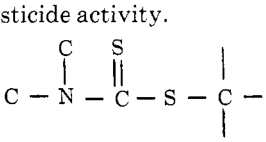
Figure 7. Question 3 (modified)

The indicated screens are:

(1) SMF: 1/C6(0,6)
(2) Subscreen: *C/Q (one or more)
(3) Screen: A/Q/*C/BZ

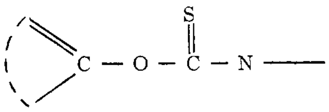
The index shown in Figure 8 indicates a series of screens of the form A/Q/*C/BnX _____, for any n, and where X is N, M, or Z. These screens represent a modified inquiry and produced the five variants in Figure 9 as an augmented response set.

Question 5: The following is a basic structure associated with potential pesticide activity.



The principal search screen is C/N/S/Sα, where α is either nothing or a carbon symbol. By using the browsing feature, compounds with a similar substructure were found that may also exhibit such activity.

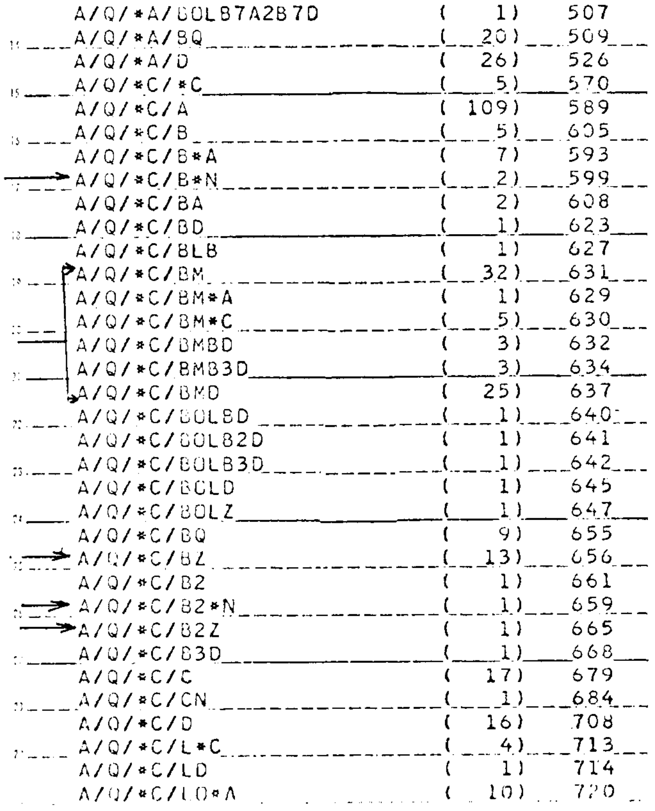
Figure 10 presents the relevant part of the index. Three compounds satisfied this inquiry, as shown in Figure 11,*a* but in addition, the index also indicates that there is a compound containing the substructure



(Screen 2946)

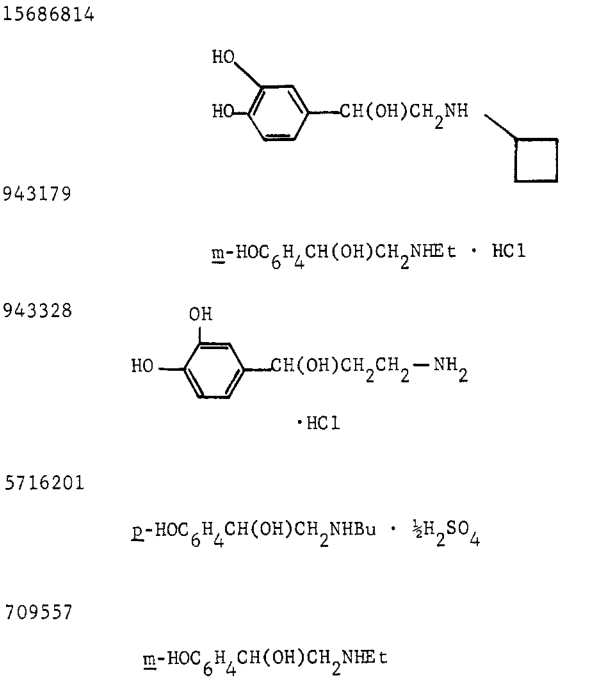
which the chemist recognized as being applicable. The compound containing this screen is shown in Figure 11,*b* and also may exhibit pesticide activity.

The fact that screen 2946 is so close physically to the others in the index is an accident, because the letter



```
  A/Q/*A/BOLB7A2B7D           (   1)   507
  A/Q/*A/BQ                   (  20)   509
  A/Q/*A/D                    (  26)   526
  A/Q/*C/*C                   (   5)   570
  A/Q/*C/A                    ( 109)   589
  A/Q/*C/B                    (   5)   605
  A/Q/*C/B*A                  (   7)   593
→ A/Q/*C/B*N                  (   2)   599
  A/Q/*C/BA                   (   2)   608
  A/Q/*C/BD                   (   1)   623
  A/Q/*C/BLB                  (   1)   627
  A/Q/*C/BM                   (  32)   631
  A/Q/*C/BM*A                 (   1)   629
  A/Q/*C/BM*C                 (   5)   630
  A/Q/*C/BMBD                 (   3)   632
  A/Q/*C/BMB3D                (   3)   634
  A/Q/*C/BMD                  (  25)   637
  A/Q/*C/BOLBD                (   1)   640
  A/Q/*C/BOLB2D               (   1)   641
  A/Q/*C/BOLB3D               (   1)   642
  A/Q/*C/BOLD                 (   1)   645
  A/Q/*C/BOLZ                 (   1)   647
  A/Q/*C/BQ                   (   9)   655
→ A/Q/*C/BZ                   (  13)   656
  A/Q/*C/B2                   (   1)   661
→ A/Q/*C/B2*N                 (   1)   659
→ A/Q/*C/B2Z                  (   1)   665
  A/Q/*C/B3D                  (   1)   668
  A/Q/*C/C                    (  17)   679
  A/Q/*C/CN                   (   1)   684
  A/Q/*C/D                    (  16)   708
  A/Q/*C/L*C                  (   4)   713
  A/Q/*C/LD                   (   1)   714
  A/Q/*C/LO*A                 (  10)   720
```

Figure 8. Index entry for Question 4



15686814

943179

$\underline{m}$-$HOC_6H_4CH(OH)CH_2NHEt \cdot HCl$

943328

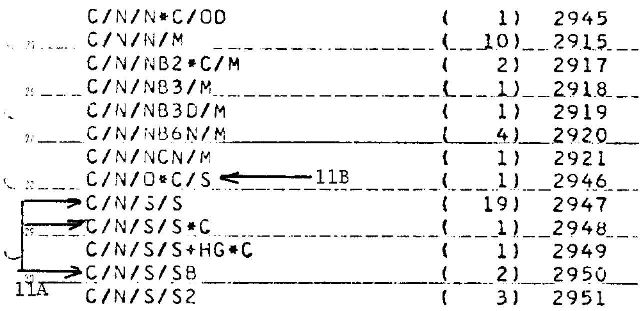
5716201

$\underline{p}$-$HOC_6H_4CH(OH)CH_2NHBu \cdot \frac{1}{2}H_2SO_4$

709557

$\underline{m}$-$HOC_6H_4CH(OH)CH_2NHEt$

Figure 9. Responses to Question 4



C/N/N*C/OD	(1)	2945
C/N/N/M	(10)	2915
C/N/NB2*C/M	(2)	2917
C/N/NB3/M	(1)	2918
C/N/NB3D/M	(1)	2919
C/N/NB6N/M	(4)	2920
C/N/NCN/M	(1)	2921
C/N/O*C/S	(1)	2946
C/N/S/S	(19)	2947
C/N/S/S*C	(1)	2948
C/N/S/S+HG*C	(1)	2949
C/N/S/SB	(2)	2950
C/N/S/S2	(3)	2951

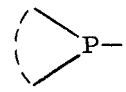
Figure 10. Index entry for Question 5

O is not too far from S, and, for this file, no intervening screens were produced. In general, however, one would have to scan the entire class C/N ______/α ______/S, where the letter α precedes S, alphabetically, and C/N/S/β ______, where the letter β follows S.

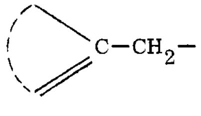
Question 6: The following substructure is associated with antineoplastic activity:

$$-N[(CH_2)_2Cl]_2$$

The screen is N/B2G/B2G/ ______, and the index is shown in Figure 12. The chemist was able to see nine possible attachments to the nitrogen, and he made a selection from these of five that were of interest—namely:

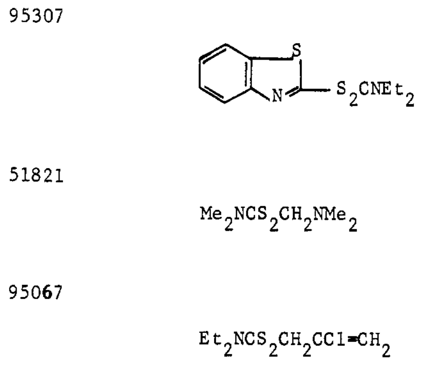
(1)  (Screen 3223)

(2) $-CH_2-$ (Screen 3282)

(3)  (Screen 3376)

(4) $-(CH_2)_3-$ (Screen 3379)

(5) —P— (Screen 3382)



95307

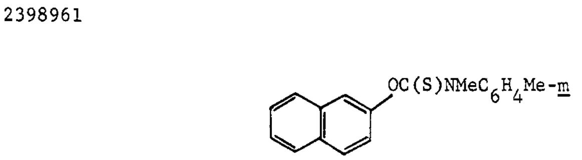
51821

$Me_2NCS_2CH_2NMe_2$

95067

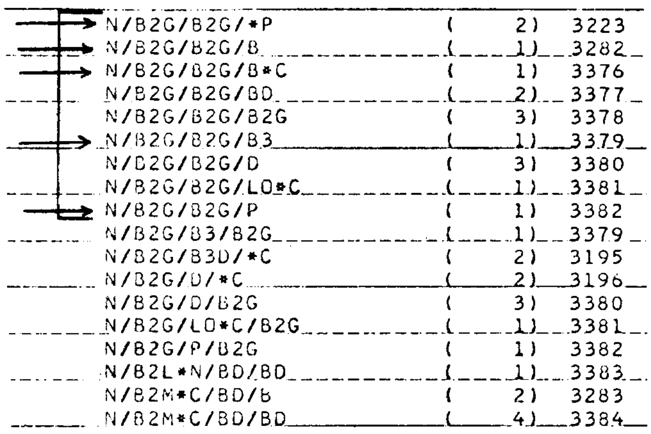
$Et_2NCS_2CH_2CCl{=}CH_2$

Figure 11,*a*. Responses to Question 5



2398961

Figure 11,*b*. Additional response to Question 5



N/B2G/B2G/*P	(2)	3223
N/B2G/B2G/B	(1)	3282
N/B2G/B2G/B*C	(1)	3376
N/B2G/B2G/BD	(2)	3377
N/B2G/B2G/B2G	(3)	3378
N/B2G/B2G/B3	(1)	3379
N/B2G/B2G/D	(3)	3380
N/B2G/B2G/LO*C	(1)	3381
N/B2G/B2G/P	(1)	3382
N/B2G/B3/B2G	(1)	3379
N/B2G/B3D/*C	(2)	3195
N/B2G/D/*C	(2)	3196
N/B2G/D/B2G	(3)	3380
N/B2G/LO*C/B2G	(1)	3381
N/B2G/P/B2G	(1)	3382
N/B2L*N/BD/BD	(1)	3383
N/B2M*C/BD/B	(2)	3283
N/B2M*C/BD/BD	(4)	3384

Figure 12. Index entry for Question 6

This produced the six compounds shown in Figure 13. Five of them were acceptable, and one was undesirable. This was compound 4,420,795 because of the —OH groups on the aromatic ring.

In summary, the system is able to produce responses that might not otherwise be produced, without the aid of the classification of fragments. In Question 5, compound 2,398,961 (Figure 11,*b*) would not have responded to the original question unless the questioner had anticipated (or had specifically been interested in) the oxygen in lieu of the sulfur bonded to the carbon.

In Question 2 the questioner can see at once all possibilities for R in Figure 5; these are the four instances (arrows) of *C/*C/LO ______. If he wants to remove the O, he sees two more screens—namely, *C/*C/L*N and *C/*C/LBD.

It is possible that the greatest value in this search technique will derive from such a browsing capability.

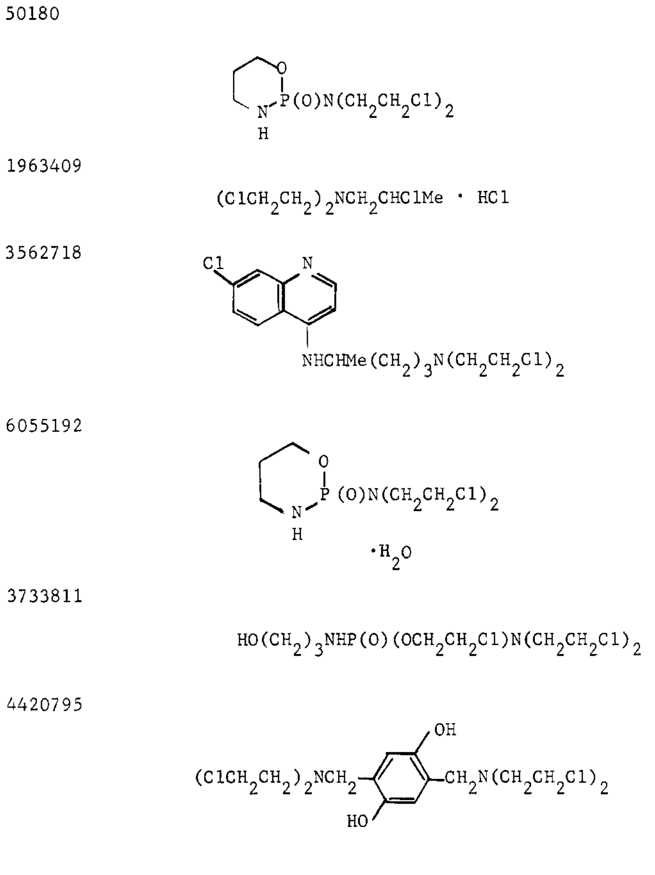


Figure 13. Responses to Question 6

SCREENING EFFICIENCY

In this part of the analysis, 35 questions were answered using the TSS index for the Common Data Base file.[6] The screened responses were all examined (as microfilmed pictures), and false drops eliminated. A measure of screening efficiency was taken to be the average ratio of valid-to-screened responses, where, as stated, the screened responses are those resulting from application of the TSS index search, and the valid responses are those that respond exactly to the question (the response set of Figure 1). This measure results in a number between 0 and 1, where 1 indicates perfect screening (no atom-by-atom search is required on any question included in the statistic), and an approach to 0 indicates very coarse and imperfect screening. 0 itself is a degenerate statistic because it indicates that all of the questions had no answers where one or more were actually screened. If there are no valid responses and there is none screened, the ratio is taken, by definition, to be unity.

Table I presents a summary of the results for these 35 questions. (The 35 questions and further discussion are presented in reference 6.) Column 1 contains the question number; column 2 contains the number of TSS screened responses; column 3 contains the number of valid responses, and column 4 contains the ratio of valid-to-screened responses. At the bottom of column 4, the average is shown to be 0.67, which means that, on the average, 2 out of 3 screened responses were valid.

Table I. Experimental Results for the Measure of Screening Efficiency

Question	TSS Screened Responses	Valid Responses	Valid/Screened
1	15	14	.93
2	18	4	.22
3	3	3	1.00
4	8	7	.87
5	4	4	1.00
6	27	3	.11
7	3	3	1.00
8	0	0	1.00
9	0	0	1.00
10	12	12	1.00
11	6	4	.66
12	1	1	1.00
13	2	1	.50
14	5	5	1.00
15	50	0	.00
16	9	8	.88
17	5	0	.00
18	3	2	.66
19	11	11	1.00
20	58	35	.60
21	5	3	.60
22	135	98	.73
23	211	13	.06
24	0	0	1.00
25	39	33	.85
26	3	3	1.00
27	30	20	.66
28	14	1	.07
29	134	14	.10
30	17	11	.65
31	150	76	.50
32	25	17	.68
33	4	4	1.00
34	2	0	.00
35	0	0	1.00
			Av. 0.67

CONCLUSIONS

The TSS indexes can be used readily for medium-sized files as a manual desk top tool if the TSS fragments in a question are relatively few in number and if no more than one very generic fragment is involved in each conjunction of terms. For example, a question containing the fragment $-CH_2COO-$ involves 87 different TSS fragments, from A/BLO through ZBLOLD in the index under BLO, and 288 different TSS fragments, from A/OLB through A/B2OLB9D in the index under OLB for the Common Data Base file of 20,000 compounds. Collectively, the "BLO or OLB" question represents a disjunctive search for 375 different screens. If the question were only as stated above, then the job of manually merging all registry numbers listed under 375 lists would be tedious, but not inordinately difficult; however, if the question were

$$-CH_2COO- \quad \text{AND} \quad =\overset{\overset{\displaystyle H}{|}}{C}-O-,$$

where the two fragments had to be present but not necessarily attached directly to each other, the merger of 375 lists of CH_2COO would have to be intersected with the merger of 62 lists of AO. For such questions the easy use of these indexes as manual desk top tools is clearly limited. The appropriate solution is to automate the list merge and intersection process. The inverted lists can be stored on a magnetic disk file and the rotated or classified index retained for manual use. The most optimal processing strategy for the above question would be to merge the 375 and 62 lists, respectively, and then to intersect the two resultant merged lists. This is called the logical product of sums and results in only 438 list processes, whereas the alternate procedure of intersecting, two by two, the 375 and 62 lists and then summing the resultant lists (called the logical sum of products) results in 23,251 list processes. The system, therefore, becomes more generally useful when (1) the inverted lists are loaded onto magnetic disk files and (2) a product of sums program is written to process these lists.

A further refinement is to load the rotated index onto a disk file and to program a scan that would find all screens containing a given sequence of letters. In the above example, the 87 BLO screens would be located by randomly entering the index at BLO and the 288 OLB screens would be located by randomly entering the index at OLB. The output of this program would then become the input for the product of sums list processor. (These two programs have been written for the IBM 7040 as part of experimental apparatus to be used by Richard Haber in his doctoral dissertation at the University of Pennsylvania.)

The automated list processor is also helpful when two or more long lists are to be intersected or merged. For example, the benzene ring [ERP = 6/C6(R)] has more than 10,000 postings, and the chlorine substitution (*C/G) has 1512 postings in a test file of 20,000 compounds. The manual intersection of these two lists would take about an hour, while the computer takes about 5 seconds.

ACKNOWLEDGMENT

The author acknowledges the major participation of Morris Plotkin in the refinement of the TSS algorithms and techniques for its application. Plotkin was also involved in the more extended analysis upon which the work in this paper is based. The systems and programming work of Joseph Willson and Norman London is also gratefully acknowledged. The discussion, consultations, and encouragement from the NSF sponsor, particularly Sarah Rhodes, Paul Olejar, and Tom Quigley was greatly appreciated, as were consultations with Anthony Petrarca and Don Rule of Chemical Abstracts Service and W. Lysle Alderson of Dupont. Others who cooperated in the analysis were Henry Kissman, Bruno Vasta, Alan Gelberg, and Mel Spann of the Food and Drug Administration and Edward Arnett of the University of Pittsburgh.

LITERATURE CITED

(1) Lefkovitz, D., "Use of a Nonunique Notation in a Large-Scale Chemical Information System,"J. CHEM. Doc. **7**, 192 (1967).

(2) Lefkovitz, D., "A Chemical Notation and Code for Computer Manipulation," J. CHEM. Doc. **7**, 186 (1967).

(3) Lefkovitz, D., "Substructure Search in the MCC System," J. CHEM. Doc. **8**, 166 (1968).

(4) Litofsky, B., "Utility of Automatic Classification System for Information Storage and Retrieval, Dissertation in Computer and Information Sciences," University of Pennsylvania, Pittsburgh Pa., 1969.

(5) Lefkovitz, D., "File Structures for On-Line Systems," Appendix B, Spartan Book Co., New York, 1969.

(6) Lefkovitz, D., and M. Plotkin, "An Evaluation of the MCC Topological Search System," Tech. Rept. to NSF, Contract **NSF-C547**, August 1969.

(7) Hyde, E., "Computer Generated Open Ended Fragment Code," *Proceedings of the Wiswesser Line Notation Meeting*, J. P. Mitchell, Ed., **EASP 400-8** (**AD 665 397**), p. 57–67 1968.

(8) Granito, C. E., "A Method of Analyzing Structural Fragments Using the Wiswesser Chemical Line Notation," Division of Chemical Literature, 155th Meeting, ACS, San Francisco, Calif., March 31–April 5, 1968.

(9) Bonnett, H. T., "Use of the Wiswesser Line Notation at the Searle Laboratories: Motivation and Status," *Proceedings of the Wiswesser Line Notation Meeting*, J. P. Mitchell, Ed., **EASP 400-8** (**AD 665 397**), p. 15–23, 1968.