On-Line Storage and Retrieval of Chemical Information. II. Substructure and Biological Activity Searching[†]

V. LYNN BOND, CARLOS M. BOWMAN, LINDA C. DAVISON, PATRICIA F. ROUSH,* ROGER D. MCGREW, and DANIEL G. WILLIAMS

Systems Research, The Dow Chemical Company, Midland, Michigan 48640

Received February 20, 1979

An improved interactive system for searching substructure and biological activity data has been developed. Features of the system include a two-level substructure search (fragment screen and atom by atom) and an expanded biological activity data base. The system operates on a file of about 150 000 compounds.

An on-line system for retrieving information on compounds with prescribed structural features was developed at Dow in 1969.[1] It initially operated on a Burroughs B-5500 and was modified in 1972 for an IBM 360. This system and the batch system[2] which preceded it were based on substructure fragmentation codes which were coordinated using Boolean logic. As the size of the file grew to well over 150 000 compounds, and as more refined techniques for substructure searching were reported in the literature,[3-5] the need for a more flexible search program became apparent. The first part of this paper describes the development of this program and the search strategy associated with it. The second part discusses changes to the biological activity data base and search, which are directly interfaced with the substructure search.

I. SUBSTRUCTURE SEARCH

It has been shown that an effective mechanism for searching a large structure data base is to phrase the request at varying levels of specificity.[3] The first level approach is usually a fragmentation code of some sort which effectively screens the file of irrelevant compounds but does not in and of itself specifically define the query structure. An atom-by-atom search is often used to further refine the results of a fragment screen, but used alone on an entire data base is prohibitively expensive. Dow's experience with fragmentation codes, combined with an increase in computing power and storage capacity, led to the decision to develop a combined fragmentation/atom search. An initial criterion of the system was to define the search vocabulary (i.e., fragments and atom codes) in terms that virtually any chemically trained person in Dow could understand. Secondly, the system had to be flexible enough to allow for repeated modifications of search results. Finally, the cost per search had to be low enough to make the entire system economically feasible. The resulting system is described in the following sections. At present, only Dow's internal compound file (Dow Registry System) is included in the data base. This represents about 150 000 defined structures, of which about 130 000 are also in the biological activity data base. About 8000 to 12 000 new structures are added per year.

FRAGMENTATION CODE

The use of several Dow-generated codes over the years, combined with a knowledge of the compound data base and the general areas of Dow research, led to the development of a fairly comprehensive fragment code. About 300 substructural features are represented. As in the past, a bit string (with each bit representing an individual code) was chosen as the most compact storage medium. The codes, along with the atom connection table, are generated by the WLN analysis program described in the preceding paper.[6] Since the WLN is the basic storage record for all defined structures, some of its symbol definitions were retained in identifying both fragment codes and connection table atom values.

Fragments for more common elements and functional groups (e.g., carbon, nitrogen, oxygen, carboxyl group) were defined in some detail. This includes codes for ring substitution, molecular formula count, multiple occurrence, and (where applicable) hydrogen substitution or unsaturation. Fragments related to oxygen are illustrated in Figure 1. A fairly elaborate scheme was worked out for carbon atoms based on all possible hydrogen/unsaturation configurations (Figure 2). Less frequent elements were defined by a single fragment to record their presence, or in some cases, grouped together as one code (e.g., transition metals).

A series of fragments was defined to identify ring characteristics. These include identifiers for ring type (e.g., heterocycle vs. carbocycle), degree of fusion, saturation, and ring size. Some common rings such as pyridine and piperazine were defined specifically. Benzene rings were divided into monosubstituted and multisubstituted benzenes, and multiple occurrences of each of these species are recorded. Some of the fragments generated for a typical compound are illustrated in Figure 3.

The user selects fragments for his search from a printed list (given to all users). The procedure for constructing a search query is described below. The use of these fragments successfully eliminates at least 96% of the file at the first level.

CONNECTION TABLE

A connection table is used for an atom-by-atom search. The table contains a record of each atom in the compound, its character value, and the atoms to which it is connected. In selecting atom values for the table, the Wiswesser symbols V, M, Z, G, E, F, I, and Q retain their original definition. "R" has been modified to define monosubstituted benzene. Multisubstituted benzenes are expanded into their individual carbon atoms. Carbon atoms have been defined to reflect their hydrogen/unsaturation configuration in a manner analogous to the fragment codes. These carbon symbols are commonly referred to as "dot-plot" symbols.[7,8] Methylene chains of two or more carbons are recorded by a number reflecting the number of carbons. All other elements retain their element symbols except those five altered by the Wiswesser code (K = -KA-, W = -WO-, V = -VA-, U = -UR-, Y = -YT-). In addition to the connectivity of each atom, a record of its charge and ring environment (whether it is in or on a ring) is also made. These additional attributes aid in eliminating possible candidate atoms from an atom search. A portion of the atom code dictionary, representing the various codes for carbon, is

[†] Presented before the Division of Chemical Information, 175th National Meeting of the American Chemical Society, Anaheim, Calif., March 15, 1978.

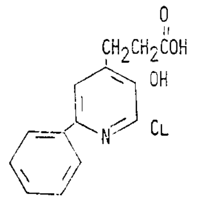
0095-2338/79/1619-0231$01.00/0

Definition		Fragment no.
Carbonyl (>C=O)		79
	More than one acyclic in compound	82
	On ring	81
	More than one on ring	252
	In ring	80
	More than one in ring	288
Ether(-O-)		73
	In ring	74
	More than one in ring	287
	On ring	75
Molform		
	O1	83
	O2	84
	O3	85
	O4	86
	O>4	87
Generic oxygen		72
	On ring sidechain	284
	On benzene	272
	On heterocyclic	273
	On carbocyclic (non benzene)	274
Hydroxyl(-OH)		76
	More than one in compound	78
	On ring	77
	More than one on ring	251

Figure 1. Oxygen fragments.

Definition		Fragment no.
Carbon on benzene		263
Carbon on heterocyclic		264
Carbon on carbocyclic (non benzene)		265
Carbon on sidechain		281
Molform		
C1-C4		30
C5-C6		31
C7-C12		32
C13-C20		33
C21-C35		34
C>35		35
$-\overset{\mid}{\underset{\mid}{C}}-$	Acyclic	1
	Spiro	2
	In ring (non spiro)	3
	On ring	4
=C<	Acyclic	5
	In ring	6
	On ring	7
$-\overset{\overset{H}{\mid}}{\underset{\mid}{C}}-$	Acyclic	8
	In ring	9
	On ring	10
=C= or -C≡		11
-CH=	Acyclic	12
	In ring	13
	On ring	14
$-CH_2-$	Generic	15
	In ring	16
	On ring	17
$-(CH_2)_{\geq 9}$		20
$-(CH_2)_2$-to-$(CH_2)_n$		22-29

Figure 2. Carbon fragments.



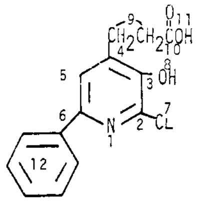
FRAGMENTS = APPROX 35

PYRIDINE RING	OXYGEN ON HETEROCYCLE
-N= IN RING	CARBOXYLIC ACID
CHLORINE ON RING	$-CH_2-$ ON RING
HALOGEN ON HETEROCYCLE	$-CH_2CH_2-$
HYDROXYL ON RING	MONOSUBSTITUTED BENZENE

Figure 3. Some fragments for a typical compound.

CODE	DEFINITION	CODE	DEFINITION
X	$-\overset{\mid}{\underset{\mid}{C}}-$	C	=C= OR -C≡
		D	-CH=
T	=C<	1	$-CH_3$
		L	$-CH_2-$
Y	$-\overset{\overset{H}{\mid}}{\underset{\mid}{C}}-$	N	$-(CH_2)_N-$

Figure 4. Atom codes—carbon.



1	N	2,6	IN RING
2	T	1,3,7	IN RING
3	T	2,4,8	IN RING
4	T	3,5,9	IN RING
5	D	4,6	IN RING
6	T	5,1,12	IN RING
7	G	2	ON RING
8	Q	3	ON RING
9	O2	4,10	ON RING
10	V	9,11	ON RING SIDECHAIN
11	Q	10	ON RING SIDECHAIN
12	R	6	ON RING

Figure 5. Connection table.

illustrated in Figure 4, and a sample connection table is found in Figure 5.

SEARCHING

An on-line interactive program provides access to the data base at two levels of specificity. The user has the ability to correct and modify his request, to save the results of any or all searches for further searching, and to restrict his search to a more current part of the file. Search results are available as chemical names, Wiswesser Line Notation, and/or accession number. A file of accession numbers may be created for retrieving biological activity data through a separate program (see below). A program to display structures is currently being developed.

Fragment codes for the first level search are coordinated using Boolean operators. Four specific search types are available:

AND	One series of fragments, all must be present
AND/OR	Several series of fragments, at least one from each group must be present
NOT	One series of fragments, all must be absent
AND/NOT	Two series of fragments, all in first group must be present, all in second group must be absent

A simple AND search is illustrated in Figure 6. A first level search of the entire data base of 150 000 compounds may take anywhere from 2 to 5 min of elapsed time, depending upon the load on the system. Results are typically under 1500 compounds and rarely over 5000.

If the first level search results are sufficiently large, the user may wish to proceed to an atom search. Some users familiar

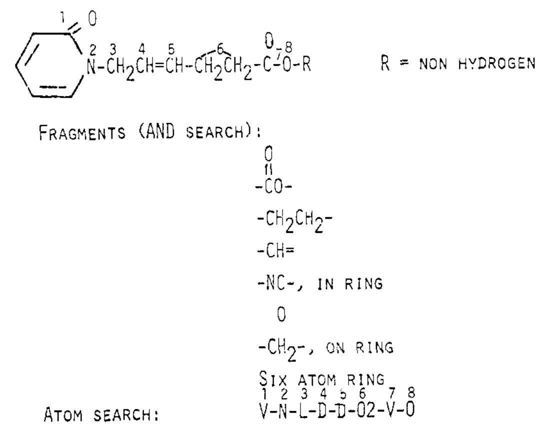


Figure 6. Search query.

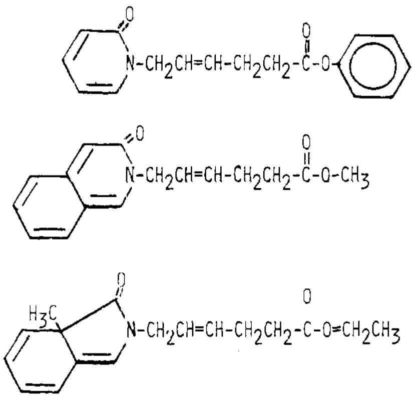


Figure 7. Search results.

with WLN or nomenclature will instead choose to screen the level one results visually if the desired features are easily recognized (e.g., rings) in the output mode selected. An atom search related to the preceding fragment search is also illustrated in Figure 6. The program prompts the user for information on the number of atoms (or nodes) in the request, the node values, their connections, whether they are in or on a ring, and their charge (where applicable). Up to eight nodes and eight values within each node are allowed.

All atom searches must be preceded by a level one fragment screen. The level one results are read sequentially from a file, and a connection table for each compound is retrieved from a random access file. The connection table and query table are matched using a set reduction technique developed by Sussenguth.[9] A typical search of about 500 compounds requires about 2 h of elapsed computer time. For this reason, level two searches are usually run after hours or on weekends. Currently, however, they comprise less than 5% of all searches run.

Some results of the above query are illustrated in Figure 7. About 100 substructure searches are run per year. Level one searches average from $15 to $20 per search. The cost of level two searches is directly proportional to the number of compounds searched, and may be from $100 to $500. The results of a substructure search may be stored in a file and further checked for biological activity via the biological activity search system (see following section). Similarly, results from an activity search may be stored and searched for structure correlations.

II. BIOLOGICAL ACTIVITY SEARCH

An improved version of the on-line biological search system[1] has been developed for screening data in the herbicide, fungicide, and insecticide areas. All biological test data are stored on master file magnetic tapes. Each test result occupies a single record and contains the compound accession number, the test method, the organism, the concentration, and the activity. On-line storage of these files is impractical because of their size, so for this purpose the files were condensed to approximately 10% of their original size by eliminating redundancies and nonactive records. A note that a particular test method–organism combination had been run was kept so very little information was lost in the compression.

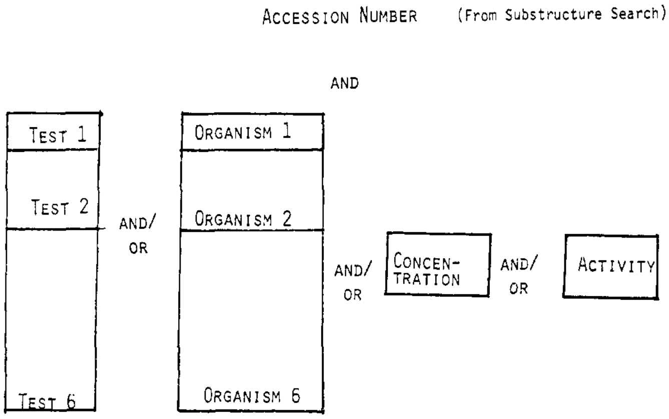


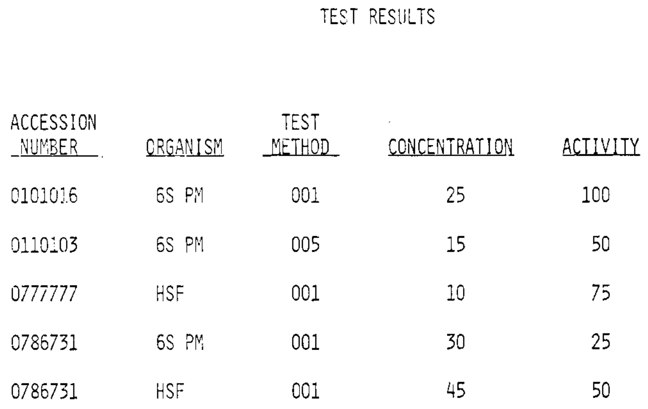
Figure 8. Search parameters for activity search.

```
        WELCOME TO THE

      INSECTICIDE SEARCH


>: TEST
ENTER TEST METHODS
+ : 001
+ : 005
+ :


>: ORGANISM
ENTER ORGANISM CODES
+ : 6S PM
+ : HSF
+ :


>: CONCENTRATION
ENTER LOWER CONCENTRATION
+ :   50
>: GO
```

Figure 9. Activity search protocol.

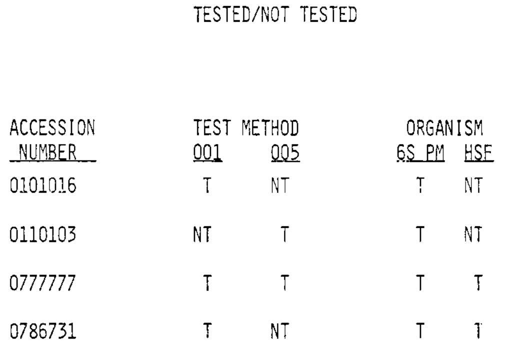
The search program serves two general functions. First it may be used to retrieve test data on a list of accession numbers. These may be entered by the user at the keyboard or they may be entered from a file created by the substructure search program. Although this information may be obtained from a microfilm, the computer listing is usually more convenient since the data spans several reels of microfilm.

The second function of the program allows users to ask biological questions. All of the file parameters (test method, organism code, concentration, activity) may be included. For example, the user might want to see a list of accession numbers that have activity against a particular organism at less than 50 ppm. Quite complex questions may be asked by logically combining the various parameters. Queries may involve a search of the whole compound data base, or may be restricted to a set of compounds resulting from a previous substructure



```
                          TEST RESULTS


ACCESSION                  TEST
 NUMBER      ORGANISM     METHOD     CONCENTRATION     ACTIVITY

0101016       6S PM        001            25              100

0110103       6S PM        005            15               50

0777777       HSF          001            10               75

0786731       6S PM        001            30               25

0786731       HSF          001            45               50
```

Figure 10. Activity search results.



```
                   TESTED/NOT TESTED


ACCESSION       TEST METHOD         ORGANISM
 NUMBER         001    005         6S PM  HSF
0101016          T      NT           T    NT

0110103          NT     T            T    NT

0777777          T      T            T     T

0786731          T      NT           T     T
```

Figure 11. Test method—organism data for selected compounds.

or activity search. Figure 8 shows the various possibilities.

The program operates in interactive mode. The options are selected by entering commands as shown in Figure 9. The program prompts the user for a command with a ">:". The user's first parameter is test methods. The "+:" is used to prompt for further information. In this case the test method numbers were entered. For the organism parameter, two test organisms, six-spotted peach moth (denoted 6SPM), and horse-shoe fly (HSF), were selected. A concentration of <50 ppm was also specified. And finally the "GO" signaled the program to begin the search.

The answers are shown in Figure 10. All records satisfying the search criteria are displayed. The test conditions and activity are also listed. The user is given the option of seeing whether or not a particular test method–organism combination has been run. This is shown in Figure 11.

Generally the user cannot accurately anticipate the number of answers to this type of search. If a question has a large number of answers, it needs to be refined in order to be useful. This is also true of the substructure search. For this reason, both programs have extensive "recycling" capabilities, whereby a user can "narrow down" his search by introducing additional parameters. This is a common feature in most on-line systems.

The use level and cost of activity searches are comparable with substructure searches. About 75 searches are run per year at a cost of about $10 per search.

CONCLUSIONS

A system for on-line substructure and biological activity searching has been enhanced through the introduction of a multilevel substructure search and expanded activity data base. Response from users has been enthusiastic, and consequently usage of the system has increased significantly.

ACKNOWLEDGMENT

The authors wish to thank Dr. Ted. Holmsen and his colleagues for their assistance in this project.

LITERATURE CITED

(1) Bond, V. B.; Bowman, C. M.; Lee, N. M.; Petersen, D. R.; Reslock, M. H. "Interactive Searching of a Structure and Biological Activity File", *J. Chem. Doc.*, **1971**, *11*, 168–170.

(2) Bowman, C. M.; Landee, F. A.; Lee, N. W.; Reslock, M. H.; Smith, B. P. "A Chemically Oriented Information Storage and Retrieval System. III. Searching a Wiswesser Line Notation File", *J. Chem. Doc.*, **1970**, *10*, 50–54.

(3) Eakin, D. R.; Hyde, E. "Evaluation of On-Line Techniques in a Sub-Structure Search System", in "Computer Representation and Manipulation of Chemical Information", Wipke, W. T., Ed.; Wiley: New York, 1974; pp 1–30.

(4) Feldmann, R. J.; Milne, G. W. A.; Heller, S. R.; Tein, A.; Miller, J. A.; Koch, B. "An Interactive Substructure Search System", *J. Chem. Inf. Comput. Sci.*, **1977**, *17*, 157–163.

(5) Tomea, Albert V.; Sorter, Peter F. "On-Line Substructure Searching Utilizing Wiswesser Line Notations", *J. Chem. Inf. Comput. Sci.*, **1976**, *16*, 223–227.

(6) Bowman, Carlos M., Davison, Linda C., Roush, Patricia F. "On-Line Storage and Retrieval of Chemical Information. I. Structure Entry", preceding paper in this issue.

(7) Hyde, E.; Matthews, F. W.; Thomson, L. H.; Wiswesser, W. J. "Conversion of Wiswesser Notation to a Connectivity Matrix for Organic Compounds", *J. Chem. Doc.*, **1967**, *7*, 200–204.

(8) Wiswesser, W. J. "The "Dot Plot" Computer Program", Division of Chemical Literature, 152nd National Meeting of the American Chemical Society, New York, Sept 1966.

(9) Sussenguth, E. H., Jr. "A Graph-Theoretic Algorithm for Matching Chemical Structures", *J. Chem. Doc.*, **1965**, *5*, 36–43.