See the example.) Next delete all the bonds that were incident to l in the original graph G, and we get subgraphs A, B, ⋯ F. The value of Z for G is given by

$$Z_G = Z_L Z_M \quad + \quad Z_A Z_B Z_C Z_D Z_E Z_{F'} \qquad (2)$$

and is shown to be independent of the choice of bond l. For graph 1 or 2 let us choose bond 2—6 (asterisked) for deletion as in Figure 5, where the procedure for obtaining Z is illustrated. The Z values for smaller graphs are taken from Tables I and II.

Relations of $p(G,k)$ and Z with the characteristic polynomial $P(X)$ are discussed elsewhere.[4,5]

ACKNOWLEDGMENT

The author is indebted to Kenzo Hirayama of Fuji Photo Film Co. for arousing his interest in this problem and esepcially in the preparation of Figure 1. Shizuo Fujiwara and Takeo Yamamoto of the University of Tokyo are gratefully acknowledged for their encouragement and discussion.

LITERATURE CITED

(1) Balaban, A.T., and Harary, F., "The Characteristic Polynomial Does Not Uniquely Determine the Topology of a Molecule," *J. Chem. Doc.* **11,** 258–9 (1971).

(2) Frome, J., and O'Day, P.T., "A General Chemical Compound Code Sheet Format," *Ibid.*, **4,** 33–42 (1964).

(3) Gluck, D.J., "A Chemical Structure Storage and Search System Developed at Du Pont," *Ibid.*, **5,** 43–51 (1965).

(4) Hosoya, H., "Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons," *Bull. Chem. Soc. Japan* **44,** 2332–9 (1971).

(5) Hosoya, H., "Graphical Enumeration of the Coefficients of the Hückel Molecular Orbitals," *Theor. Chim. Acta* **25,** 215–22 (1972).

(6) Hosoya, H., Kawasaki, K., and Mizutani, K., "Topological Index and Thermodynamic Properties. I. Empirical Rules on the Boiling Point of Saturated Hydrocarbons," submitted to *Bull. Chem. Soc. Japan.*

(7) Kawasaki, K., Mizutani, K., and Hosoya, H., "Tables of Non-Adjacent Numbers, Characteristic Polynomials and Topological Indices. II. Mono- and Bicyclic Graphs," *Natural Science Report, Ochanomizu Univ.* **22,** 181–214 (1971).

(8) Mizutani, K., Kawasaki, K., and Hosoya, H., "Tables of Non-Adjacent Numbers, Characteristic Polynomials and Topological Indices. I. Tree Graphs," *Ibid.* **22,** 39–58 (1971).

(9) Morgan, H.L., "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service," *J. Chem. Doc.* **5,** 107–13 (1965).

(10) Patterson, A.M.P., Capell, L.T., and Walker, D.F., "Ring Index," 2nd ed., American Chemical Society, 1960.

(11) Plotkin, M., "Mathematical Basis of Ring-Finding Algorithms in CIDS," *J. Chem. Doc.* **11,** 60–3 (1971).

(12) Smith, E.G. ed., "The Wiswesser Line Formula Chemical Notation," McGraw-Hill Book Co., New York, 1968.

(13) Spann, M.L., and Willis, D.D., "A Comparative Study of a Fragmentation vs. a Topological Coding System in Chemical Substructure Searching," *J. Chem. Doc.* **11,** 43–7 (1971).

(14) Spialter, L., "The Atom Connectivity Matrix (ACM) and its Characteristic Polynomial (ACMP): A New Computer-Oriented Chemical Nomenclature," *J. Amer. Chem. Soc.* **85,** 2012–13 (1963); *J. Chem. Doc.* **4,** 261–74 (1964).

Search of CA Registry (1.25 Million Compounds) with the Topological Screens System

MARGARET MILNE,* DAVID LEFKOVITZ, HELEN HILL, and RUTH POWERS
Office of Engineering Research, University of Pennsylvania,
Philadelphia, Pa. 19104

Received June 2, 1972

The TSS (Topological Screens System) for substructure search was applied to the CAS Registry file of 1.25 million compounds, making it searchable on-line. The TSS screens and the use of the screen indexes are described. Statistics on screen assignment are provided, and the strengths and weaknesses of the system in general and in particular for a large file are discussed.

The Topological Screens System (TSS) is a screening system for substructure search which was developed at the University of Pennsylvania under NSF support. The system has been applied to the Chemical Abstracts Service 1.25 million compound Registry File making it searchable on-line. This application was carried out under the U.S. Army Chemical Information and Data Systems (CIDS) contract with the University and used a version of the CIDS on-line search system which allowed single terminal access of the file. This paper describes the TSS as applied to the CAS file, and relates preliminary experiences in its search.

The TSS was applied to the CAS Registry file using an IBM 7040. Because of limited disk storage capacity (one 1301 disk module), the CAS registry file was divided into two parts. The first contained the compounds with CAS

*To whom correspondence should be addressed.

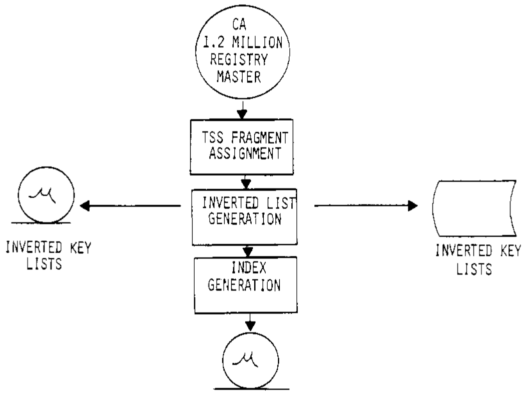


Figure 1. Generation of a searchable file from the 1.25 million compound CA Registry file

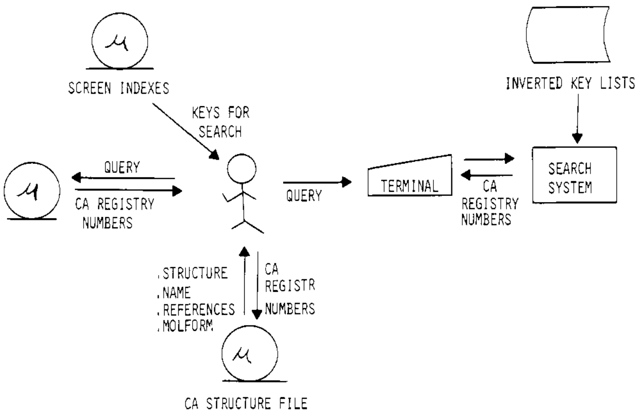


Figure 2. Search of the TSS Indexes for the 1.25 million compound CA Registry file

registry numbers through 13,743,492; the second contained those from 13,743,505 on. The two halves are treated as separate files and must be searched independently. The description that follows applies to each half of the file.

SYSTEM OVERVIEW

Figure 1 is a diagram of the screening of the compounds and their organization for search. The CAS connection tables for the 1.2 million compounds were first reformatted for the screening programs. The TSS screens (described later) were assigned, and inverted lists for the screens were generated. (An inverted list for a screen, or *keylist* is a list of all of the compounds in the file that have been assigned that screen.) The keylists were stored both on random-access computer storage and on microfilm. A KLIC (key letter in context) index of the screens was generated and stored on microfilm. The index also contains an "idiot number" that identifies the keylist for each screen.

Figure 2 diagrams search of the file, which may be carried out either manually or by machine. The user consults the microfilm indexes to identify the screens appropriate to his query, and extracts from the microfilm the idiot numbers of the corresponding keylists. In manual search, he uses these idiot numbers to access directly the microfilmed keylists containing the CAS Registry Numbers of compounds assigned the screen. In automated search, any boolean combination of idiot numbers is input from a terminal. The search programs retrieve the keylists from random access storage, carry out the stipulated intersections, merges, and negations of lists, and output the CAS registry numbers of responses. The registry numbers can be used to enter a microfilm file containing full structural diagrams of the compounds in registry number sequence.

DESCRIPTION OF THE SCREENS

Four kinds of screens are used in the TSS:

Acyclic screens and

Acyclic subscreens, both used for search of acyclic portions of compounds

Elementary ring population (ERP) screens for search of individual (isolated or imbedded) rings

Skeleton molecular formula (SMF) screens for search of complete ring systems

Two characteristics regarding these screens should be noted, namely

They are graph-exhaustive, that is, every atom in the compound graph is included in a fragment.

They are analytically assigned, that is, they are assigned according to an algorithm and not from a pre-defined list of screens. Thus the set of screens is open-ended and at any time directly reflects the contents of the file to which they are assigned.

Use of MCC Symbols in Acyclic Screens and Subscreens. Each Acyclic Screen and Acyclic Subscreen in TSS corresponds to a specific topologically defined region of a compound. In the (automated) process of screen generation, every region of the compound that satisfies the topological definition of an Acyclic Screen becomes an acyclic screen for that compound, and likewise for Acyclic Subscreens.

The first step in assigning TSS acyclic screens and subscreens is to replace certain atom symbols and combinations of atoms with their Mechanical Chemical Code (MCC) symbols. The atom symbols and atom combinations along with their MCC equivalents are given in Table I. Figure 3 depicts the replacement of standard element symbols of a structure with the corresponding MCC symbols. The Mechanical Chemical Code (MCC) is a nonunique but unambiguous linear notation for representing chemical structures. The code consists of symbols for representing the atoms of a structure, plus a set of rules for citing the atoms so that their connectivity is indicated. For representing certain atoms or groups of atoms within the molecule, special symbols—namely, those defined in Table I—are used; all other atoms are represented by their standard element symbols. It should be emphasized that the TSS screening system being described here uses the MCC *symbols* but does not involve

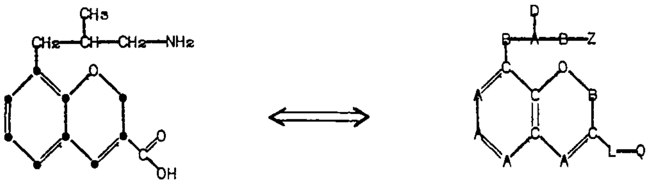


Figure 3. Replacement of standard chemical symbols with MCC symbols

the MCC rules for citing the symbols for a complete compound. A detailed description of MCC is given in D. Lefkovitz, *J. Chem. Doc.* **7,** 186 (1967).

Critical to the definition of the TSS acyclic screens and subscreens is the concept of a central or branching atom. A *central atom* is defined as (1) any nonring MCC symbol with three or more non-H attachments; or (2) any ring symbol attached to two or more acyclic bonds, where L, +JX and +SX (representing $-\overset{O}{\overset{/\!/}{\mathrm{C}}}-$, $-\overset{O}{\overset{/\!/}{\mathrm{N}}}{=}\mathrm{O}$ and $-\underset{\backslash\!\backslash\, O}{\overset{O}{\overset{/\!/}{\mathrm{S}}}}-$ respectively), are considered nonbranching atoms.

Assignment of Acyclic Screens. A TSS acyclic screen consists of a central atom as defined above plus all of the branches or "rays" emanating from it until one of the following is reached:

(1) A terminal atom, i.e., an atom with exactly one non-H attachment

(2) Another branching atom not in any ring. Two cases are distinguished. If the other branching atom is not bonded directly to the central atom, it is not represented in the acyclic screen, and the string of symbols representing the branch ends in a dollar sign ($) to show that the next atom, not cited, is a branching atom. If the other branching atom is bonded directly to the central atom, its symbol does appear in the acyclic screen, preceded by $.

(3) A ring atom. The ray in this case includes the ring atom preceded by an asterisk (*).

Example 1: The structure from Figure 3 above has one central atom (arrow) and hence one acyclic screen (circled).

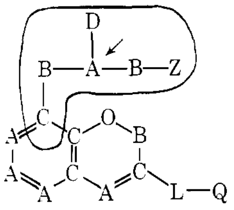
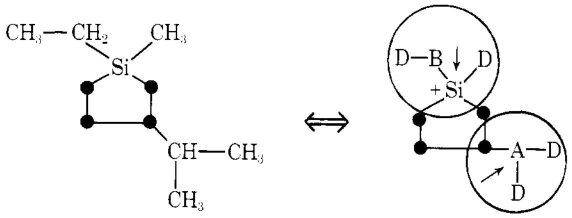


Table I. Atom Symbols and Atom Combinations Represented by MCC Symbols in the TSS Screens

MCC Symbol	Atom or Combination of Atoms
A	$-CH<$
B	$-CH_2-$
D	$-CH_3$
E	—Br
G	—Cl
J	any pentavalent nitrogen
L	$>C=O$ where C is not in a ring
M	$>NH$
Q	—OH
Z	$-NH_2$
+JX	$-NO_2$
+SX	$-SO_2$
Two letter element symbol preceded by "+"	All two-letter element symbols

Example 2: The structure



has two central atoms (arrows) and two acyclic screens (circled).

In writing out the acyclic screens, the rules below are followed:

(1) Cite the central atom first, followed by the branches in lexocographic order, each branch, being preceded by a slash and cited in the direction moving away from the central atom. Thus the acyclic screen for the structure in Example 1 is written

A/B*C/BZ/D

(2) Repeated symbols within a branch are represented by means of the symbol plus a subscript. Thus the *n*-butyl branch

$-CH_2CH_2CH_2-CH_3 \leftrightarrow -B-B-B-D$

is represented . . . $/B_3D$ . . .

(3) Multiple occurrences of the same branch are indicated by citing the branch followed by a hyphen followed by a multiplier. Thus the acyclic screen for the isopropyl group in Example 2 is written

A/*C/D-2

where "/D-2" represents the two $-CH_3$ groups.

(4) A central atom which is a ring atom is preceded by an asterisk (*). Thus Example 2 above contains the acyclic screen

*–SI/BD/D

In assigning acyclic screens bond multiplicities, charges and isotopic labels are ignored. Except as implied by the symbols A, B, D, M, Q, and Z, hydrogen is ignored unless the CAS connection table shows a hydrogen atom to be bonded to two or more other atoms, in which case it is exhibited like any other atom, using the symbol H.

An acyclic screen may represent an entire compound. For example A/D/LBD/MZ is assigned to the structure of Figure 4 (CAS R.N. 3,990,203).

Assignment of Acyclic Subscreens. An acyclic subscreen consists of a central atom plus one of its rays or branches. Thus Example 1 above, which contains the acyclic screen

A/B*C/BZ/D

is also assigned the three acyclic subscreens

A/B*C
A/BZ
A/D

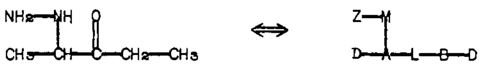
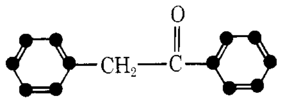


Figure 4. Illustration of a complete compound (CAS R.N. 3, 990,203) encompassed by a single acyclic screen

In addition to this, there are four other kinds of subscreens:

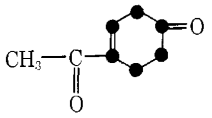
(1) A compound containing no rings and no branching atoms is described by a single subscreen, in which the atom symbols are listed in sequence beginning with either end. For example the acyclic subscreen GBLOD represents the compound $BrCH_2COOCH_3$, CAS No. 96,322. No slash-marks are used in whole-compound subscreens.

(2) An acyclic link between two ring atoms, neither of which is the central atom of an acyclic screen, containing no branching atoms (or no atoms at all, if the ring atoms are joined directly by an acyclic bond) is described by *two* acyclic subscreens. For example the compound



has the acyclic subscreens *C/LB*C and *C/BL*C. The slash-mark is used after the first ring atom. If the two rings had been bonded directly without intervening atoms, the subscreens(s) would have been *C/*C, seen twice by the programs that determine the acyclic subscreens.

(3) An unbranched attachment to a ring atom is described by a subscreen consisting of the asterisked ring atom followed by a slash followed by the unbranched substituent. The structure



has the acyclic subscreens *C/O and *C/LD.

(4) Any of the above three kinds of acyclic subscreens that would be too long to fit in the microfilm indexes are broken in pieces that are considered distinct acyclic subscreens.

Assignment of ERP Screens. The ERP screens specify the total number of atoms, the element types and counts, and the type of bonding in the primary rings of a compound. The rings keyed may be isolated (as the C_6 ring of benzene) or fused (as the C_5N ring of quinoline). (By primary rings are meant those rings contained in the smallest set of smallest rings (SSSR) that describes a complete ring system. (The SSSR for naphthalene is C_6-C_6.) For some ring systems, notably certain bridged systems, more than one SSSR can be defined.)

The MCC symbols A, B, D, J, L, M, Q, Z, +JX, and +SX are not used in expressing the elementary ring population. All carbon atoms are C and all nitrogen atoms are N. However, the symbols E for bromine and G for chlorine are used, together with the MCC representations +BO for boron, +SI for silicon, etc.

Figure 5 illustrates the assignment of ERP screens to a sample compound. The compound has two primary rings, hence is assigned two ERP screens. Each screen consists of the number of atoms in the ring followed by a slash followed by the ring formula followed by the sum of the double and alternating bond in the ring. An alternating bond is any bond in a *closed* path of alternating single and double bonds. For example, the benzene ring contains six alternating bonds.

Assignment of SMF (Skeleton Molecular Formula) Screens. The SMF screens tag characteristics of complete ring systems, that is, clusters of rings in the compound to which no additional rings are fused. The characteristics of each ring system tagged by each screen are: the number of rings in the system, the total formula of the ring system, the number of double bonds, and the number of alternating bonds.

One SMF screen containing all four of the above characteristics is assigned to each total ring system in the compound. Figure 6 illustrates the assignment of SMF screens to a sample compound.

TSS SEARCH: THE SCREEN INDEXES

As illustrated in Figure 2, search in TSS is conducted by use of the screen indexes and the key lists. The screen indexes are ordered lists of all the screens assigned to the entire file. Four such indexes (one for each screen type) are used. Associated with each screen in the index is an "idiot number" referencing the key list for the screen. The key list is a list of the registry numbers of the compounds assigned that screen, along with the number of times the screen is assigned to each compound.

The ordering of screens in the screen indexes is a vital aspect of the total system. Most of the TSS keys are precise. Often, query structures do not specify all of the information contained in a screen. Thus, a query might ask for a ring of a particular size and containing at least two nitrogen atoms. Generation of all possible ERP keys that might describe such rings, or perusing an alphabetic list of screens for such an inquiry is inconvenient because the screens that would be responsive to such a question are

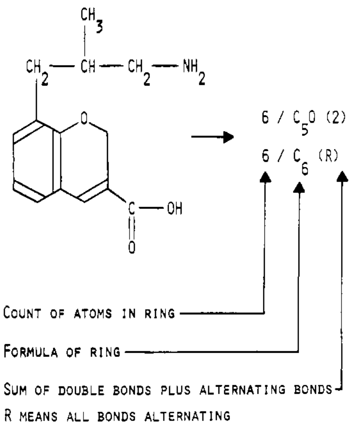


Figure 5. Assignment of ERP individual ring screens

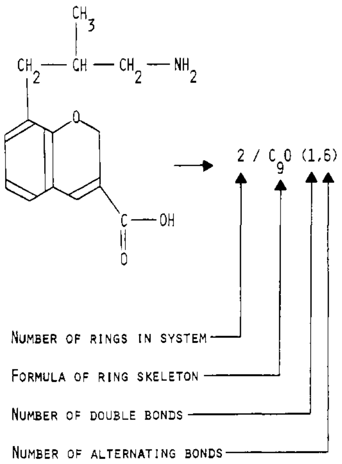
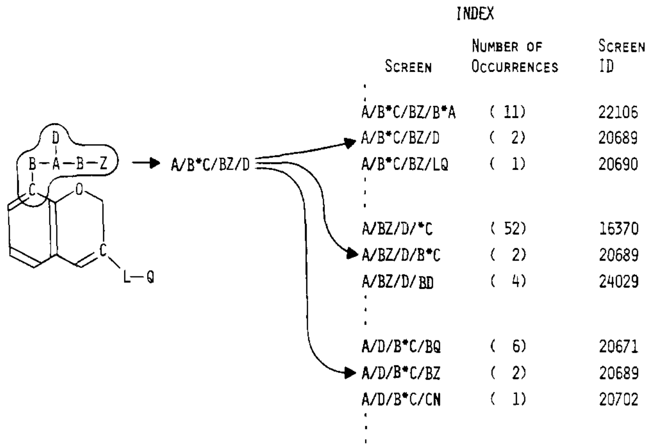


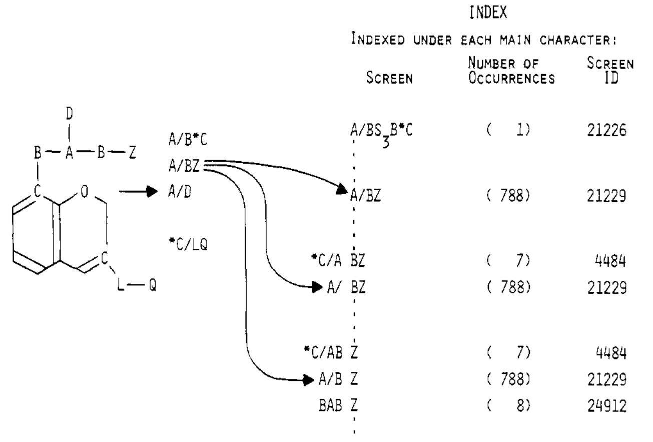
Figure 6. Assignment of SMF complete ring system screens



INDEX

Screen	Number of Occurrences	Screen ID
A/B*C/BZ/B*A	(11)	22106
A/B*C/BZ/D	(2)	20689
A/B*C/BZ/LQ	(1)	20690
A/BZ/D/*C	(52)	16370
A/BZ/D/B*C	(2)	20689
A/BZ/D/BD	(4)	24029
A/D/B*C/BQ	(6)	20671
A/D/B*C/BZ	(2)	20689
A/D/B*C/CN	(1)	20702

Figure 7. Indexing of an acyclic screen

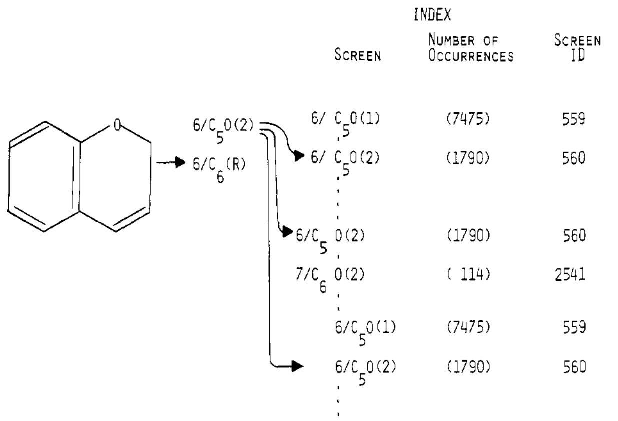


INDEX

INDEXED UNDER EACH MAIN CHARACTER:

Screen	Number of Occurrences	Screen ID
A/BS$_3$B*C	(1)	21226
A/BZ	(788)	21229
*C/A BZ	(7)	4484
A/ BZ	(788)	21229
*C/AB Z	(7)	4484
A/B Z	(788)	21229
BAB Z	(8)	24912

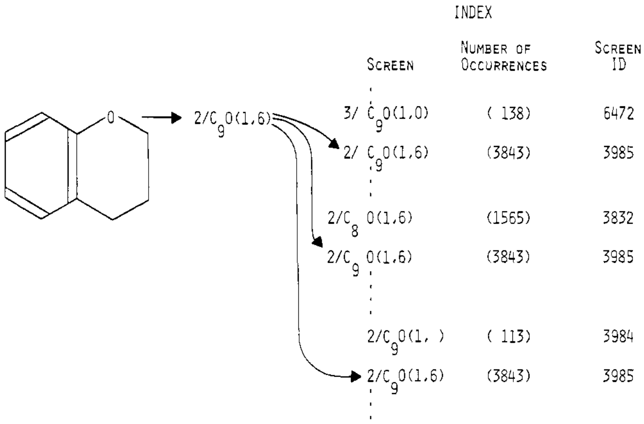
Figure 8. Indexing of acyclic subscreens



INDEX

Screen	Number of Occurrences	Screen ID
6/ $C_5O(1)$	(7475)	559
6/ $C_5O(2)$	(1790)	560
$6/C_5$ O(2)	(1790)	560
$7/C_6$ O(2)	(114)	2541
$6/C_5O(1)$	(7475)	559
$6/C_5O(2)$	(1790)	560

Figure 9. Indexing an ERP screen



INDEX

Screen	Number of Occurrences	Screen ID
3/ $C_9O(1,0)$	(138)	6472
2/ $C_9O(1,6)$	(3843)	3985
$2/C_8$ O(1,6)	(1565)	3832
$2/C_9$ O(1,6)	(3843)	3985
$2/C_9O(1,\ )$	(113)	3984
$2/C_9O(1,6)$	(3843)	3985

Figure 10. Indexing an SMF screen

scattered throughout the list. Therefore, rotated KLIC (key letter in context) indexes, containing multiple entries for each screen, are used for all four screen types, as described below.

The acyclic screens are indexed by central atom and then by branch, with each branch listed first in one entry as in Figure 7.

The acyclic subscreen index has an entry under the central atom and under each atom in the branch, as illustrated in Figure 8. In addition, an entry is made under "*" (thus grouping ring atoms) and "+" (grouping two letter atom symbols, mostly metals), if either is present.

The ERP screens are indexed under the number of rings and under each of the element symbols in the ring formula, as illustrated in Figure 9.

The SMF keys are indexed under the number of rings in the system and under each of the element symbols in the formula of the ring system. Figure 10 gives an example of the indexing of an SMF key.

TSS SCREEN ASSIGNMENT STATISTICS FOR 1.25 MILLION COMPOUNDS

Table II gives statistics on the number of screens per compound and the total number of screens of each type that resulted from application of the TSS to the 1.25 mil-

Table II. TSS Screen Assignment Statistics for 1.2 Million Compounds

(Total number of compounds, 1,256,337)

	ACYCLIC SCREENS	ACYCLIC SUB-SCREENS	ERP	SMF
Total number of different screens[a]				
First Half (625,385 cpds.)	95,834	45,864	2943	13,426
Second Half (630,952 cpds.)	109,650	48,549	5439	17,773
Total File (estimated)	150,000	67,500	7500	25,000
Screens per compound	1.03	6.75	2.35	1.53
Total Screens/compound		11.66		
Median number compounds per screen	1	2	2	2
Average number compounds per screen	6	90	400	70

[a] The available storage (one 1301 disk module) was insufficient to hold all of the key lists for the entire file simultaneously. Therefore, the total file was broken approximately in half and the two halves processed independently. Table II gives statistics for the two halves, along with estimated statistics for the full file. The estimates are based on a sampling of both halves to determine the percent of screens of each type that are common to both halves.

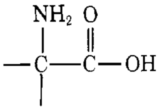
lion compound CAS Registry file. Roughly a quarter-million different screens were generated. The very low median number of compounds per screen is indicative of the very high precision of the TSS screens.

UTILITY OF TSS

The TSS offers a number of advantages as a substructure search method in general and in particular for application to the large CAS data base, as follows:

1. For the most part, the screens are precise. Since only a small number of compounds are assigned each screen, an inverted-list file structure and interactive, on-line search is reasonable, even for file sizes of the same order of magnitude as the total number of compounds currently known.

Moreover, precise screens allow the characteristics of query structures to be phrased exactly, rather than as a combination of generic characteristics. This exactness helps to reduce the number of false drops to screen search. Thus for example, in searching for the characteristic fragment of α-amino acids



both the amino group and the carboxyl group, and their exact displacement are all contained in the same TSS fragment. Many less precise screening systems must request the two groups independently, with no indication of their displacement, thereby producing many false drops. In addition, the CAS structuring conventions are well suited to topological screens since these conventions favor explicit display of bonding as in, for example, coordination compounds and inorganics.

2. Owing to their topological character, the set of screens assigned to a file parallels the file contents. This is particularly useful in dealing with the CAS file, which includes some rather esoteric structures.

3. The organization of the TSS indexes encourages off-line browsing and permits manual search for some questions even on very large files.

The indexes, in combination with the topological nature of the screens, effectively lay out before the user the variety of environments in which a given substructure occurs in file compounds. Serendipitous discovery of useful but unexpected variations on the original substructure are encountered which without browsing would have been missed.

In many cases, users of substructure search systems have no accurate method for estimating the response to a given search. As a result, queries are asked which are hopelessly broad or needlessly narrow with respect to the file searched, and which must subsequently be refined in repeated runs until a suitable response is achieved. The problem is exaggerated with a large, broad-based file such as the Registry file. The TSS indexes, which include the number of assignments of each screen, provide a basis for off-line query adjustment prior to search.

FACILITY OF TSS ON VARIOUS QUESTION TYPES

The TSS is particularly competent with query structures containing ring or nonring atoms with a number of acyclic branches. Consider for example a user interested in diamminedichloroplatinum.

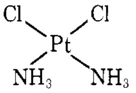


Table III gives statistics on some platinum compounds which might be useful in comparative study with the original compound of interest. These compounds could be retrieved manually by simply looking up the indicated key-lists.

Long unbranched or moderately branched acyclic fragments are also very readily retrievable with TSS. Figure 11 illustrates such a substructure for which the screens produced 276 responses from the total file—all valid responses.

One common type of substructure inquiry, for which the TSS screens are quite effective, is the search for a specific ring system. Where the required ring system is not permitted to have additional rings fused, the SMF keys are especially useful; for very characteristic ring systems, a single or a very small set of SMF screens may correspond precisely to the requested ring system. If this is the only or the most stringent requirement of the question, the response can be obtained manually from the microfilm file of key lists. The organization of the indexes allows specification of bonding in the ring system to be optional.

Where a ring system contains a rather characteristic ring, an ERP screen may be used in conjunction with the SMF screen. For example, a request for the (isolated) ring system

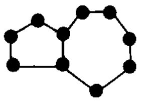


Table III. Some Statistics on Platinum Compounds

Number of different rings containing platinum	127
Number of different ring systems containing platinum	216
Number of dichloro platinum compounds	1069
Number of dibromo, diiodo and difluoro platinum compounds	499

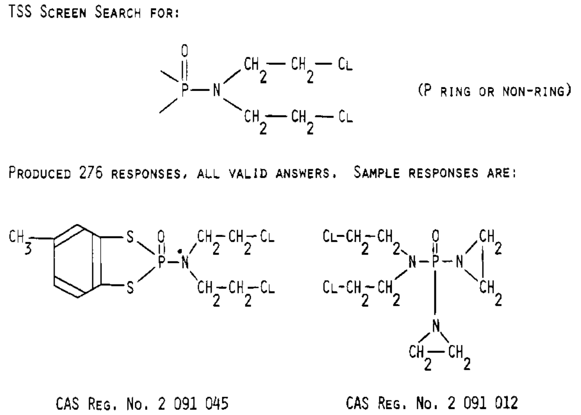
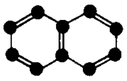


Figure 11. Search for moderately branched acyclic fragment

where any bonding is permitted in the rings would utilize the SMF screens

$$2/C_{10} \text{ (variable)}$$

Many of these screens would also be assigned to derivatives of naphthalene



which are not valid responses. If the ERP screens for C_7 rings are demanded in conjunction with the SMF screens, the majority of false drops are eliminated.

In searching for ring systems to which additional rings may be fused, the ERP screens for the individual rings in the system are used. This approach is particularly effective when one of the rings imbedded in the system is very distinctive, or if the system contains a number of less distinctive rings, the conjunction of which is very discriminating.

The TSS screens, as described above, have serious shortcomings in four characteristic inquiry areas:

Hetero and substituent positioning in a ring

Variable or don't care branching

Specification of fragment connectivity—i.e., the ability to specify that a given fragment (screen) is directly attached to another

Nonspecification of the ring/non-ring character of one or more atoms in the inquiry

A consequence of weakness in these areas is that more generic screens must be selected to answer questions of this type, resulting in a need for an atom-by-atom (iterative) search.

John Tinker at Eastman Kodak has augmented the TSS with an additional screen type called the "ring cut code" and with an atom number association technique for handling some of these problems. In the ring cut code, the atoms of each ring are cited in a string in sequence from an arbitrary starting point, with an indication for each atom of whether or not the atom is substituted. The codes are organized in a KLIC index for search, and thus allow search of hetero and substituent positions in rings.

In the atom number association technique, the connection table atom number is associated with the atom citation in each of the TSS screens so that, at search time, an overlap requirement between or among screens can be specified. Two screens are considered in this way as overlapping if they have at least one atom number in common; furthermore, the precise atom of overlap can also be specified. This technique allows overlap of screens to be examined without resorting to full atom-by-atom search. It does require that the atom-number associations be stored for each screen within a record relating to the compound.

SUMMARY

In summary, the following may be said about the TSS as applied to the 1.2 million compound CAS file

The specificity of screens is conducive to on-line operation on large files.

The fragment indexes are useful for browsing, inquiry refinement, and some manual searching.

Algorithmic generation of screens makes them relatively economical to assign.

The ring-cut code for positioning heteroatoms and substitutions is a desirable extension on rings. The atom association number is a desirable extension in the absence of an atom-by-atom search.

Variable branching and variable ring membership in queries can present difficulties.

File update requares regeneration of KLIC indexes.

ACKNOWLEDGMENT

The authors wish to acknowledge the contributions of the U.S. Army, Edgewood Arsenal, particularly Stanley Goldberg, in assisting and supporting this application; and of the National Science Foundation who sponsored the original development of TSS. The assistance of R. Thomas Swem in preparation of the manuscript and diagrams is also acknowledged.