In terms of the information in the document, some of these descriptors, and certainly combinations of some of the descriptors, are likely to yield false retrievals. The Multiterms of Figure 1, on the other hand, are directly related to the information in the document and cannot lead to a false retrieval. Furthermore, the Multiterm tends to extend the searcher's ability to find documents of interest easily. For example, if we were interested in the reactions of ozone with any chemical, we need merely to scan the Multiterms in which OZONE -R is the first term for the other reactant term—e.g., BUTADIENE -R, or STYRENE -R, or VINYL CHLORIDE -R.

ADVANTAGES OF THE INFORMATION SYSTEM

Hercules interest in Government reports in the broad area of chemical propulsion results in the accession of about 130 documents (without duplication) per week at four locations. Centralization of an information system, at the Research Center, for the preparation of a weekly awareness bulletin and computer-produced indexes has resulted in a more timely and more comprehensive system at considerably less cost than the original four systems.

The awareness bulletin, which is distributed to 180 scientists and engineers at the four locations, has promoted an appreciable increase in the use of the Government report literature with a favorable feedback from the readers. Since it was initiated, the system has informed the readers of the total documents received at all locations, including those received at their own location with their own document control numbers. The system has reduced almost completely the need for many scientists and engineers to examine TAB, USGRDR, and STAR, as they did before the centralized system was set up.

A valuable product of the information system was the Multiterm indexing system which was conceived and developed in response to the problems that confronted us in taking on this challenging task.

ACKNOWLEDGMENTS

We acknowledge the advice and encouragement of R. Steinberger and the assistance of W. R. Payson, R. H. Petty, T. M. Norback, and W. G. Young who were involved in the early stages of setting up the information system.

REFERENCES

(1) Caponio, J. F., and T. L. Gillum, "Practical Aspects Concerning the Development and Use of ASTIA's Thesaurus in Information Retrieval," J. CHEM. DOC., 4, 5–8 (1964).
(2) Day, M. S., "The Scientific and Technical Information Program of the National Aeronautics and Space Administration," J. CHEM. DOC., 3, 226–8 (1963).
(3) Gillum, T. L., "Compiling a Technical Thesaurus," J. CHEM. DOC., 4, 29–32 (1964).
(4) Green, J. C., "The Role of the Department of Commerce," J. CHEM. DOC., 3, 223–6 (1963).
(5) Hicks, M. S., "Government-Sponsored Research Reports in Three Areas of Physical Chemistry," J. CHEM. DOC., 3, 144–8 (1963).
(6) Skolnik, H., "The Multiterm Index: A New Concept in Information Storage and Retrieval," J. CHEM. DOC., 10, 81–5 (1970).
(7) Skolnik, H., Book Review of NASA Thesaurus, J. CHEM. DOC., 8, 53 (1968).
(8) Skolnik, H., and W. R. Payson, "A New Posting Method for the Preparation of a Cumulative List," J. CHEM. DOC., 3, 21–4 (1963).
(9) Vann, J. O., "Defense Documentation Center (DDC) for Scientific and Technical Information," J. CHEM. DOC., 3, 220–2 (1963).
(10) Wente, V. A., and G. A., Young, "Selective Information Announcement Systems for a Large Community of Users," J. CHEM. DOC., 7, 142–7 (1967).

A Comparative Study of a Fragmentation vs. a Topological Coding System in Chemical Substructure Searching*

MELVIN L. SPANN[1]
Science Information Facility, Food and Drug Administration, 200 C St., S.W., Washington, D. C. 2020
DELORES D. WILLIS
Statistical Data Branch, Bureau of Medicine, Food and Drug Administration, Washington, D. C.

Received June 26, 1970

For the past six years, the Food and Drug Administration has been utilizing a fragmentation coding system for the storage and retrieval of chemical structure information pertaining to Investigational and New Drug Applications. After installation of the Chemical Abstracts Service Substructure Search System by FDA's Science Information Facility, a three-month comparative study was conducted using both systems for the retrieval of chemical compound information. This paper presents many of the observations made during this study with particular emphasis on the degree of specificity available in question phrasing and the precision in retrieving chemical compounds containing desired substructures.

Recognizing the need to automate the storage and retrieval of information in Investigational New Drug Applications (INDs) and New Drug Applications (NDAs), the former Bureau of Medicine of the Food and Drug Administration, in 1963, instituted an information retrieval system called RAPID (Retrieval of Automatically Processed Information on Drugs). This electronic accounting machine-based system was established to handle chemical, medical, and management data received by FDA in conjunction with drug applications. This paper is limited

* Presented at the Fifth Middle Atlantic Regional Meeting, ACS, Newark, Del., April 1, 1970.
[1] Present address, Food and Drug Administration, Bureau of Drugs, 5600 Fishers Lane, Rockville, Md. 20852

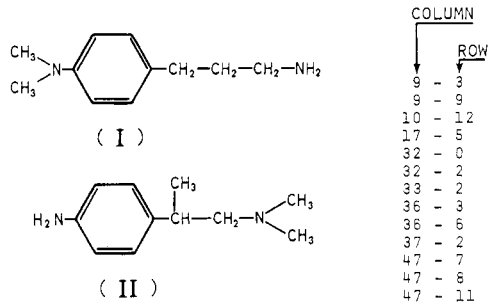


Figure 1

to the chemical structure aspects of the RAPID system.

After reviewing and analyzing existing punch-card oriented chemical coding systems, a special FDA task force decided that the General Chemical Fragmentation Code by Frome and O'Day[1] would best fulfill the requirements of the chemical retrieval component of the proposed drug information system. Thus, fragmentation was established as the Bureau of Medicine's method of chemical information handling.

Within the Bureau, use of the fragmentation method of searching for chemical structures was adequate in that total recall of most relevant chemical compounds could be achieved. However, the high degree of ambiguity characteristic of most fragmentation codes greatly affected the search results. Difficulty was encountered in uniquely defining specific chemical compounds or partial structures. Figure 1 illustrates this point.

With the fragmentation code used, compounds (I) and (II) would receive identical codes. Therefore, querying the system for parasubstituted *N,N*-dimethylaniline derivatives would cause the retrieval of both compounds, compound (II) representing a false drop. This example illustrates the inability of a fragmentation code to retrieve, in most cases, only those structures that are actually desired. Consequently, when the size of the chemical file exceeded 20,000 punch cards, this ambiguity placed an unnecessary burden on the searcher. In order to eliminate the false drops (usually 20 to 30% of the answers retrieved), a listing of the chemical names for those compounds meeting the search requirements was prepared for a review by the searcher. After the false drops were eliminated, a second listing was provided to the original requester. Providing these two listings on EAM equipment represented a considerable loss of time.

A second area of concern, which developed after experience with the fragmentation method of chemical searching, was the recognition that the system was greatly limited in performing substructure searches. Since the value of a collection of chemical structures depends primarily on the versatility in providing for specific as well as generic searches, the users of the RAPID system desired more capability in retrieving compounds with certain definite arrangements of atoms and bonds in common.

Recently, some sophisticated atom-by-atom and bond-by-bond substructure search programs have become available. The Substructure Search System, proposed by Gluck[2] at du Pont and further developed by the Chemical Abstracts Service,[3] is one such program. These topological methods represent atoms as nodes and connections as branches in a network. The chemical structure is considered a graph from which a connectivity table is built which completely represents the topology of a two-dimensional structure. Unlike fragmentation codes, which dissect chemical structures at input, topological codes do not subordinate any aspects of a structure to another, and the degree of discrimination is determined at the time of the search. Thus, a topological coding system represents maximum flexibility in chemical searching—the ability to retrieve specific compounds as well as unlimited capability in performing substructure searches. Because of this versatility in topological coding systems, FDA's Science Information Facility (SIF) acquired the CAS Substructure Search System (SSS) to serve as an agency-wide chemical retrieval system.

The Substructure Search System operates at two levels of search specificity. The fragment search acts as a screen that selects only those compounds meeting the minimum requirements established through fragment coding. After having greatly reduced the number of potential answers to a search query, a detailed iterative search is then performed on those compounds passing the fragment search.

The CAS Substructure Search System was installed by FDA in August 1968. After several months of testing and familiarization with the system, SIF considered the Substructure Search System to be an extremely effective chemical structure search tool. To confirm this, a study was undertaken to compare the effectiveness of RAPID's fragmentation system to the CAS SSS in retrieving chemical compounds.

The chemical compounds searched by the RAPID system are all components of INDs and NDAs; each entity is linked to every document containing that compound. There are approximately 6000 unique chemical compounds in about 20,000 documents resulting in over 28,000 punch cards representing chemical structures in the RAPID system. At the time of the study, the Substructure Search System at FDA contained 20,246 structural chemical compounds, which comprised the Common Data Base. To establish commonality between the two systems, the IND and NDA numbers of each chemical entity in the RAPID file were linked to CAS Registry numbers through a name and structure match.

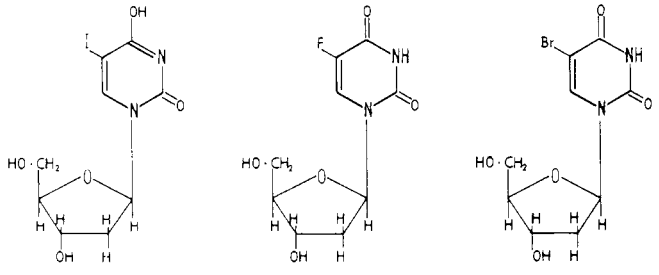
Beginning in April 1969, and continuing through July 1969, concomitant searches were conducted by the SIF and the Statistical Data Branch (SDB) personnel for compounds emanating from the incoming INDs and NDAs. The SDB group selected the searches to be run and defined the structures or substructures that were desired as search answers. SIF coded each search according to the specifications outlined by the SDB personnel without modification. This assured that each unit was conducting the same type of search.

A total of 60 unique queries was selected for searching during the study period; these were random in that they reflected structures found in new submissions to FDA. The searches utilizing the fragmentation method were run on EAM equipment, averaging three searches per run. The searches using the Substructure Search System were conducted in batches of 20 questions on an IBM 360/40. Table I gives the average times per search involved. At the end of the three-month period, the searches were compared and analyzed.

Table I.

		RAPID	SSS
Search Coding	(P)	2 min	8 min
Key punching	(P)	...	6 min
Searching the chemical structures	(M)	31 min. IBM 101 Sorter	2.3 min. (CPU time) IBM 360/40
Extraction of chemical names	(M)	15 min.	
Listing	(M)	5.2 min	6 sec
Review	(P)	16 min	5 min
Relisting	(M)	3.8 min	...

P = personnel time. M = Machine Time.



(III) (IV) (V)

Figure 2

Tautomerism. One of the more interesting observations made during this study concerned the handling of tautomers by the two systems. Using the fragmentation code, it is possible to retrieve data linked to one tautomeric form of a compound and miss information associated with the other form. For example, one search conducted by RAPID was for any IND or NDA application containing either 5-iodo-2′-deoxyuridine (III) or 5-fluoro-2′-deoxyuridine (IV) in combination with 5-bromo-2′-deoxyuridine (V) (Figure 2). One application containing the combination product was retrieved, but 27 applications containing the iododeoxyuridine were missed. The reason for this poor result was that at input, the bromo- and fluorodeoxyuridines were encoded as the keto forms and the iododeoxyuridines were encoded as the enol. The searchers, having checked the coding for two of the three compounds, looked only for the keto forms (that is, two carbonyl groups attached to a *N*-hetero ring). This, of course, caused the loss of all the deoxyuridines coded as enols.

This particular tautomeric situation was handled in the SSS search by looking for the oxygen atom attached to the *N*-hetero ring by a chain bond, and having a nonhydrogen count of one (only one other nonhydrogen atom attached to it). This allowed for the retrieval of either a carbonyl or hydroxy group attached at the position specified. If a tautomeric string involves all cyclic or all acyclic bonds, the SSS program automatically recognizes the tautomeric system and assigns all bonds an equivalent value.[1] This allows both forms of a tautomer to be retrieved if either one is sought.

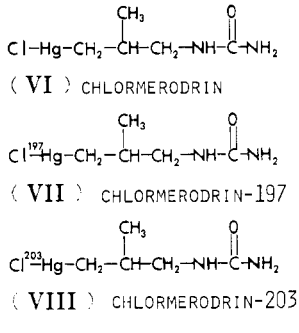
Radioactivity. The RAPID fragmentation code cannot discriminate between an unlabeled compound and its radioactive isomer; both would be retrieved and would have to be manually separated. The Substructure Search System is able to make this distinction and even has the capability of specifying a definite atomic mass. Figure 3 exemplifies this situation.

The SSS system would retrieve compounds (VII) and (VIII) if radioactive chlormerodrin were desired or could specifically retrieve either compound (VII) or (VIII) if a particular atomic mass were sought.

Ring Unsaturation. The degree of unsaturation within a ring system cannot be specified with the RAPID fragmentation code. This code uses the Patterson Ring Index number to describe a ring system without a further description of the types of ring bonds involved. Since one Ring Index number is used to describe many differently saturated rings, this procedure causes false drops in many instances. An illustration of the disadvantage of being unable to specify ring unsaturations was encountered by the RAPID group in a search for all 21-alkylesters of prednisolone (IX), as seen in Figure 4.

Since prednisolone differs from hydrocortisone only in the unsaturation at the 1,2 positions, RAPID had no way of discriminating between these two compounds. Accordingly, the search results showed 32 relevant hits and 93 hits that had to be considered false drops. Problems of this nature are avoided with the Substructure Search System since, with that system, one is able to specify the degree of unsaturation in a ring system.

Ring Substitutions. Another area that lacks specificity and is the cause of many false drops in the RAPID system concerns the ability to specify substitutions on rings. Although the fragmentation code has a number of punches to indicate that certan ring positions are sub-

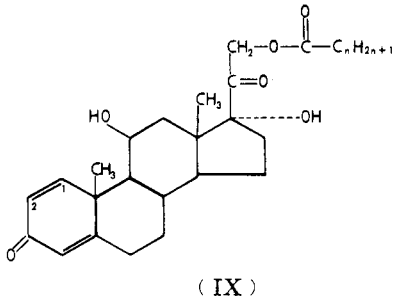


(VI) CHLORMERODRIN

(VII) CHLORMERODRIN-197

(VIII) CHLORMERODRIN-203

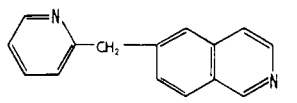
Figure 3



(IX)

PREDNISOLONE-21-ALKYLESTER

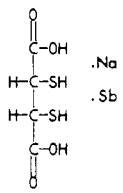
Figure 4



(X)

5-(2-PYRIDYLMETHYL) ISOQUINOLINE

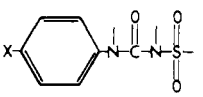
Figure 5



(XI)

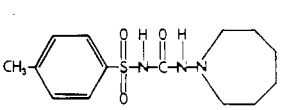
SODIUM ANTIMONY DIMERCAPTOSUCCINATE

Figure 6

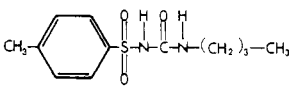


(XII)

Figure 7

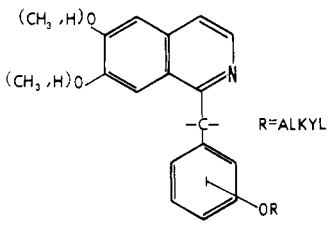


(XIII) TOLAZAMIDE



(XIV) TOLBUTAMIDE

Figure 8



(XV)

Figure 9

stituted, the punches are effective only when a single ring system is present. It the compound contains two or more ring systems, there is no way of indicating which ring is to have the desired positions substituted. In one of the RAPID searches, which attempted to retrieve all isoquinoline derivatives having the nitrogen atom substituted, a punch was used to specify that the 2-position of the ring had to be substituted. Compound (X) (Figure 5) fulfilled the specified punches, but was considered a false drop because the 2-position of the pyridine ring, rather than the isoquinoline ring, was substituted. Since the environment of each atom is described in the Substructure Search System, it was possible to specifically retrieve only those compounds having the 2-position (i.e., the nitrogen atom) of the isoquinoline ring substituted.

Polymers. The fragmentation code assigns definite punch positions for the fragments appearing in a polymer; therefore, it has some capability for searching such compounds. However, polymers appearing in the FDA file are not searchable through the CAS SSS. Accordingly, compounds such as dimethylpolysiloxane or methylcellulose are searchable with the RAPID system but not with the SSS system. The only means for retrieving such compounds would be a manual look-up of the assigned Registry numbers through a name match.

Indefinite Chemical Structures. Since the position and connection of every atom in a chemical structure must be known in order to encode it in a connection table, the following types of indefinite compounds cannot be handled by the connection table-based CAS SSS:

Compounds with indeterminant locants, such as "dichlorophenol"

Compounds with unstructurable salts, such as chlorpheniramine tannate

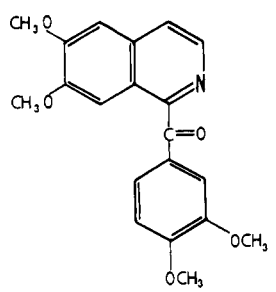
Metal derivatives of compounds with uncertain replaceable hydrogen atoms, such as sodium antimony dimercaptosuccinate (XI), shown in Figure 6.

The RAPID fragmentation code is capable of handling these indefinite chemical structures.

Functional Group Arrangement. Although the RAPID fragmentation code has several miscellaneous dictionaries that specifically describe certain combinations of hetero atoms as an entity, there is no way of depicting the arrangement of the group within the total chemical structure. In searching for the substructure (XII), as seen in Figure 7, RAPID was able to specify a para-substituted benzene ring being attached to a sulfonyl urea group; however, the code was unable to state that the benzene ring had to be attached to the nitrogen atom of the sulfonyl urea moiety. This search resulted in several false drops containing the proper fragments but with the wrong molecular arrangement. Compounds (XIII) and (XIV) represent two of these (Figure 8). This sort of false drop did not occur with the SSS system, and only those compounds that fulfilled the exact requirements of the search were retrieved.

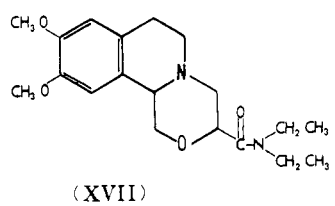
Functional Group Dependence. As previously stated, a fragmentation code divides a chemical compound into predetermined functional groups. Retrieval is then based on the absence or presence of these groups. However, chemical searching based solely on functional groups at times may be too restrictive and could cause the elimination of many relevant compounds. For instance, one of the RAPID searches sought all 6,7-dihydroxy- or 6,7-dimethoxyisoquinoline derivatives having an alkoxybenzene group one carbon removed from the 1-position of the isoquinoline ring; represented by substructure (XV) in Figure 9.

To include compounds with branching off the linking carbon atom, the RAPID search asked for any compound with an alkylene group connecting the benzene ring to the isoquinoline ring. However, compound (XVI) (Figure



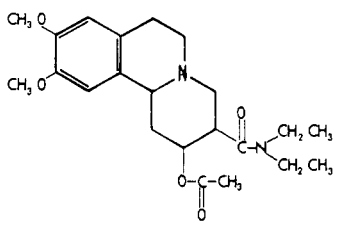
(XVI)

Figure 10



(XVII)

Figure 11



(XVIII)

Figure 12

10), retrieved by the SSS system and of interest to the searcher, was missed by the RAPID strategy because the carbonyl group was not sought.

In general, a fragmentation code is quite limited in its ability to perform true substructure searches. To locate the co-occurrence of certain atoms in a specific relationship requires a great deal of ingenuity on the part of the RAPID searcher. This capability is built into the SSS system; any substructure that can be adequately defined can be coded and found through SSS.

Another search that illustrates the diversification offered by a topological coding system pertains to compound (XVII) (Figure 11). In looking for related compounds, the RAPID search utilized two approaches; the first strategy sought any compound having the same ring system as compound (XVII) irrespective of substitution. The second strategy looked for those compounds containing a 6,7-dihydroxy- or a 6,7-dimethoxyisoquinoline ring with a diethylamide group two carbon atoms removed from the nitrogen atom of the ring. Although the best available strategies were devised, a negative search resulted. The same substructure was coded for SSS and one of the hits obtained was compound (XVIII) (Figure 12).

SUMMARY AND CONCLUSIONS

Analysis of the results of this comparative study convincingly demonstrated the greater capability of the CAS Substructure Search System over the Frome and O'Day fragmentation code in retrieving chemical compounds. The degree of specificity available in phrasing a query with the SSS system is greater than with this fragmentation code. This stems from the fact that SSS utilizes more fragments (1800 vs. 892) in screening compounds and, in addition, has an iterative search capability for those compounds passing the fragment screen. The topological search enhances the ability of the system to retrieve only pertinent answers. In the present study, 25% of the answers retrieved by RAPID were false drops; the SSS system did not lead to a single false drop.

An over-all comparison of the two methods of chemical searching showed that the Frome and O'Day fragmentation code demands more skill on the part of the searcher, That is to say, the searcher must be very familiar with the compounds in the data base and how they were encoded, he must anticipate every possible variation in functional groups or ring systems, and he must manually eliminate false drops in most searches.

The prejudgment of compounds in the data base that would be of interest in a chemical search was not required with the Substructure Search System. The only requirement was a precise definition of the chemical substructure desired in the compounds being searched.

ACKNOWLEDGMENT

The authors would like to show appreciation for the assistance provided by the personnel in the Statistical Data Branch/Bureau of Drugs: Joyce Hinckley, Carlos Smith, and Henri Williams. Thanks also to Henry Kissman, Alan Gelberg, Bruno Vasta, and Gerard Guthrie of the Science Information Facility for their encouragement and technical guidance. Finally, we acknowledge the valuable cooperative efforts of Charles E. Simmons, Elizabeth McNamee, and Eloise Ingram of the Division of Data Processing.

LITERATURE CITED

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

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

(3) 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).

(4) Vasta, B. M., M. L. Spann, and G. T. Guthrie, "Experience with the CAS Substructure System." Paper presented at Fourth Middle Atlantic Regional Meeting, ACS, Washington, D.C., 1969.

Chemical Structure and Substructure Search by Set Reduction*

TAO-KUANG MING and STEPHEN J. TAUBER
National Bureau of Standards, Washington, D. C. 20234

Received April 22, 1970

As part of a computerized system for handling chemical and related information[1] we have included routines for handling chemical structure information. Among these routines is one for structure and substructure search by set reduction, based directly on the method of Sussenguth.[2,3] We have introduced the following refinements: separation of structure search and substructure search into distinct subroutines and inclusion of first-order degree and second-order degree[2] in the control vector for use in structure search. This feature performs much the same function as the connectivity code described by Penny.[4] In addition our routine has processed structures with greater symmetry than any which Sussenguth's routine is

* Presented in part before the Computers Division of the 3rd Great Lakes Regional Meeting, ACS, DeKalb, Ill., June 5–6, 1969.

This work was supported by the National Institutes of Health under an interagency agreement. Contribution from the National Bureau of Standards. Not subject to copyright.