retrieval becomes outdated. Improvements are taking place so fast (in databases, search programs, and communications techniques and equipment) that there is little which can be said now which will remain unchanged for more than a few months. Users are advised to stay in close contact with the database producers, on-line brokers, and communications vendors to keep abreast of the changes.

LITERATURE CITED

(1) G. V. O'Bleness and G. Cohen, "Presentation of "SOLD" Computer Retrieval System", paper presented at Central Patents Index Subscribers Meeting, Washington, D.C., May 1973.
(2) R. Donati (Lockheed Information Systems), "Overview of Patents Covered by DIALOG Retrieval Service", presented at Session on Patent Literature Systems from the Symposium on Intellectual Property as Sources for Scientific, Technical and Business Information sponsored by the Canadian Patent Office and others, Ottawa, Ontario, Nov 24, 1976.
(3) M. Bonner (System Development Corporation), private communication, Dec 1976.
(4) M. E. Williams and S. H. Rouse, Ed., "Computer-Readable Bibliographic Data Bases - A Directory And Data Sourcebook", American Society for Information Science, Washington, D.C., 1976.
(5) J. B. Hare, "On-Line Searching for Patent Information—A Comparison Of The CHEMCON And WPI Data Bases", paper presented at Pharmaceutical Manufacturers Association Science Information Subsection Meeting, Hot Springs, Va., March 1976.

An Interactive Substructure Search System

R. J. FELDMANN and G. W. A. MILNE

National Institutes of Health, Bethesda, Maryland 20014

S. R. HELLER*

Environmental Protection Agency, Washington, D.C. 20460

A. FEIN, J. A. MILLER, and B. KOCH

Fein-Marquart Associates Inc., Towson, Maryland 21212

Received April 29, 1977

A family of programs for searching on the basis of chemical structure through data bases of chemical information has been assembled and is now publically available on a commercial computer network. The design of and results obtained with these programs are reported, and the status of the system is described and discussed with particular reference to the NIH–EPA Chemical Information System (CIS) and the Toxic Substances Control Act (TSCA).

INTRODUCTION

The ability to use a computer to search for a particular chemical structure or substructure in files of chemical data has for some time been sought after by chemists, and the need for such capability is currently becoming very pressing. A widening interest in the relationships between chemical structure, on the one hand, and various properties, such as toxicity, pharmacological activity, and mutagenicity, on the other, has led in recent years to considerable efforts to generate computer programs which will enable the scientist to locate all occurrences of a given structure or substructure in chemical databases. Further decisive pressure behind these developments has been provided by the enactment, in November 1976, of the Toxic Substances Control Act (Public Law 94-469). This law will require that chemical compounds whose use in commerce is envisaged must first be located within Governmental regulatory files. If they are not in these files, they are deemed "new" and their use in commerce becomes subject to a series of regulations, depending upon their respective toxicities.

In this paper, we describe the NIH–EPA substructure search system, a family of interactive computer programs which allow the user to define a chemical structure or substructure and then to search for occurrences of the structure or substructure in the various databases of the NIH–EPA Chemical Information System.

During the past 25 years, a considerable number of methods of machine representation and handling of chemical structure have been proposed and studied for their utility in manual and automatic data retrieval methods. Some of the better known among these include the German GREMAS system,[1] the British CROSSBOW system,[2] and, in the U.S., the programs developed at the National Cancer Institute,[3] Walter Reed Army Institute of Research,[4] Chemical Abstracts Service,[5] and the Army's Chemical Information Data System.[6]

With the resulting progress in the area of computer-handling of chemical structures, it has become clear that structure records in the form of two-dimensional connection tables are absolutely necessary for structural representation and that both linear notations and chemical nomenclature are at a serious disadvantage vis-à-vis connection tables as far as unambiguity and completeness are concerned. For many years, however, there was no adequate means of screening such connection tables and so, in spite of their intrinsic value, they were not used in any retrieval system.

In the area of structure retrieval, most effort was expended in the development of systems that were designed to fulfill a specific local need. A system of this type that is currently perhaps the most widely used by the chemical industry is the CROSSBOW program,[2] a dozen or so versions of which are in operation around the world. Other systems, such as the GREMAS system[1] or the Walter Reed system[4] require special equipment that is not generally available. While the larger industrial organizations can often afford such luxuries and often also demand in-house facilities of this sort, such systems are of little value to the general chemist. It is this dilemma which led to the development of the NIH nested tree structure searching system, which can operate on a connection table database and which is susceptible to wide dissemination and

Table I. Databases That Can Be Searched by the Substructure Search System.

Cambridge (Xray) Crystal File.
CPSC Chemicals in Consumer Products.
EPA AEROS SOTDAT File.
EPA Las Vegas Chemical Spill File.
EPA Storage and Retrieval of Air Data.
EPA Pesticide Standards.
EPA STORET Water Data Base.
EPA-FDA Pesticide Repository Standards.
EPA Inactive Ingradients in Pesticides.
EPA Oil and Hazardous Materials File.
EPA Pollutants in Drinking Water.
EPA Pesticides File.
EROICA Thermodynamics Data File.
Merck Index.
NBS Gas Phase Proton Affinities.
NBS Heats of Formation of Gaseous Ions.
NBS Single Crystal File.
NCI-SRI Industrial Chemicals File.
NCI PHS-149 File of Carcinogens.
NIMH File of Psychotropic Drugs.
NIH-EPA Carbon-13 Nuclear Magnetic Resonance Search System.
NIH-EPA Mass Spectral Search System.
WHO International Non-proprietary Name File of Drugs.

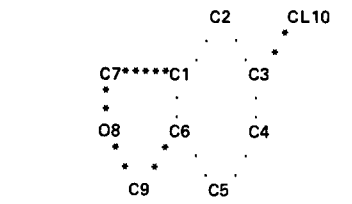
use on a time-shared computer. This system was originally conceived by Feldmann in 1971,[7] and since that time there has been considerable use, testing, and further development. These developments, described below, have been in various different directions, and the more significant features are the following. (1) The programs have been extensively rewritten to reduce CPU demand and improve the user-machine dialog; the new software has been completely documented. (2) A comprehensive User's Manual has been written. (3) The Chemical Information Data System (CIDS) structure codes have been introduced into the substructure search programs where they can be used for searching, or, more importantly, as the basis of fragment screen procedures. (4) While the interactive programs run on a DEC PDP-10, all the database preprocessing is now accomplished on an IBM 370/168. (5) Other additional search features based upon user feedback and on successful features of other systems have been added.

DATABASES

The substructure search programs are designed to support the searching of a number of separate and independent databases. In the current version of the software, a decision must be made by the user as to the identity of the database that will be searched. There is, however, considerable overlap between separate databases and so work is now in progress to merge all the databases, remove duplicates, and arrive at a single consolidated file.

There are currently 23 distinct databases associated with the substructure search system. These are listed in Table I. Many of these files of chemicals have been derived from the files of the EPA. This agency, alone in the U.S. government, has an internal regulation[8] requiring registration of all agency databases containing chemical information. A second major source of files of connection tables is the NIH/EPA Chemical Information System.[9] This is a collection of databases containing spectroscopic and other data that relate to chemical compounds. A decision fundamental to the development of the CIS has been to register all chemicals in the component files of the system. Finally, other databases that are used by the substructure search system have been obtained from other U.S. government agencies. If necessary, they have then been registered for inclusion into the substructure search system.

The process of registration of a compound by CAS takes place in three steps which have been described in detail elsewhere.[10] When a database is obtained for registration and inclusion into the CIS and the substructure search system, the



1 C	2	6	7	0	0	0	9	9	1	0	0	0
2 C	1	3	0	0	0	0	9	9	0	0	0	0
3 C	2	4	10	0	0	0	9	9	5	0	0	0
4 C	3	5	0	0	0	0	9	9	0	0	0	0
5 C	4	6	0	0	0	0	9	9	0	0	0	0
6 C	1	5	9	0	0	0	9	9	1	0	0	0
7 C	1	8	0	0	0	0	1	1	0	0	0	0
8 O	7	9	0	0	0	0	1	1	0	0	0	0
9 C	6	8	0	0	0	0	1	1	0	0	0	0
10 CL	3	0	0	0	0	0	5	0	0	0	0	0

Figure 1. A connection table. The structure is not normally included but is given here for the sake of clarity.

names of all chemicals in the database, together with the appropriate accession numbers, are delivered to CAS. The first step of registration is an attempt to find the compound name in the master CAS nomenclature files. If this name is found, then the "correct" name and the CAS registry number can be extracted and labeled with the accession number. If the name is not found, then CAS, in the second step of the registration process, establishes a structure for the compound and seeks to locate that structure in its structure files, and so arrive at the correct name and registry number. If this step fails, the compound is assumed to be absent from the CAS master authority files and is then given a name and registry number and incorporated into the files.

Ultimately, all the identifiable chemicals in the NIH/EPA CIS files are registered by one or another of these methods and, at that point, a database of accession number, CAS registry number, the name under which it is listed in the CAS (8th or 9th) Collective Index, and the connection table is returned to NIH/EPA for merging into the CIS. A separate file of registry number and synonyms is also obtained from CAS and used in the CIS.

Finally, as a general procedure, all the information that is derived from the CAS files is subject to annual updating. In this way, errors and refinements such as additional synonym information may be incorporated in the substructure search files.

The connection table that is supplied by CAS is not used directly by the substructure search system. Rather, it is translated into a derived connection table which is itself merged into the substructure search system files. This articulation leads to an important advantage in that any changes made by CAS in the format of their connection tables can be handled simply by an adjustment in the program that carries out the translation. The more complex programs that actually handle the substructure searching do not have to be changed.

The basic connectivity information for a compound is supplied by CAS in the form of three separate data elements in a single structure record. These data elements, which are designated the "graph", "nodes", and "bonds" elements, define the element type of each of the nodes, the connections between nodes, and the bond type of each of the connections.

The basic connectivity information present in the structure records supplied by CAS is used to generate the tabular connection table that is shown in Figure 1 and which is of the type that is employed by the substructure search system. The main feature of this derived connection table is that it permits rapid access to all connectivity information associated with a particular node or atom. Generation of the derived connection table is relatively straightforward, although a few minor problems arise. For example, dot-disconnected structures (e.g., of anion–cation pairs) are independently

Table II. Structure Generation Commands.

COMMAND	EFFECT
AATOM n1 m1	Insert an atom between atom n1 and atom m1.
ABOND n1 m1	Insert a bond between n1 and m1.
ABRAN l1 at n1	Add a branch of length l1 at atom n1.
ALINK n1 l1 m1	Insert a chain of length l1 between n1 and m1.
ALTBD n1 m1	Define alternate bonds in the smallest ring containing n1 and m1 as aromatic bonds.
ARING n1 m1 l1	Create a ring of l1 atoms between n1 and m1.
CHAIN l	Create a chain of l atoms.
CLEAR	Erase the existing query structure.
CRING n1 l1	Create a ring of l1 atoms including atom n1.
DATOM n1	Delete atom n1.
DBOND n1 m1	Delete the bond joining nl and m1.
MORGA	Renumber the query structure by the Morgan algorithm.
NUC 66	Create a structure of two fused six-membered rings.
REG	Retrieve the structure corresponding to a specific registry number.
REST	Negate the effect of the previous command.
RING l	Create a ring of l atoms.
SATOM n1	Define the elemental nature of atom n1.
SBOND n1 m1	Define the nature of the bond joining n1 and m1.
SPIRO n1 l1	Create a spiro-attached ring of (l1 +1) atoms at n1.
WISBD n1 m1	Define alternate bonds in the smallest ring containing n 1 and m1 as double bonds.

numbered and stored by CAS but must be assembled and renumbered for processing within the substructure search system.

In addition to the basic connectivity information described above, CAS also provides information describing a large number of more unusual structural features, such as charge, abnormal mass, abnormal valency, stereochemistry, and so on. These additional features are not currently handled by the substructure search system.

STRUCTURE SEARCHING

A search through a database of connection tables for a specific complete structure, as opposed to an imbedded partial structure, can take advantage of different design techniques and is considered here independently of substructure searching.

It is vital to the implementation of the Toxic Substances Control Act (PL 94-469) that the government-maintained inventory of chemicals used in commerce can be examined rapidly and accurately for the presence or absence of specific compounds. This requires a search for a full structure and is performed effectively by searching through the database of connection tables for a specific connection table, defined by the user.

To do this, the system permits the user to generate a "query structure", as described in the next section. Once the query structure is complete, an identity search can be requested. At that time, a modified form of the connection table corresponding to the query structure is converted to a hash-coded form. The process of hash-encoding involves conversion of the connection table to a single number which is a probable, but not guaranteed, unique representation of the compound's structure. The searching programs then scan a set of searchable files which are indexed by that number for an exact match.[11] If the match is found, the registry number of the matched entry is reported.

SUBSTRUCTURE SEARCHING

The substructure search process involves three distinct steps. The first of these is the generation of a query structure. This is followed by the actual search, and the final step is the display of the structures that have been retrieved from the database by the search. Each of these steps will be treated separately below.

```
OPTION ? NUC
SPECIFY NUCLEUS LINE CODE

LINE CODE =  65
OPTION ? ABRAN  1 AT  3
OPTION ? ALTBD  1 2
OPTION ? SBOND  3 10
BOND TYPE (H FOR HELP)    = CS
OPTION ? SBOND  1 7 7 8 8 9 6 9
BOND TYPE (H FOR HELP)    = RS
OPTION ? SATOM  8
SPECIFY ELEMENT SYMBOL  = O
OPTION ? SATOM  10
SPECIFY ELEMENT SYMBOL  = CL
OPTION ? D
        2      10CL
       . .    *
      .   .  *
7*****1     3
*     .     .
*     .     .
8O    6     4
 *   *  .   .
  * *    . .
   9      5
```

Figure 2. Use of the structure generation commands to define a structure.

1. Query Structure Generation. One of the more difficult problems in the design of an interactive substructure search system is how the chemist can enter a chemical structure or partial structure into the machine. In the present case, this is accomplished by means of a family of programs that permit the generation of a structure at the computer terminal. These programs have been designed in such a way that the simplest of computer terminals is adequate for their use and so the substructure search system can be accessed by a simple teletype terminal as well as by an advanced graphics terminal. In either case, the user must, with the commands that are given in Table II, generate the query structure of interest. As each command is received, the system generates or modifies a connection table so that it reflects the current structure. The connection table is invisible to the user, although a version of it can be printed out at a command. A more useful option, however, is the DISPLAY command (D), which, working from the current connection table, produces a drawing of the corresponding structure, using procedures designed to produce an unambiguous two-dimensional representation. With the commands given in Table II, the chemist can generate rings of specified sizes, add branches to existing atoms, and specify the identity of specific atoms or bonds. Various other, more powerful, commands include the NUCLEUS command, with which one can generate a fused multi-ring system, such as that common to steroids, in one step. As a query structure is being developed, it is often necessary to be able to inspect it so as to identify the numbers assigned by the program to various nodes (atoms), and it is here that the display command is very useful. When certain modifications are made to the query structure, the program will renumber the atoms, and so it is necessary to examine the structure in order to proceed. An example of the dialog involved in the generation of a query structure is shown in Figure 2. The basic ring system is provided by the NUC command, a branch is added by the ABRAN option, bonds are all specified by the ALTBD and SBOND options, and noncarbon atoms are defined by the SATOM command. The result is the query structure that is finally examined with the help of the display command "D".

Once the query structure has been generated, the second step, a search through the database, may be undertaken. This searching may be accomplished in a variety of different ways. The most trivial of these are the special property searches which scan the database for all compounds whose molecular

formula corresponds to that of the query structure, or which have a given molecular weight, and so on. The most used of the structural searches are those in which a specific atom-centered fragment from the query structure is searched for in the database (FPROB) and those (RPROB) in which a particular ring or rings from the query structure is the object of the search. Finally, the SUBSTRUCTURE SEARCH, which is the most exhaustive of all the searches, examines every connection table in subsets of the database on an atom-by-atom, bond-by-bond basis in a search for an exact or imbedded match for the query structure. These options are described in more detail below.

2. Molecular Formula Search. Historically, the molecular formula of each compound was derived from the connection table by summing the entries in column 2, the element column. The molecular formula so obtained contains no hydrogen atoms because hydrogen atoms are not explicitly described in the connection table.

This method has now been supplanted by a simpler process which uses the molecular formulas supplied by CAS as a part of the compound identification. These molecular formulas do, of course, include the number of hydrogens in the compound.

The molecular formulas are hash-encoded as has been described previously,[12] and the file of hash-encoded formulas vs. registry numbers is sorted, primarily upon the hash-encoded formulas, and secondarily upon the registry number. Pointers to the sorted file are generated, and the file of pointers together with the file of encoded formulas and registry numbers become the basis of the molecular formula search.

3. Special Properties Searches. A number of different items are included in the general category of special properties searches. All of these items are organized into a single set of hierarchically ordered files for searching purposes, but the different items are separately identified by property type. Currently, five different property types are included as follows.

a. Molecular Weight. The molecular weight of each compound is derived from the molecular formula provided by CAS as part of the compound identification. Searches may be conducted for a specific molecular weight or for all compounds whose molecular weights fall within a specified range of molecular weights.

b. Total Atom Count. The total atom count for a compound, as used here, is simply the total number of nonhydrogen atoms in the compound, which is the total number of atoms defined in the connection table. The ACOUN search may be used to identify all compounds having a specific total number of nonhydrogen atoms or with a total atom count that falls within a specified range.

c. Atom Population. The atom count for each element in the compound is also extracted from the molecular formula provided by CAS. This information forms the basis of the partial and ranged molecular formula search options which can be used to identify compounds having a specified partial molecular formula. In such a case, specific requirements are defined for some elements, but any number of other elements is also permitted. Alternatively, a partial formula can be defined as the permissible ranges of the appropriate elements.

d. Total Ring Count and Ring Population. These last two types of special property are concerned with the smallest set of smallest rings (SSSR) present in the compound. Both the searchable files and the query structure are examined by algorithms which can identify an SSSR correctly. These algorithms trace pathways through the connectivity section of the connection table (columns 3–8 of Figure 1) and locate the different smallest rings. Standard techniques for starting the tracing and continuing it once a ring has been closed are used in order to ensure that the rings located do indeed constitute an SSSR. Thus, in the example given in Figure 1,

4	C	C	9	C	9	C	1	0	0	2
3	C	C	9	C	1	0	0	0	0	4
3	C	C	9	C	9	0	0	0	0	6
4	C	C	L5	C	9	C	9	0	0	1
3	C	C	L5	C	9	0	0	0	0	2
3	C	O	1	C	1	0	0	0	0	2
3	O	C	1	C	1	0	0	0	0	1

Figure 3. Fragment table corresponding to the connection given in Figure 1.

the algorithms will discover that node 8 is connected to node 9. Node 9, in turn, is connected to node 6 which is connected to node 1, which is connected to node 7 which is connected back to node 8. In this way, the five-membered ring is traced, and in the same way the six-membered ring in the structure can be identified, but the larger nine-membered ring is ignored. The total ring count in this case, then, is two, the total number of rings constituting the SSSR. Similarly, the ring population information provides the numbers of rings of different sizes in the SSSR (one five-membered ring and one six-membered ring in the example of Figure 1). The RCOUN search permits location in the database of all compounds containing either a specified total number of rings or a given number of rings of a specific size. The same command can also be used to impose range requirements upon either of these criteria.

In the generation of the special properties file, the property type, property value, and registry number for each compound are combined into a file in that form. This file is then sorted, first upon property type, secondarily upon the property value, and finally upon the registry number. For search purposes, this inverted file is reorganized into a set of hierarchically ordered files which are accessed by the various commands described above (MW, ACOUN, partial MF, ranged MF, and RCOUN) to identify compounds in the database that fulfill the criteria described by the user.

4. Fragment Probe Search. The fragment probe permits a search through the database for all occurrences of a specific fragment defined by the user. This search operates on a file of fragment properties vs. registry number which is derived from the connection table in the following way. The connection table is scanned, an atom at a time, and a corresponding fragment table is built by a process involving an analysis of that atom and all its immediate neighbors. Each line in this derived fragment table contains (1) the size of the fragment, i.e., the total number of nonhydrogen atoms in the fragment, including the central atom; (2) the nature of the central atom; (3) the nature of the first neighbor; (4) the nature of the bond joining the central atom to the first neighbor; (5) the nature of the second neighbor, and so on until all neighbors have been described. The first neighboring atom to be considered is the one representing the least common element. In the substructure search files, carbon is clearly the most common element, followed by oxygen and nitrogen in that order. In the event that two neighbors are found to represent the same element, the one joined to the central atom by the less common type of bond is considered first. Thus if all the neighbors are carbon, then that carbon that is doubly bonded to the central atom is taken before any that are singly bonded to the central atom. Ten fields are available for this description, unused fields are filled with zeroes, and an eleventh field contains the number of times the fragment occurs in the molecule. When a fragment table entry is generated, the first ten fields of that entry are compared with all existing entries in the part of the table that has already been generated. If a complete match is found, the occurrence count for that entry (field 11) is incremented by one rather than creating a new entry in the table.

Storage of the fragments in this manner permits the

identification, during a search, of any compounds containing a sufficient number of occurrences of the fragment in question. One additional possibility must, however, also be considered. When a search is performed for a particular fragment, the occurrence of that fragment in a compound as a subset of a larger fragment as well as in a stand-alone manner, must be detected. To accomplish this, every node in the structure will be described in the new fragment table several times when the compound is entered into the database. It will first be entered with all its neighbors described. Then the node with one neighbor removed will be described, then a different neighbor will be removed, and so on, until all fragments from which one neighbor has been removed have been entered. If the atom, as it appears in the molecule, has four neighbors, then all the possibilities that result from the stripping of two neighbors will be computed and entered into this new fragment table. Every nonhydrogen atom is therefore described in this table at least once and possibly as many as ten times, depending upon the number and nature of its neighbors. Many of the multiple entries will, of course, be the same and can be merged by incrementing the counter in field 11. For this reason, the fragment table shown in Figure 3, which is derived in this way from the original connection table in Figure 1, has only seven lines.

Once the derived fragment table shown in Figure 3 has been generated, each line, with the appropriate registry number appended to it, is entered into the master database. When this database has been completely assembled, it is sorted on all fields, and the resulting inverted file is hierarchically organized for searching by the FPROBE option. Each level in the hierarchy is associated with one of the eleven fields described above and contains pointers for progressing to the next level of the search. At each level, the appropriate property in the query structure is sought in the database. If it cannot be located, the search fails, but if it is found, the search proceeds to the next level. The eleventh, and final, level contains a pointer to a list of registry numbers of compounds that satisfy the set of properties that were satisfied individually at each level.

If the bonds in the query structure are not all specified, it is possible for more than one database property to satisfy the query requirement. Multiple matches also normally occur at the eleventh level, since any number of occurrences of a fragment in a database structure in excess of the number of occurrences in the query structure will also satisfy the requirement. In these cases, multiple paths are traced to the bottom of the hierarchical structure, and each of the resulting registry number lists is merged to provide the composite set of responding compounds.

5. Ring Probe Search. The ring probe search permits the locating of structures containing a specific ring or rings, with or without heteroatoms, whose identity and position may be specified or not, and having a given substitution pattern. The search is hierarchical and properties are sought in the order given above. Compounds may be retrieved either because the query structure feature is imbedded in them or because they represent an exact match for the query structure.

As in the case of all the other search options, RPROB uses a file derived directly from the connection table shown in Figure 1 by the following series of steps. First the connection table is scanned to identify each nucleus (i.e., set of contiguous rings) in the compound. This is accomplished through a simple determination of which atoms are connected to which other atoms by ring bonds rather than chain bonds, for example. The numbering of the nodes is then reorganized according to a modified Morgan algorithm,[13] which tends to cause the numbering to radiate from the most central node. Figure 4 shows the nucleus from Figure 1, numbered arbitrarily (left) and according to the Morgan algorithm (right).

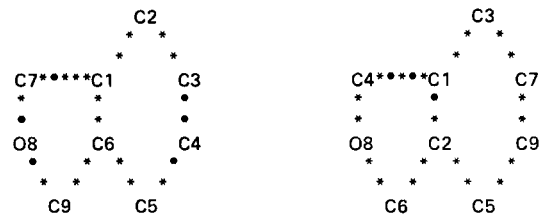


Figure 4. Query structure before (left) and after (right) application of the Morgan algorithm.

In the third step, the connection table for the renumbered nucleus is generated and fields 3–8, which describe the connectivities, are hash-encoded. It is this hash-encoded information that constitutes the first level in the hierarchical search. Next, the heteroatom positions are noted, then the heteroatom types, and finally the substitution pattern around the rings. These four levels of information for each compound are then assembled into a single table entry containing the hash code, a list of 12 heteroatom positions, a list of 12 heteroatom types, a list of 12 substituent positions, and, finally, the appropriate registry number.

The information described above is sufficient to permit the identification of complete nucleus structures within a database. However, to provide the information necessary to permit the identification of imbedded ring structures, i.e., a ring or a set of contiguous rings that is imbedded within a larger ring structure, additional processing is necessary. Path-tracing algorithms are used to identify the individual rings present in the nucleus. All possible combinations of contiguous rings are then formed, and each of these combinations is treated as described above. Thus a unique numbering is assigned using the modified Morgan algorithm, the connectivity information is hash-encoded, and heteroatom and substituent information is noted (bonds to other cyclic nodes which are not included in the set of rings being processed are treated as substituents). This information is then assembled and entered into the table as before.

When the entire database has been formatted in this way, it is sorted. The primary sort fields are those devoted to the hash code, and the subsorts are on heteroatom positions, heteroatom types, substituent positions, and registry numbers. This file is then hierarchically organized in a logically equivalent manner to the fragment files described above, and the resulting files are those that serve as the basis of the RPROBE search.

6. Substructure Search. While it is very useful to be able to learn that a specific fragment or ring is present in various compounds in the database, a more demanding query is whether or not a given complete structure is present, either per se, or imbedded in a larger structure. This is accomplished using the option SUBSTRUCTURE SEARCH.

This program conducts an atom-by-atom, bond-by-bond comparison between the connection table corresponding to the query structure and each of the connection tables in a selected subset of the database. This comparison is done without any subtlety and, in the case of the connection table shown in Figure 1, would proceed as follows. First, atom 1 in the table, i.e., C1 in the structure, is compared to node 1 in the first database connection table that is to be examined. The first check is that both C1 and node 1 represent carbons. If this is not so, then C1 of the query connection table is compared to node 2 of the database connection table. Once two identical atom types, one from each connection table, are found, their neighbors are compared to each other. If the neighbors are not the same, then that node of the database connection table is dropped and a new node is examined. If the neighbors do match, then the various bonds between the central atoms and

their respective neighbors are checked. Again, lack of correspondence would cause this pair of atoms to be abandoned, but if the match is perfect, the program proceeds to the next atom in the query structure and repeats the entire process. In this way, every atom in the query structure will, if necessary, be compared to every atom in the database structure, and all bonds will also be compared. Only if the query structure is exactly the same as the database structure or, if an exact copy of the former is located within the latter, will the structure be retrieved from the database as a positive response to the user's question.

The ability of this program to locate and produce compounds in which a specified partial structure is imbedded is very powerful because this is precisely the type of query that chemists are prone to pose. A very typical request, for example, is for all compounds containing a 1-fluoro-3-bromophenyl ring. The fragment probe will allow retrieval of all structures containing an aromatic fluorine and an aromatic bromine, but this search will not guarantee that the fluorine and bromine be in the same ring. The ring probe will permit retrieval of all meta-substituted aromatic ring compounds, but again, the halogen atoms need not be in the same ring. Only the substructure search can limit this list to just those compounds containing the 1-fluoro-3-bromophenyl moiety.

Because of the bluntness of the substructure search program, it can use considerable amounts of processor time, and the most sensible way in which to use this program appears to be to anticipate its application by performing the appropriate fragment or ring probes. In this way, the large database can be reduced to a smaller file of candidate structures known to contain the desired fragments and/or rings. The substructure search program can operate on such a file without incurring intolerable expense, and it can extract from this file just the compounds which represent correct responses to the original query.

7. Structural Feature Code Search. As a somewhat different approach to the problem of identifying compounds having various combinations of structural properties, the substructure search system also contains a structural feature code search capability. The structural feature codes, extracted from the CIDS chemical search keys of the U.S. Army CIDS System,[14] consist of a large, somewhat open-ended set of both generic and specific predefined structural characteristics. Generally, several thousand unique codes will be found in a database of a reasonable size. The codes applicable to each compound in the database are automatically assigned. An exhaustive analysis and comparison of the connection table for the compound with the set of predefined structural feature codes is necessary to accomplish this assignment. As in the other types of searches, these codes, along with the registry numbers to which they apply, are sorted to create an index with respect to code. From that index, a set of hierarchically organized files are created for searching purposes.

To use the structural feature code search capability, one must first establish which of the codes are appropriate to the question at hand. The codes can then simply be entered into the search system, which will in turn identify the compounds to which these codes have been assigned. Through the intersection and merging of the results associated with the searches for individual codes, compounds possessing any arbitrarily complex logical combination of the codes can be identified.

RESULTS

In this section, a number of examples of use of the substructure search system are given. All the examples given here are of searches through the structural data base that corresponds to the NIH–EPA Mass Spectral Search System.[15] This database currently contains just under 30 000 distinct compounds and their respective low-resolution mass spectra.

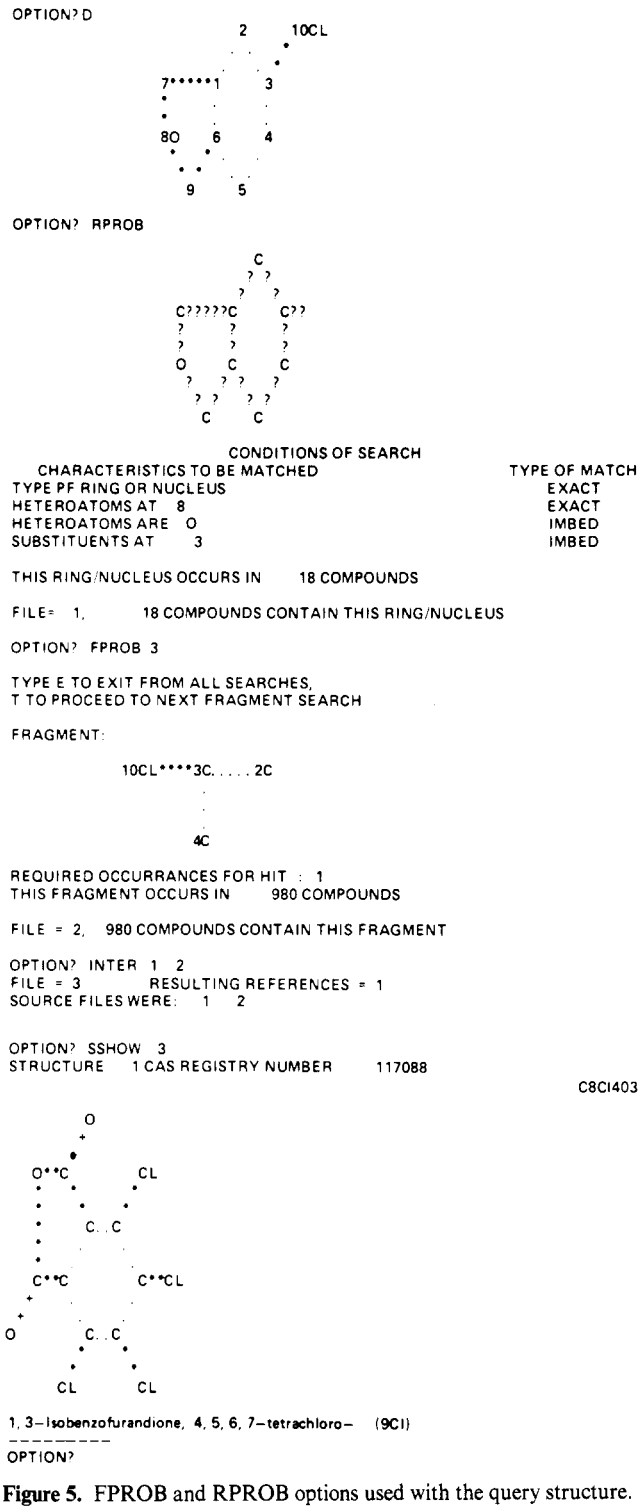


Figure 5. FPROB and RPROB options used with the query structure.

In the first case, which is shown in Figure 5, several searches were carried out using the query structure given in Figure 1. The first search used RPROB with the following match conditions implied: the ring systems retrieved should be identical with that in the query structure, no imbedment was to be permitted, node 8 and no other node may be a noncarbon, and, finally, there should be substituents at least at node 3. The search for this nucleus produced scratch file number 1 with 18 candidate structures. Next, a fragment probe for all compounds containing a node identical with C3 of the query structure (i.e., the chlorine-bearing carbon) was carried out. This resulted in scratch file number 2, with 980 structures.

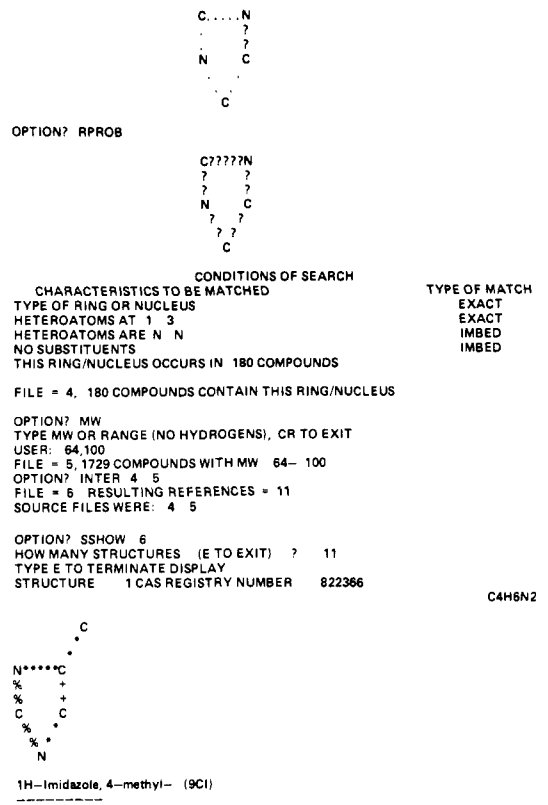


Figure 6. Use of RPROB in conjunction with a molecular weight range specification.

Intersection of these two files resulted in a third file containing a single compound, 1,3-isobenzofurandione, 4,5,6,7-tetrachloro-, that satisfied all the criteria defined. In a final command, SSHOW, this structure, its name, registry number, and molecular formula were displayed.

In a different method of searching, shown in Figure 6, the intention was to locate all low molecular weight derivatives of imidazole. An RPROB search led to the retrieval of 180 compounds containing a five-membered ring with two noncarbon atoms in a 1,3 disposition. This was followed by a command to locate all compounds in the database that have molecular weights between 64 and 100. There are 1729 such compounds, but the subsequent intersection showed that only 11 of these were also in the file that resulted from the RPROB search. The first of these, 1*H*-imidazole, 4-methyl-, registry number 822366, was printed out with the SSHOW command.

Finally, use of the powerful structural feature codes is shown in Figure 7. In this search, the objective was to find all the carotenes in the database. These are compounds that contain two cyclohexene rings joined together by a long (e.g., C18) olefinic chain. The first code that was used, 13,4, implies at least four occurrences of chain branching, i.e., a carbon bonded to at least three other carbons. Eighty compounds fulfilled this criterion. Next, a much more stringent requirement, the occurrence of 9 olefinic bonds was defined with code 11,9, and this produced only 16 hits. In a final criterion, code 71,50,2,2 two, and no more nor less than two, cyclohexene rings were requested. This gave 18 hits. The automatic intersection led to just seven compounds, all carotenes, that met all three requirements, and the first of these, β,ϵ-Carotene-3,3′-diol, (3*R*,3′*R*,6′*R*)-, registry number 127402, was listed with the SSHOW command. In this case, the structure was drawn on a CRT terminal, which permits a better picture.

As can be seen from the foregoing examples, the ways in which queries can be posed to a chemical structure searching system are varied; the chemist can often use nonstructural information, such as an upper limit to molecular weight, to

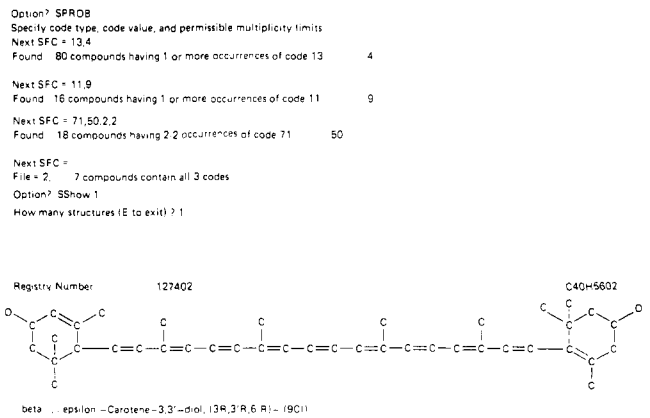


Figure 7. Search using structural feature codes.

aid in the convergence of a search. This substructure search system has been developed with this in mind. The various search options that are available permit the user to make the greatest use of all the information at his disposal and so complete searches rapidly and efficiently.

SUMMARY

The system that has been described is interactive, and it is this property that is one of its most important features. It is not difficult to learn to use, and once some experience has been gained, queries can be framed very rapidly. At this point, the speed with which answers to the queries are provided becomes extremely valuable because the answers often form the basis of the subsequent query.

As is clear, this is a fairly large program package, and it has not been designed for facile export or transfer from one computer to another. Rather, it is expected that the substructure search system will be most accessible via a networked time-shared computer system, and it is, in fact, already available in this form.[16]

REFERENCES AND NOTES

(1) R. Fugmann, W. Braun, and W. Vaupel, *Angew. Chem.*, **73**, 745 (1961); R. Fugmann in "Chemical Information Systems", J. E. Ash and E. Hyde, Ed., Wiley, New York, N.Y., 1975, Chapter 13.
(2) D. R. Eakin and E. Hyde in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldmann, and E. Hyde, Ed., Wiley, New York, N.Y., 1974, pp 1–30; D. R. Eakin in "Chemical Information Systems", J. E. Ash and E. Hyde, Ed., Wiley, New York, N.Y., 1975, Chapter 14.
(3) G. F. Hazard and S. Richman, paper presented at the 176th National Meeting of the American Chemical Society, San Francisco, Calif., Sept 1976.
(4) D. P. Jacobus, D. E. Davidson, A. P. Feldman, and J. A. Schafer, *J. Chem. Doc.*, **10**, 135 (1970).
(5) R. J. Rowlett and F. A. Tate, *J. Chem. Doc.*, **12**, 125 (1972).
(6) M. Milne, D. Lefkovitz, H. Hill, and R. Power, *J. Chem. Doc.*, **12**, 183 (1972).
(7) R. J. Feldmann and S. R. Heller, *J. Chem. Doc.*, **12**, 48 (1972).
(8) EPA Internal Regulation No. 2800.2, 1976.
(9) S. R. Heller, G. W. A. Milne, and R. J. Feldmann, *Science*, **195**, 253 (1977).
(10) G. W. A. Milne and S. R. Heller, *J. Chem. Inf. Comput. Sci.*, **16**, 232 (1976).
(11) R. J. Feldmann in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldmann, and E. Hyde, Ed., Wiley, New York, N.Y., 1974, pp 55–81.
(12) S. R. Heller, *Anal. Chem.*, **44**, 1951 (1972).
(13) H. L. Morgan, *J. Chem. Doc.*, **5**, 107 (1965).
(14) "Handbook of CIDS Chemical Search Keys", Fein-Marquart Associates, Inc., Towson, Md., Nov 1973.
(15) S. R. Heller, H. M. Fales, and G. W. A. Milne, *Org. Mass Spectrom.*, **7**, 107 (1973); S. R. Heller, D. A. Koniver, H. M. Fales, and G. W. A. Milne, *Anal. Chem.*, **46**, 947 (1974); S. R. Heller, R. J. Feldmann, H. M. Fales, and G. W. A. Milne, *J. Chem. Doc.*, **13**, 130 (1973); R. S. Heller, G. W. A. Milne, R. J. Feldmann, and S. R. Heller, *J. Chem. Inf. Comput. Sci.*, **16**, 176 (1976).
(16) The system is available for general use via the TYMSHARE computer network. For further details, please contact AEF.