Substructure Search in the MCC System*

DAVID LEFKOVITZ
The Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pennsylvania

Received May 9, 1968

A Monitor system based upon computer-produced printed indexes is described in this paper. The system is intended to monitor a large scale experiment in file organization for a real-time, interactive chemical information system. It is based upon a topological screen system that ensures the inclusion of every compound atom (including H) in at least one screen assignment, that appears to be responsive to a broad range of query types, and that is amenable to random-access techniques. The Monitor system is also considered, in this paper, as an independent interim approach toward fulfilling, in an effective and economic manner, the functional requirements of a small- to medium-sized chemical information system. Thus it could be used to encode and generate structure files, to assign search screens, and to provide manual substructure search capability via microfilm or hard copy printed indexes.

The ultimate objective of the work described in this paper is the development of a system that will provide on-line, real-time access to all known chemical compounds with certain associated nonstructural data. It is envisioned that this system would function in a man/machine interactive mode so that the chemist himself could communicate readily with the automated system. This would not only greatly reduce the query-response cycle time but would add the significant dimension of browsing, which can only be performed effectively by placing the chemist in direct communication with the automated retrieval system. This objective is being pursued in close cooperation with the Chemical Abstracts Service of the American Chemical Society under the coordination of the National Science Foundation.

The approach taken by the author was first to conceive certain fundamental design principles relating to the system's functional requirements. Such concepts have been published.[1] The next step was to propose some specific solutions toward implementation of these concepts and to commence experiments that would adequately test the proposals on a sufficiently large scale. To aid in the design of the experiments and to reduce their cost, it was decided to construct a Monitor system which enables the file organization techniques intended for large scale experimentation to be tested and studied via computer-produced printed indexes prior to their implementation on the disk files. This mock or simulated system has been programmed in FORTRAN and has already provided valuable design information for the larger scale experiments.

This paper discusses the system's structure screens and their organization into indexes intended ultimately for storage in mass random access disk files; however primary emphasis is placed in the paper upon the manual Monitor system because of its present functional status.

* Presented before the Division of Chemical Literature, Symposium on Notation Systems, 155th Meeting, ACS, San Francisco, Calif., April 4, 1968.

The system is based upon computer storage of the connection table in a highly concise linear notational format called the Mechanical Chemical Code[2] and upon retrieval screens derived in terms of this notation.

An examination has been made[1] of five system functional requirements: input, registry, file storage, substructure search, and display, and a number of conclusions were drawn in that paper with regard to the representation of a structural formula in digital storage and with regard to the mode of operation in an on-line interactive system. A non-unique notational specification was then provided[2] for concise storage of the structural formula (the MCC) which is also operationally amenable to automatic screen assignment and an atom by atom search. This paper also indicated how this non-unique notation could be used to register compounds.

Whereas these former papers were addressed to the functional requirements of input, registry, and file storage, the present paper addresses itself more particularly to the fourth system requirement—substructure search—by describing a screening system that possesses the properties of (1) completely automated and economical screen assignment, (2) ease of programming, and (3) both coarse and fine screening characteristics. Furthermore, the screen system appears to be responsive to a broad range of substructure query types, although only further testing and extensive usage can ultimately make this determination.

The block diagram in Figure 1 illustrates the experimental system. Compounds from the CAS registry system are the primary data source. These are converted in Block 1 from the nested canonical connection table (CT) into the MCC for disk file storage. In Block 2, coordinates will be added to the MCC to enable display of the structure. This procedure assures uniformity of display and also enables the system to construct files from connection tables without coordinates. Thus CAS registry compounds, including those that are manually structured or automatically translated from nomenclature, will be assigned coordinates in Block 2. The compound

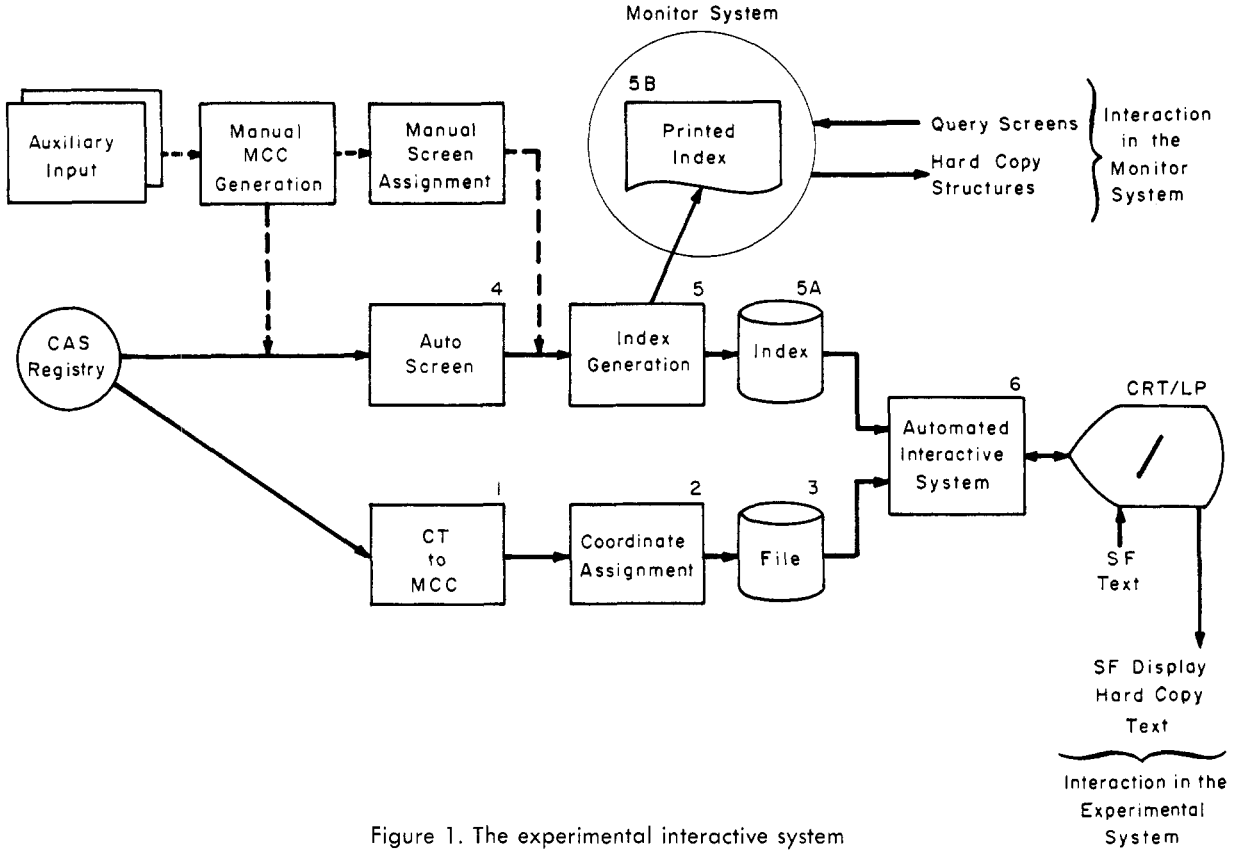


Figure 1. The experimental interactive system

file and certain associated nonstructural data will then be stored in the random access file. The CAS registry connection tables are also processed in Block 4, where the screens to be discussed in this paper are automatically assigned. In Block 5, a series of indexes are generated and stored in the disk file (5A) for on-line retrieval via the Automated Interactive System (Block 6) and the printed indexes of the Monitor system (5B).

An auxiliary input mode has also been provided which at present is being utilized for the generation and testing of small files by manual MCC and screen generation, and the trial of 4000 compounds reported in this paper was based on file composition via this route.

The experimental interaction with the larger file will be via cathode ray tube, light pen, and keyboard (CRT/LP). The chemist will enter a query in the form of a structural formula (or partial structure) and text using the light pen and keyboard. Responses will appear on the face of the CRT as structural formula displays and text.

Interaction in the Monitor system is via the printed indexes where structural queries are first transformed by inspection to query screens, which can be searched in the indexes. The responses from the printed indexes are CAS Registry Numbers that may be used to retrieve structures from a microfilm or hard copy file.

The primary objective of this paper is to report on the Monitor system, which has been completed, because it simulates the full scale experiment, and also because in its present form it can be used for the low cost generation of a compound file with a powerful substructure search capability via the indexes that could be stored on microfilm for files in the range of 100,000 compounds.

A sample of 4000 compounds was manually encoded into the Monitor system, and the following statistics and facts were revealed:

(1) Average number of non-H atoms per compound = 16.9
(2) Average number of MCC characters per compounds = 11.9
(3) Number of MCC characters per non-H atom = 0.7
(4) Average time to generate an MCC (after steady state) = 30 seconds
(5) Average time to generate the screens (after steady state) = 50 seconds
(6) Number of encodings to reach steady state in (4) and (5) = 300 compounds
(7) Background of coder, High school graduate
(8) Time to teach coder MCC = 6 hours
(9) Time to teach coder screen system = 4 hours

The MCC encoding either from typewriter or manual input is also subject to the parity check presented in the Appendix of a previous paper,[2] which makes it a highly reliable file building procedure. Furthermore, if the manual procedures were to be coupled with direct on-line input via typewriter keyboard, the parity check would be made immediately thus avoiding recycling controls.

THE MCC SCREEN SYSTEM

The MCC screens are at present divided into two categories, the acyclic and cyclic screens. The acyclic screens are based strictly upon the topology of the graph that represents the structure, while the cyclic screens, although topologically based upon the rings, are also

meaningful in terms of more natural classifications already in use by chemists. Some preliminary work on extending the generality of the acyclic screens to cover cyclic components has been started so that ultimately the entire structure could be screened by a common topological screen type, but the present paper is confined to this dual screening structure.

Two basic principles underlie the screen system. First, the screens map the entire structure so that every atom (including hydrogen) appears in at least one of the assigned screens. Then, according to the system design, screens assigned to a query will appear as subscreens of an existing screen or set of screens (if this substructure exists in the file), and the conjunction of this existing set of screens will point to all compounds in the file containing this substructure. Furthermore, a query may be completely arbitrary in terms of its atom connections.

The second design principle is that all screens are analytic in the sense that they are assigned to a particular file of compounds by an algorithmic or analytic procedure, so that the actual repertoire or vocabulary of screens is a function of the specific compounds in the file and not a predetermined tabulation. Hence one can expect that the variety of screen types will increase rapidly at the begining of file generation and will approach a point of saturation in number after a certain file size has been reached.

The Acyclic Screens. The acyclic screens are assigned according to the following three step procedure:

(1) Identify all *central atoms* of the structure, where a central atom is (1) any acyclic MCC branching symbol,[5] (2) a ring attachment onto which is substituted one or more straight (unbranched) chains, and (3) either end (but not both) of a straight chain compound.

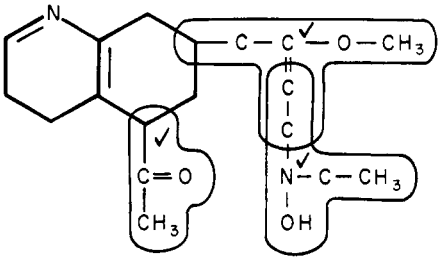
(2) Identify the *neighborhood* of each central atom as the set of MCC symbols that radiate in a straight chain from the central atom up to and including a terminal symbol, up to but not including another central atom, and up to and including a ring attachment.

(3) Write one *acyclic screen* for each central atom and its neighborhood in the format $x/\alpha/\beta/\ldots\omega$, where x is the central atom symbol and α, β, etc. are the MCC notations for the respective straight chains that emanate from x and which are included within the boundary of the neighborhood. (If the central atom is the terminal atom of an acyclic straight chain compound, the slash is omitted after the citation of the central atom.) Furthermore, to distinguish particularly ring attachments, the symbol representing the ring attachment, whether it appears as x or in α, β, etc., is prefixed by an asterisk.

The screen is then expanded into a series of subscreens of the form x/α, x/β, $\ldots x/\omega$. Two indexing systems are developed in the next section, one for the screens $(x/\alpha/\beta/\ldots/\omega)$ and one for the subscreens $(x/\alpha, x/\beta, \ldots x/\omega)$.

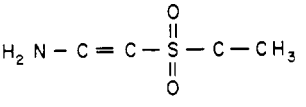
Examples 1 and 2 in Figures 2 and 3 illustrate all possible interpretations of these three generative rules.

The checked atoms in Example 1 indicate the central atoms, and the neighborhoods are defined by the indicated boundaries. In Example 2 either of the screens may be used without prejudicing the system's retrieval capability.



Acyclic Screens	Cyclic Screens
C/b*a/Oc/ab	$a\,b_2\,C_2\,N$
N/ba/bc/Q	$*\,a_2\,b_2\,C_2$
*a/Lc	

Figure 2. Screen assignments: example 1



Acyclic Screen

$Z\,a_2\,S\,X\,b\,c$

or

$c\,b\,S\,X\,a_2\,Z$

Figure 3. Screen assignments: example 2

The subscreens that are generated from the screens shown in Figures 2 and 3 are:

Screen No.	Screen	Subscreen	Subscreen No.
F1	C/b*a/Oc/ab	C/b*a	1
		C/Oc	2
		C/ab	3
F2	N/ba/bc/Q	N/ba	4
		N/bc	5
		N/Q	6
F3	*a/Lc	*a/Lc	7
F4	Za_2SXbc or $cbSXa_2Z$	Za_2SXbc or $cbSXa_2Z$	8

The screens and subscreens are also assigned a unique serial identification, as shown above to facilitate their references in the indexes.

The Cyclic Screens. The cyclic screens are mapped onto the compound nuclei and therefore overlap the acyclic screens only at the point of substituent attachment. In the present system they represent a coarser screen type than the acyclics and are essentially adaptations of standard ring classification codes,[1] commonly called the elementary ring population of each ring of the set of smallest rings in each nucleus expressed in MCC symbols.

The cyclic screen assignments of Example 1 are:

Cyclic Screen	Screen No.
ab_2C_2N	R1
$*a_2b_2C_2$	R2

A number of additional ring screens might also be considered desirable to satisfy more or less generic requirements, such as the number of nuclei, the ring counts of each nucleus, the numeric ring population of each ring in a nucleus, etc., all of which are currently being investigated in the U.S. Army CIDS; however, the system described here utilizes only the above defined screen.

THE INDEXES OF THE SYSTEM

After assignment of the acyclic and cyclic screens, two index sets are generated, one for the subscreens and the other for the screens. Each index set consists of a Rotated Index of the screen types and an Inverted Index that lists all compound references by a registry or accession number in which the given key appears. The cyclic keys appear only in the subscreen index set.

The Subscreen Index Set. The formation of a rotated index coupled to an inverted index was suggested for this type of screen by Hyde.[4] The rotated index is generated by shifting the subscreen so that every symbol of each subscreen in the system will appear in the principal index column, and the index is maintained in a sorted alphabetic sequence from the principal column rightward and subsequenced within this, leftward from the principal column. Figures 4 and 5 present the Rotated Acyclic and Cyclic Indexes, respectively, for the screens of Examples 1 and 2.

Line		↓	Subscreen No.
1	C/b	*a	1
2		*a/Lc	7
3	N/b	a	4
4	C/	ab	3
5	Z	a_2SXbc	8
6	C/a	b	3
7	C/	b*a	1
8	N/	ba	4
9	N/	bc	5
10	Za_2SX	bc	8
11	N/b	c	5
12	Za_2SXb	c	8
13	*a/L	c	7
14	C/O	c	2
15		C/ab	3
16		C/b*a	1
17		C/Oc	2
18	*a/	Lc	7
19		N/ba	4
20		N/bc	5
21		N/Q	6
22	C/	Oc	2
23	N/	Q	6
24	Za_2	SXbc	8
25		Za_2SXbc	8

Figure 4. Rotated subscreen index for acyclic screens

Line	↓	Screen No.
1	*a_2 b_2 C_2	R2
2	a b_2 C_2 N	R1
3	*a_2 b_2 C_2	R2
4	a b_2 C_2 N	R1
5	*a_2 b_2 C_2	R2
6	a b_2 C_2 N	R1
7	a b_2 C_2 N	R1

Figure 5. Rotated index for cyclic screens

If the symbol / is retained in the sort, all central atoms of a given type will be collected (for example C/, N/), and it might actually be desirable to generate a separate subindex of these. Other special considerations in the sort are (1) X in the combination αX should not appear in the principal column since retrieval would always be via α; (2) only the first letter of two-letter element symbols would appear in the principal column.

The other index in the set is an inverted list of compound reference numbers for each subscreen. That is, for each subscreen of the system, which may be identified in this index by its serial number, there appears an ordered listing of all compound registry numbers that contain the screen. In the disk stored index, the compound addresses, rather than registry numbers, would be stored in the list.

This index is, therefore, entered by one or more subscreen identification numbers, which would have been determined from a search of the Rotated Index, and it decodes to lists of compound reference numbers that can be merged or intersected, depending upon the query screen logic, as will be discussed further in the next section.

The Screen Index Set. The screen index set like the subscreen set contains a rotated and an inverted index. A portion of the rotated screen index for Example 1 appears in Figure 6. The screen of Example 2 would appear only in the subscreen index.

In terms of size, the rotated indexes will become quite large but can be expected to approach saturation as the file size increases, while the inverted compound reference indexes will increase linearly with the file size; however the inverted screen lists will be relatively short while the inverted subscreen lists will tend to be relatively long. Experiments currently in process will provide statistics for a file of 100,000 CAS Registry compounds.

For the 4000-compound test file the ratio of file size to subscreen vocabulary size was 0.97 when the file had 300 compounds, 0.72 at 700 compounds, 0.51 at 2000, and 0.42 at 4000 compounds. The corresponding ratios for the cyclic screens were 0.15, 0.12, 0.11, and 0.08, respectively. The average number of cyclic and acyclic screens (not subscreens) assigned per compound in the test file was 6.5.

SUBSTRUCTURE SEARCH STRATEGIES

There are two basic search strategies that are employed with these indexes, each addressable to one of the two index sets described above. These strategies, called I and II, are characterized by Figures 7 and 8.

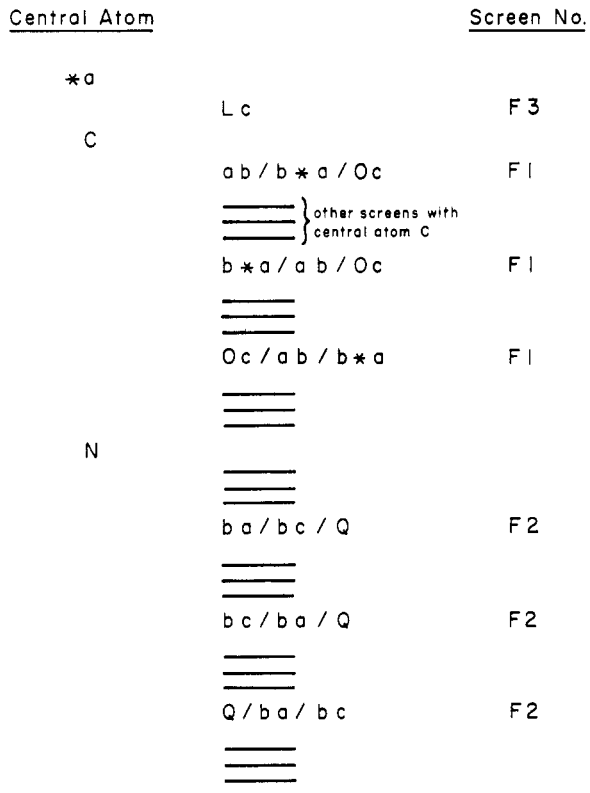


Figure 6. Rotated screen index

Strategy I is characterized by having no more than a single chain emanating from a central atom of the query or by having no identifiable central atom, although, as indicated by Figure 7e, there may be more than one central atom in the query. Strategy II is characterized by having two or more chains emanating from a central atom, and again there may be more than one central atom (Figure 8b).

The indexing sequence for Strategy I is shown in Figure 9, and the sequence for Strategy II is shown in Figure 10.

These procedures are now illustrated by applying the queries of Figures 7 and 8 to the Examples of Figures 2 and 3.

Query 7a would be screened as C/a. The Rotated Index of Figure 4 could then be entered via C/ or a. In general, it would be preferable to enter by the symbol with fewest principal column entries, and a central atom is usually easier to scan since the eye only has to move to the right. The disk stored index on the other hand, would be entered by a standard random access decoding tree, and criteria such as tree depth (also related to least frequent occurrence) would be used. The query screen C/a could thus be located either on line 4 (entry by a) or on line 15 (entry by C/), and in either case the File subscreen appears as No. 3 (C/ab). If the Inverted Index on subscreens were then entered with No. 3, all compound references containing the screen C/ab would be produced.

Query 7b is screened as ab or ba. Both searches must be performed because it cannot be determined at this

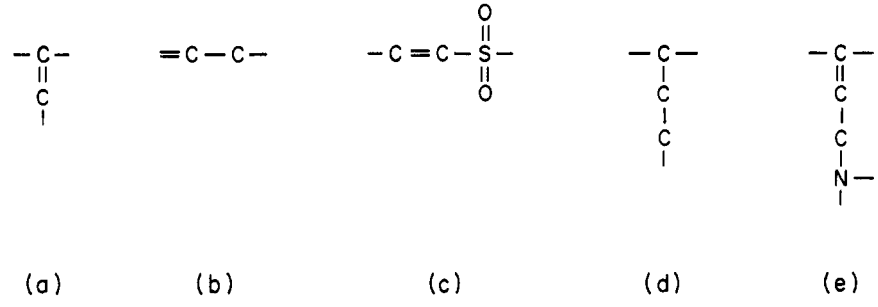


Figure 7. Examples of search strategy I

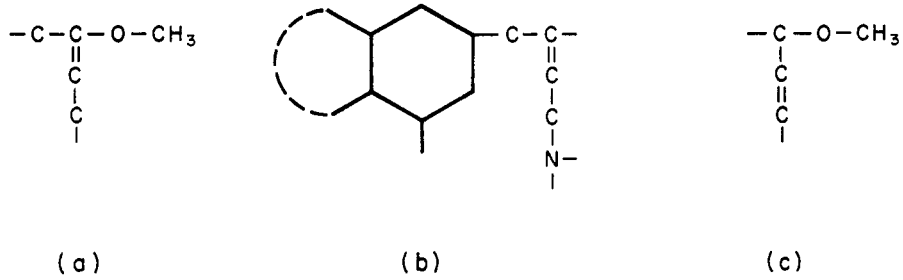


Figure 8. Examples of search strategy II

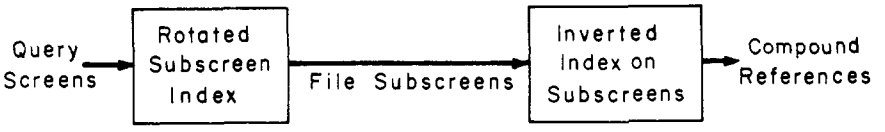


Figure 9. Index sequence for search strategy I

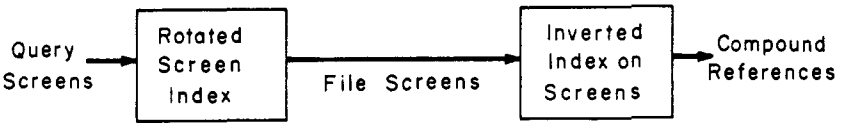


Figure 10. Index sequence for search strategy II

point whether a central atom lies to the a or b side. (Central atoms could, of course, lie to both sides in which case the compound would be retrieved by either query screen.) The Rotated Index search produces a hit for ab on line 4 or 6 (Subscreen 3) and for ba on line 3 or 8 (Subscreen 4). The total response to this query would then be obtained by merging the respective inverted lists of subscreens 3 and 4. In the case of Example 1, this compound would be found on both inverted lists.

Query 7c is screened as a_2SX or SXa_2 and is located on line 24 by entry via SX. Note that there is no entry for SXa_2 since the compound screen (Example 2) was assigned only in the direction a_2SX, which is completely adequate since we require that the query be asked in both directions.

Query 7d is screened as a/b_2, and since there is no a/ entry neither of the example compounds responds. In the actual index for a real file, if a subscreen cannot be located, then a definite null response to the query is indicated.

Query 7e is screened as a logical conjunction of (C/ab) AND (N/ba). This means that both of the screens must be found in the rotated index and the resulting inverted lists must be intersected. However, since each query screen could produce more than one file subscreen, the procedure would be first to merge the lists generated by the individual query screens—i.e., C/ab generates a merged list and N/ba generates a merged list—and then to intersect these two resultant merged lists.[5] In the index, C/ab is located via C/ on line 13 (subscreen 3) and N/ba is located via N/ on line 17 (subscreen 4). The intersection of the subscreen 3 and 4 lists would then produce the compound of Example 1.

Two remarks should be made, however, at this point. First, this conjunction does not require (as search strategy II does) that these two screens actually overlap. That is, the compound of Figure 11 would respond to this query and could only be further qualified by visual inspection or by an atom by atom search.

The second remark is that this query illustrates the fact that any arbitrary connected acyclic substructure or set of substructures will respond to this key system, because this query does not lie wholly within either of

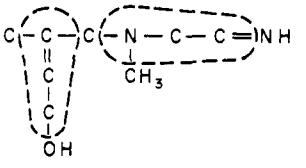


Figure 11. A false drop to query 7e

the responding subscreens, thus demonstrating the independence of query substructures and compound screen assignments.

Strategy II is applied to the queries of Figure 8, where in every case at least one central atom appears with two or more branches. This type of query may therefore be addressed to the Screen Rotated Index (Figure 6). The query screen of Figure 8a is C/ab/b/Oc, which is located as screen F1, indexed within the central atom Csection under, b*a or Oc. The Inverted Screen Index would then be entered with F1, indexed within the central atom C section under ab, b*a, or Oc. The Inverted Screen Index would then be entered with F1 to produce the relevant registry numbers. Query 8b is screened with acyclics C/ab/b*a AND N/ba AND cyclic $*a_2a_2b_2$. The two acyclics are found in Figure 6, but the cyclic is not found in Figure 5, the Rotated Cyclic Index. This means that no compound in the file satisfies this query; however, $*a_2b_2C_2$ is found, which indicates that this query with somewhat less saturation might have a response.

Finally, query 8c would be screened as $a/a_2/Oc$, which does not appear in the Rotated Index of Figure 6, and which therefore means that the compound is not in the file.

CONCLUSIONS

A technique has been developed for monitoring a large scale file organization experiment involving a minimum of 100,000 compounds to project on-line system feasibility for 3 million compounds. Such feasibility depends upon the ability to devise a screen system that will partition the files for list structured, random access retrieval, where the lists, or intersections of these lists, produce few enough

Table 1. Advantages of the MCC System

System Function	Advantage of the MCC System
Input	A. Manual Input The code is non-canonical and mechanical; therefore, the entire set of rules can be taught in a few hours to a clerical non chemist.[a] Coding is rapid and reliable. Code verification is automatic by built in MCC parity check. B. Connection Table Input The CT to MCC translation is readily performed. A FORTRAN program now exists. Code verification (MCC parity) can be used to validate the CT.
Registry	The MCC generated from the CAS canonical CT is also canonical, and therefore can be used for registry. Alternatively, the CMF, [2] which is readily generated from the MCC, can be used for isomer sort registry The CMF generation program in FORTRAN now exists.
File Storage	The average number of MCC characters per non hydrogen atom of the compound is 0.7, which means that a file of 1 million compounds with an average of 25 non-hydrogen atoms per compound could be stored in 17.5 million characters, or 1 reel of 800 BPI magnetic tape, exclusive of any control information.
File Search	The screens of this system are designed to produce few false drops and yet be responsive to a broad range of query types, so that atom by atom search is generally not required. The MCC can be converted to a CT in a one pass operation for atom by atom search. A FORTRAN program now exists to make this conversion.

[a] Reference 1 demonstrates that neither registry nor search require a canonical structure representation.

query responses to minimize or eliminate the requirement for atom by atom iterative search. The Monitor system, described in this paper, facilitates the examination of retrieval strategies intended for random access disk implementation by simulation of the file indexes as computer generated print-outs. The screens themselves are single pass analytic screens which require no table look-ups or iterative searches and hence are economical to assign by machine. Also, because they are purely topological, they can be assigned manually by a clerical process to produce auxiliary test files in the Monitor system.

The Monitor system may also be viewed as an immediately useful spinoff of this research, because the addition of a registry procedure such as the coded molecular formula (CMF) isomer sort registry[2] would provide a low cost system for the generation and search of small- to medium-sized files. Such a system would include the four components of input, registry, file storage, and search, but would still rely on hard copy (or microfilm) for display.

The input would be manual via MCC generation. The screens would then be generated by the programs represented by Block 4 of Figure 1. Registry for this size file would be by a combination of CMF isomer sort and visual inspection of competing isomers. This would eliminate the need for an atom by atom search to distinguish the isomers. The search file would consist of (1) the coded structures which could be cataloged and printed as an MCC with identifying registry numbers, (2) the structural formulas on microfilm or in hard copy files, and (3) the two index sets printed by computer and reduced to microfilm. If nonstructural descriptors, test results, or applications were also assigned to the structures as a coordinate index, these would be incorporated into an inverted index that would enable the query logic to include structural screens in combination with nonstructural terms. The search function is performed manually via the indexes and the structure files. However, as the file grows, as indicated in the paper, the Inverted Indexes grow correspondingly, and they become increasingly awkward to manipulate; hence the first search component to automate as the file grows would be the Inverted Indexes, which are fairly easy to store and search on magnetic disks.[5] The next increment of growth would include disk storage of the Rotated Indexes, which necessitates a query language and an interactive system for on-line operation, this being a current area of investigation discussed in the introduction.

The specific advantages of the MCC system in terms of these four system functions are indicated in Table I.

ACKNOWLEDGMENT

The author wishes to acknowledge various kinds of assistance which have contributed and continue to contribute to these research efforts. First there is the administrative and financial support of the National Science Foundation (Contract No. NSF-C467), and the technical management and guidance of Paul Olejar, Director, Sarah Rhodes and Tom Quigley of the NSF Chemical Information Unit, Office of Scientific Information Service who have coordinated our work with that of others as well as contributing through technical discussions. Among these coordinated efforts are included those of Ronald C. Read of the University of the West Indies, whose valuable suggestion resulted in the present form of the acyclic screen and Don Rule and Anthony Petrarca of CAS. The author is indebted to University of Pennsylvania research colleagues Morris Plotkin and Richard Haber and Dr. Alfonso Gennaro of the Philadelphia College of Pharmacy. Finally, the valuable programming assistance of Rebecca Chao, Norman London, Lydia Tomasello, Nick Homer, Joseph Willson, and Malcohm Cohen is acknowledged.

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) Van Meter, C. T., D. Lefkovitz, and R. V. Powers, "CIDS No. 4, An Experimental Chemical Information and Data System," University of Pennsylvania; produced under Contract No. DA-18-035-AMC-288(A), Technical Support Directorate, U. S. Army Edgewood Arsenal, Edgewood, Md., 1967.

(4) Thomson, L. H., E. Hyde, and F. W. Matthews, "Organic Search and Display using a Connectivity Matrix Derived from Wiswesser Notation," J. CHEM. DOC. 7, 204 (1967).
(5) Lefkovitz, D., R. V. Powers, and H. N. Hill, "CIDS No. 5, Computer Programming for An Experimental Chemical Information and Data System," University of Pennsylvania; produced under Contract DA-18-035-AMC-299(A), Technical Support Directorate, U.S. Army Edgewood Arsenal, Edgewood, Md., 1967.

(6) An MCC branch is any element or combination element symbol with three or more nonhydrogen attachments. Therefore, the symbols L (-(CO)-) and $\alpha X(SO_2, NO_2)$ are not branches.

(7) These screens, without hydrogen denotation, are currently in use by the U.S. Army CIDS and have proved to be highly effective therein.[3]

(8) In logic, this is called a product of sums.

A Line-Formula Notation System for Markush Structures*

HELEN M. S. SNEED, JAMES H. TURNIPSEED, and ROBERT A. TURPIN, JR.
U. S. Patent Office, Department of Commerce, 1406 G St., N. W., Washington, D. C. 20231

Received May 21, 1968

A notation system has been developed in the U. S. Patent Office to handle some Markush forms. The system is presented as a supplement to the existing Hayward Notation System which was developed for specific organic chemical structures. The proposed notation system for organic Markush structures is limited to determinate structures of several isolated Markush forms, including those forms that are restricted in substitution depending on the condition of some other Markush group.

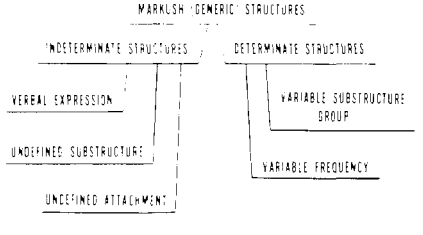
The Line Formula Notation System for Markush Structures is a supplement to a notation system developed for specific organic structures by H. Winston Hayward of the U. S. Patent Office. Efforts were made to avoid some of the rigid unique features of the Hayward system but remain within the over-all framework of the system itself.

Markush structures, as they are referred to in the Patent Office, are generic expressions of chemical structures or structure classes. The expressions can be verbal, structural, or some combination of verbal and structural statements. The Markush expressions which are in terms of structure with all points of attachment defined as *determinate* Markush structures. All other Markush expressions, namely, verbal, combinations of verbal and structural, and structures with undefined attachment, are *indeterminate* Markush structures. For convenience of notation, specific chemical structures may be reduced topologically to a graph in which each atom node (vertex) represents only one atomic structure and each bond (edge) represents only one bond type, and are further defined as a set of atom nodes and bonds such that each bond of the set is connected to two atom nodes. Likewise, determinate Markush structures can be defined in terms of a graph, reducing the Markush group to a node. A determinate Markush structure is then defined as a Markush structure in which all nodes, atom and Markush, are specifically defined by an atomic structure, or group of definite atomic structures and/or atom strings; all bonds are defined by a bond type or group of alternative bond types; and, all points of attachment are explicit.

This investigation will not venture beyond the realm of the determinate Markush structure. It is hoped, however, that the notation system for Markush structures will be modified in the near future to handle some of the indeterminate forms. The determinate Markush forms, as isolated in this report, were extracted from U. S. Patents. For Patent Office purposes, a Markush structure should be copied just as it is disclosed in the original document. This policy reduces the possibility of generating structures which may not be encompassed in the original disclosure.

ISOLATED FORMS OF MARKUSH EXPRESSIONS

To define further the forms in which Markush expressions may be disclosed, the general form may be broken down as follows:



* Presented before the Division of Chemical Literature, Symposium on Notation Systems, 155th Meeting, ACS, San Francisco, Calif., April 4, 1968. This paper reports on work advanced by Patent Office researchers as a part of that agency's continued efforts, in cooperation with the National Bureau of Standards and the National Science Foundation, to solve the problems of retrieval of information from machine-oriented systems containing generic chemical structures.