The NIH–EPA Structure and Nomenclature Search System

G. W. A. MILNE*
National Institutes of Health, Bethesda, Maryland 20014

S. R. HELLER
Environmental Protection Agency, Washington, D.C. 20460

A. E. FEIN, E. F. FREES, R. G. MARQUART, J. A. MCGILL, J. A. MILLER, and D. S. SPIERS
Fein-Marquart Associates, Towson, Maryland 21212

Received July 12, 1978

The NIH–EPA Structure and Nomenclature Search System (SANSS) contains 110776 different compounds from 40 separate files. The structure records have been merged into a single unified file as have the nomenclature records. It is possible, using the interactive programs of the Structure and Nomenclature Search System, to retrieve all references to a particular chemical name, structure, or substructure within these 40 files of chemicals. The NIH–EPA Chemical Information System, of which the Structure and Nomenclature Search System is a central part, is available to the international scientific community via a timeshared, networked computer.

INTRODUCTION

In an earlier paper in this series,[1] we described the design of the first publicly accessible version of the family of interactive programs that comprised the NIH–EPA Substructure Search System, a component of the NIH–EPA Chemical Information System.[2] Since then, some fundamental changes in the design of the Substructure Search System have been accomplished, and it is the purpose of this paper to describe these in detail.

Based upon programs developed originally by Feldmann,[3] the Substructure Search System (SSS) permitted the searching through a file of chemical substances for a specific structure or substructure. To be made searchable by this system, a file, or list of chemical substances, was first sent to the Chemical Abstracts Service (CAS) where every compound was assigned a CAS Registry Number,[4] using methods that have been described in detail elsewhere.[5] Once the Registry Number was known the nomenclature associated with the substance (index names and synonyms) was extracted from the Registry maintained by CAS, and a copy of the connection table was derived from the same source. The connection table is a two-dimensional matrix which tabulates every nonhydrogen atom in the molecule, listing each of its neighbors and each of the bonds associated with the atom. The SSS basically allowed one to search through the assembled connection tables for a specific type of atom (e.g., an aromatic carbon bearing a bromine or a carbonyl carbon flanked by a fluorine) and locate all compounds that contained such an atom. Related programs within the SSS permitted one to locate all compounds having specific types of rings, molecular weights, or partial or complete molecular formulas. Every search resulted in a temporarily stored file of Registry Numbers of those compounds which fulfilled the particular criterion; and, since the files could, upon command, be combined in the Boolean AND, OR, or NOT sense, the resulting system allowed searches for any chemical substance, defined precisely or otherwise. It was quite simple, for example, to locate all the tetracyclic compounds that contain just one aromatic ring substituted by two fluorines in an ortho relationship.

In June 1977, this version of the SSS was made generally available for use on a fee-for-service basis via a commercial networked computer. At that time, it was possible to search through five distinct files of chemicals, including the files of the CIS mass spectral and carbon-13 NMR spectral databases as well as the Inventory Candidate List that had been developed from CIS files for use in connection with the Toxic Substances Control Act (TSCA). Considerable general use has been made of that system, during which it became clear that the TSCA Candidate List was being searched far more than the other files.

Table I. Types of Search Available in SANSS

Program	Effect
IDENT	To search for a specific complete structure.
SUBSS	To search for a specific substructure.
NPROBE	To search for a complete or partial name.
FPROBE	To search for specific atom-centered fragments.
RPROBE	To search for specific rings or ring systems.
SPROBE	To search using the CIDS keys screens.
MW	To search for specific molecular weights.
MF	To search for specific complete or partial molecular formulas.
RCOUNT	To search for compounds having a specific number of rings.
ACOUNT	To search for compounds having a specific number of atoms.

It also became clear that a weakness of the system was that it was necessary, prior to the search, to decide which of the files one wished to search, and to search through all five files, five separate searches were necessary. A further problem was that it was relatively difficult to locate common compounds such as cholesterol or camphor, which have well-accepted trivial names but quite complex structures. Further, it was also impossible to search directly for an exact match to one's query structure.

These deficiencies have now been remedied and the resulting system permits searching through many files for substances identified by name, structure, substructure, or chemical "screens", as shown in Table I. The SANSS is therefore operating along precisely the lines defined in the MITRE report[6] in connection with the Chemical Structure and Nomenclature Search System. A name search has been installed

0095-2338/78/1618-0181$01.00/0

in the SSS, and there is also an IDENTITY search which will search for an exact match for the query structure. As a result, the name of the Substructure Search System has been changed to Structure and Nomenclature Search System (SANSS), which more exactly describes the capabilities of the programs.

A more substantial change is that all the previously independently searchable files, 40 in number, have been merged into a "Unified Database" thus obviating the file selection process described above. Searches are now conducted on the full Unified Database, and, as before, the answers to searches are the Registry Numbers of the retrieved compounds. In addition, however, information as to which of the 40 files contain information concerning these compounds is also available. The user can, prior to a search, specify the files in which there is interest and if this is done, substances that are only found in other files will be ignored. The fact that all searches are now conducted through the large Unified Database, rather than small, selected databases, adds slightly to the computational overhead, but as is discussed below, the increase in search costs is minor and this approach seems to represent the better of the two alternatives. Moreover, it results in a substantial reduction in storage costs. As a result of the merger, the Unified Database now contains 110 776 different compounds, representing 40 different files. These changes to the original SSS are described in detail in this paper.

DISCUSSION

1. Unified Database. When a search of any sort is carried out in the SANSS, the system will locate all compounds which fulfilled the criteria specified by the user. What actually is retrieved by the SANSS is a list of the CAS Registry Numbers of these compounds. If the user then seeks to examine this subset of compounds, the program will use the Registry Numbers, one-by-one, to look up the names and various other information describing the compounds. Whichever of these are requested are then presented to the user for inspection. In the earlier version of the system, referred to above, each file of compounds was searched independently of the others and the search resulted only in the Registry Numbers of those compounds in that one particular file. The Registry Numbers, in turn, could then be used to retrieve the names and structures of the compounds.

During 1977, it was decided to merge together all the available files and for each compound, it became necessary to retain, not only the Registry Number, but also the information as to which of the source collections contained information about that compound. Thus a substance such as benzene, which is in many files of the SANSS, will be stored once in the Unified Database, with the notation that its Registry Number is 71-43-2 and in addition that, together with its Registry Number, it is to be found in certain specific files of the SANSS. Every chemical described in the SANSS therefore has an identifier (the Registry Number) and also a subidentifier (the names or numbers of the files that contain it).

With the exception of file 7 (the Merck Index), the entire Unified Database of names, CAS and other synonyms, and connection tables is in the public domain.[7] The various component files of the Unified Database are given in Table II, which contains the name and number of each file, together with the number of substances it contains. The important advantages of this approach are that the whole SANSS can be searched with a single command and that the data associated with each unique compound need be stored only once. Prior to the merging of the component databases, there were in the SSS, 40 separate files containing a total of 173 817 entries. The number of unique compounds in the system, however, was only 110 776. Thus 36.35% of the stored data were redundant. Unification of the databases has made it possible to save almost all of this storage, and the resulting decrease in the cost to the end user has been considerable.

Table II. Component Files of the Unified Database

#	Name	Substances
1	EPA - TSCA Inventory Candidate List.	33449
2	NIH-EPA CIS Mass Spectral Data Base.	25544
3	NIH-EPA CIS Carbon-13 NMR Spectral Data Base.	6505
4	EPA - Pesticides, Active Ingredients.	1444
5	EPA - OHM-TADS.	858
6	Cambridge X-Ray Crystallography Data Base.	16742
7	Merck Index.	8959
8	EPA - Pesticides, Analytical Reference Standards.	562
9	EPA - STORET.	234
10	EPA - Chemical Spills.	577
11	EPA - SOTDAT.	572
12	NIMH - Psychotropic Drugs.	2214
13	EPA - SAROAD.	65
14	NBS - Gas Phase Proton Affinities Data Base.	514
15	CPSC - CHEMRIc File.	890
16	EPA - Pesticides, Inactive Ingredients.	735
17	NBS Heats of Formation of Gaseous Ions Data Base.	3156
18	National Fire Prevention Association List of Chemicals.	396
19	FDA/EPA - Pesticides Reference Standards.	613
21	International Trade Commission List of Chemicals.	9140
22	NBS - Single Crystal Data Base.	18286
25	EPA - Effluent Guidelines.	118
26	EPA - Organic Chemical Producers.	375
27	IPC - Chemical Production.	104
28	IPC - Chemical Plant.	103
29	NSF - List of Environmentally Hazardous Chemicals.	225
30	University of Tokyo - EROICA Data Base.	4492
31	PHS-149 List of Carcinogens.	4425
32	NIOSH Registry of Toxic Effects of Chemicals.	19891
33	NIOSH National Occupational Hazard Survey.	4560
35	Environmental Mutagen Information Center Data Base.	4030
36	Environmental Teratogen Information Center Data Base.	3250
43	EPA - Selected Organic Air Pollutants.	578
45	EPA - Section 112 of Clean Air Act.	5
58	NCTR - Potential Industrial Carcinogens and Mutagens.	91
59	EPA - List of Environmental Carcinogens.	27
66	EPA - List of Hazardous Pesticides.	22
67	EPA - Mutagenicity Studies.	25
70	CIIT - List of Candidates.	26
82	NOAA - Microconstituents of Fish and Fishery Products.	15
	Total	173817

Offsetting this positive result, there has been an impact upon the actual searching through the system. The whole database of 110 776 substances, rather than a component file of perhaps 10 000 substances, is being searched. Searching is based upon an inverted file concept,[8] and so the cpu demand of a search is not changed markedly because the file is bigger. The number of retrievals to most searches is now considerably higher, however, and this places a burden upon the intersection and merge routines that are implicit in many of the searching programs. Mainly because of this, the absolute cost of carrying out a search has increased, as a result of the merging of the databases, by between 10 and 50%. If the costs are expressed, however, on a per-compound-searched basis, they can be seen to have decreased by about 50% because the number of compounds that are searched in any operation has increased greatly. The typical commercial cost of a complete structure search or a name search is comfortably under $2.00.

The merging of the databases has had a number of other, less important effects that are apparent to users. At the beginning of a session, the user is informed that, unless the program is otherwise notified, all searches will be conducted through the full file. At this point, an opportunity is provided to select a smaller number of search files. In order to take advantage of this, the user must know the identifying numbers of the files in which searches are desired. Consequently, a new HELP file listing the files and their identification numbers has been added to the system. This HELP file identifies each of the 40 files that comprise the Unified Database.

Background information concerning each of the component files has also been added to the system. For example, the user who wishes more information about File 32, the NIOSH Registry of Toxic Effects of Chemical Substances, can type "HELP 32" in response to the "OPTION:" prompt, and a

```
OPTION? NPROBE

FRAGMENT OR WHOLE NAME SEARCH  (F/W)  (F)?F

SPECIFY FRAGMENT (CR TO EXIT):     SASSAFRAS

FILE     1,        1 COMPOUNDS HAVING FRAGMENT:  SASSAFRAS
SPECIFY FRAGMENT (CR TO EXIT):

OPTION? SSHOW 1

STRUCTURE INFORMATION IS NOT AVAILABLE FOR REGISTRY NUMBER 6179Ø-23-6
STRUCTURE    1 CAS REGISTRY NUMBER 6179Ø-23-6
TSCA Candidate List: R325-6667
U.S. International Trade Commission
                                                          W99

Oils, sassafras, hydrogenated ()
Sassafras oil, hydrogenated
---------
OPTION?
```

Figure 1. Use of NPROBE with a name fragment.

554-word description of the RTECS file will then be returned to his terminal. This description includes the name, address, and telephone number of an individual at NIOSH who can provide even more detailed information.

When a search has been completed, the command SSHOW will list the compounds that were retrieved. Also listed, at the user's option, are Registry Numbers, structural and molecular formulas, and alternative names or synonyms. It is also possible to list the component files in which each compound is represented. It is possible, at the user's discretion, to list these files by name or number, or both or neither.

A difficulty arises when a compound is present more than once in a component file. The carbon-13 NMR file, for example, often has more than one entry for a given compound, because its spectrum may have been measured in different solvents. This does not happen in many component files, but when it does, a system of local identifiers must be used to permit distinction between the various occurrences of the same compound.

2. Nomenclature Search. The structural searching capabilities of SANSS are very powerful and appear to be the search method of choice for many chemists. A need exists, however, for a program which permits a search for substances by name or synonym, and accordingly, a nomenclature search has now been developed.

Chemical names may be single words, groups of words separated by "natural" delimiters, typically spaces, groups of words separated by special characters such as hyphens or parentheses or (in the case of many TSCA class 2 chemicals) paragraph descriptions of syntheses/processes. Names in the first two of these groups may be searched for relatively easily, and, once a definition of delimiters is established, the third group presents no particular problem.

If the program is told that the name supplied by the user is a complete name, it can search through a set of inverted files for each occurrence of exactly that character string. It is likewise a simple matter to decompose a multiword name into its component words and build an inverted file from these partial names. If the user indicates that the name supplied is a partial name, the search program simply accesses an inverted file of those name fragments for retrieval purposes. While it is unlikely that irrelevant retrievals will result from a search with the full name, such an outcome is possible with a partial name. The hits from all name searches are stored in temporary files which can be inspected by the user, who may select the compounds which were actually sought. In this sense, the name search, NPROBE, operates in just the same way as, for example, the fragment probe search, FPROBE.

An example of the use of NPROBE is shown in Figure 1. Here the user enters the name fragment SASSAFRAS. This, of course, is an utterly trivial name, but it is found as part of a name of the generically registered material, hydrogenated sassafras oil. The retrieval is reported to the user who, inspecting temporary file number 1, finds that the substance does indeed have a Registry Number, is to be found in the TSCA Inventory and the International Trade Commission List, but has no known molecular formula (W99 is used to encode this fact) or structure.

```
OPTION? NPROBE

FRAGMENT OR WHOLE NAME SEARCH  (F/W)  (F)?W

SPECIFY NAME (CR TO EXIT):        DHA

FILE     1,        1 COMPOUNDS HAVING NAME:      DHA
SPECIFY NAME (CR TO EXIT):

OPTION? SSHOW 1

STRUCTURE    1 CAS REGISTRY NUMBER 52Ø-45-6
TSCA Candidate List: RØ8Ø-5132
CIS Mass Spectrometry
EPA Pesticides - Active Ingredients: 278Ø1
Merck Index
U.S. International Trade Commission
NBS Xray Crystallography: 52Ø-45-6.Ø1 TO 52Ø-45-6.Ø2
PHS-149 Carcinogenic Activity: BØ787
NIOSH RTECS: UP8Ø5ØØ
                                                          C8H8O4

          O
         * *
        *   *
    O++C     C**C
       *     +
       *     +
C**C**C      C
    +   *    *
    +    * *
    O     C
          +
          +
          O

2H-Pyran-2,4(3H)-dione, 3-acetyl-6-methyl- (8CI9CI)
Acetic acid, dehydro-
Dehydracetic acid
Dehydroacetic acid
DHA
DHS
3-Acetyl-6-methyldihydropyrandione-2,4
4-Hexenoic acid, 2-acetyl-5-hydroxy-3-oxo-, .delta.-lactone
---------
OPTION?
```

Figure 2. Search for the full name "DHA".

```
OPTION? SELECT

COMPLETE DATA BASE SELECTED

ENTER NEW SELECTION (H FOR HELP): 5 9 1Ø 31

COLLECTIONS SELECTED:   5   9  1Ø  31
OPTION? NPROBE

FRAGMENT OR WHOLE NAME SEARCH  (F/W)  (F)?W

SPECIFY NAME (CR TO EXIT):        TOLUENE

FILE     1,        1 COMPOUNDS HAVING NAME:      TOLUENE
SPECIFY NAME (CR TO EXIT):

OPTION? SSHOW 1

STRUCTURE    1 CAS REGISTRY NUMBER 1Ø8-88-3
EPA OHM/TADS: 72T16928
EPA STORET: 34Ø1Ø
EPA Chemical Spills
PHS-149 Carcinogenic Activity: AØ369
                                                          C7H8

      C
     . .
    .   .
   C     C
   .     .
   .     .
C**C     C
    .   .
     . .
      C

Benzene, methyl- (9CI)
Toluene (8CI)
Antisal 1a
Methacide
Methylbenzene
Methylbenzol
Phenylmethane
Toluol
---------
OPTION?
```

Figure 3. Search for "toluene" in selected files.

A more informative result is shown in Figure 2, where a full name, DHA, is entered by the user. This, it transpires, is an acronym for dehydroacetic acid, which is in eight of the files, has eight synonyms, including DHA, and has the structure and molecular formula shown.

As has been discussed above, it is possible to limit the retrieved information to just that pertaining to specific files.

```
OPTION? NPROBE

FRAGMENT OR WHOLE NAME SEARCH  (F/W)  (F)?W

SPECIFY NAME (CR TO EXIT):          TOLUENE

FILE     1,        1 COMPOUNDS HAVING NAME:       TOLUENE
SPECIFY NAME (CR TO EXIT):

OPTION? SSHOW 1

STRUCTURE    1 CAS REGISTRY NUMBER 1Ø8-88-3
TSCA Candidate List: RØ38-8579
CIS Mass Spectrometry
CIS Carbon 13 NMR Spectrometry: 1Ø8-88-3.Ø1 TO 1Ø8-88-3.Ø4
EPA Pesticides - Active Ingredients: 8Ø6Ø1
EPA OHM/TADS: 72T16928
Merck Index
EPA STORET: 34Ø1Ø
EPA Chemical Spills
EPA AEROS SOTDAT: 71Ø2,52Ø2
NBS Proton Affinities
CPSC CHEMRIC
NBS Gaseous Ions
NFPA Hazardous Chemicals: 49284
U.S. International Trade Commission
EPA Organic Chemical Producers: 3349
NSF Hazardous Chemicals List: A11,13
EROICA Thermodynamics
PHS-149 Carcinogenic Activity: AØ369
NIOSH RTECS: XS525ØØ
---------
OPTION?
```

Figure 4. Search for "toluene" in the Unified Database.

An example of this approach is shown in Figure 3. Here, the SELECT command was used to restrict retrievals to files 5, 9, 10, and 31. The first three of these are EPA files of pollutant chemicals, and the fourth is the PHS-149 List of Carcinogens. Use in NPROBE of the full name TOLUENE, followed by the SSHOW command, reveals that toluene, which has seven other names, is found in all four of the files. If, as is shown in Figure 4, retrieval is permitted from all files, then toluene is found to be in 19 files. In this example, the SETDC command has been used to limit the display to the names of the files in which toluene is found; the names, structure, and formula have been suppressed.

If the user wishes to conduct a search using part of a name, i.e., a name fragment resulting from the breaking of a name at some point other than a "natural" delimiter, this is possible, and, provided the beginning of the name or the name fragment is preserved, is not a particularly difficult task. The break in the name is denoted with a colon; thus BENZ: will retrieve benzene and benzophenone, but BENZO: will fail to retrieve benzene. Searches for a right-truncated name are conducted through the same inverted file as the partial name search described above and can have some interesting results. In Figure 5, an NPROBE search is shown in which the user first establishes that a name fragment is to be entered. This is typically one word from a multiword name. Then, a right-truncated fragment, DEXTROPIMAR:, of that one word, is entered. That this has been truncated is signified by the colon terminator. Two compounds are found in this search and use of the SSHOW command reveals that they are closely related to one another; the first is the naturally occurring diterpene, dextropimaric acid, and the second is the corresponding aldehyde, dextropimarinal.

The user may wish to omit specifying the beginning of a name, entering, for example, :ENZENE or :ENZ:, and in this case, the search is much more difficult. A file sorted on the full name is used for the full name search, but this approach would be very expensive for use with left-truncated names like :ENZENE, because a separate entry would be needed for each letter in each of the names for each compound. Instead, the left-truncated name, with or without right truncation, is accepted and the program simply searches sequentially through the file of full names. The search is slow and expensive, and for this reason, the system will not permit a search for a left-truncated name through the complete Unified Database. Such searches can only be carried out within temporary files that result from some other search. In this way, pressure is

```
OPTION? NPROB

FRAGMENT OR WHOLE NAME SEARCH  (F/W)  (F)?F

SPECIFY FRAGMENT (CR TO EXIT):     DEXTROPIMAR:

FILE     2,        2 COMPOUNDS HAVING FRAGMENT:  DEXTROPIMAR:
SPECIFY FRAGMENT (CR TO EXIT):

OPTION? SSHOW 2

HOW MANY STRUCTURES (E TO EXIT) ? 2

TYPE E TO TERMINATE DISPLAY
STRUCTURE    1 CAS REGISTRY NUMBER 127-27-5
                                                    C2ØH3ØO2

   O
   %
   %
O%%C     C**C
   *    *    *
   *   *      *
   C**C        C++C
  **   *      *    *
 * *    *    *      *
C  C  C**C**C     C**C**C++C
 *      *    *      *
  *    *      *    *
   C**C        C**C

1-Phenanthrenecarboxylic acid, 7-ethenyl-1,2,3,4,4a,4b,5,6,7,9,1Ø,1Øa-
     dodecahydro-1,4a,7-trimethyl-, [1R-(1.alpha.,4a.beta.,4b.alpha.,7
     .beta.,1Øa.alpha.)]- (9CI)
Podocarp-8(14)-en-15-oic acid, 13.alpha.-methyl-13-vinyl- (NT18CI)
.alpha.-Pimaric acid
.delta.8(14)-Pimaric acid
d-Pimaric acid
Dextropimaric acid
---------
STRUCTURE    2 CAS REGISTRY NUMBER 472-39-9
                                                    C2ØH3ØO

   O
   +
   +
   C
   *
   *
   C     C
  ***   * *
 * * * *   *
C  C  C     C
*     *     *
*     *     *
C     C     C
 *   ***   * +
  * * * * *   +
   C  C  C     C
         *     *
         *     *
         C  C**C**C++C
          *   *
           * *
            C

1-Phenanthrenecarboxaldehyde, 7-ethenyl-1,2,3,4,4a,4b,5,6,7,9,1Ø,1Øa-d
     odecahydro-1,4a,7-trimethyl-, [1R-(1.alpha.,4a.beta.,4b.alpha.,7.
     beta.,1Øa.alpha.)]- (9CI)
Podocarp-8(14)-en-15-al, 13.alpha.-methyl-13-vinyl- (8CI)
Dextropimarinal
Pimaral
Pimarinal
---------
OPTION?
```

Figure 5. Search for a right-truncated name.

applied to the user to conduct a preliminary screen before embarking on a search for a left-truncated name.

When NPROBE is invoked, the user must specify whether the search is to be for a whole or a fragment name, and then the name must be entered. If the first character is a colon, the program recognizes that a left-truncated name is being entered and asks for the number of the temporary file that is to be searched. After this is provided, the search begins. As with the Substructure Search option,[1] the program reports at regular intervals on the number of names searched and the number of hits, and the user may cleanly terminate the search at any time.

As with any other search option in the system, NPROBE stores retrieved substances in the form of a list of Registry Numbers in a temporary file. This temporary file can be logically combined with other files, or displayed, at the discretion of the user.

3. Selection of Files from the Unified Database. The SELECT command permits users to restrict retrievals to

```
OPTION? RPROBE

   C??C        C??C
   ?   ?      ?    ?
   ?    ?    ?      ?
   ?     C??C         C??C
   ?    ?    ?      ?    ?
   ?   ?      ?    ?      ?
 ??C??C         C??C??      C??
    ? ?       ?    ?       ?
     ?   ?    ?      ?    ?
          C??C         C??C


                CONDITIONS OF SEARCH
    CHARACTERISTICS TO BE MATCHED              TYPE OF MATCH
TYPE OF RING OR NUCLEUS                            EXACT
NO HETEROATOMS                                     EXACT
SUBSTITUENTS AT  3 13 15  6                        EXACT
THIS RING/NUCLEUS OCCURS IN   598 COMPOUNDS

FILE =  1,   598 COMPOUNDS CONTAIN THIS RING/NUCLEUS

OPTION? PULL 7

FILE  2 CREATED, CONTAINS   8959 COMPOUNDS
SOURCE COLLECTION :   7
OPTION? INTER 1 2

FILE =  3     RESULTING REFERENCES =    77
SOURCE FILES WERE:   1   2
OPTION? NPROB

FRAGMENT OR WHOLE NAME SEARCH  (F/W)  (F)?F

SPECIFY FRAGMENT (CR TO EXIT):      :STEROL

TEMPORARYFILE (E TO EXIT)?: 3

DOING DETAILED NAME SEARCH
TYPE E TO EXIT
SEARCHED   1Ø COMPOUNDS,     3 HITS SO FAR
SEARCHED   2Ø COMPOUNDS,     7 HITS SO FAR
SEARCHED   3Ø COMPOUNDS,    11 HITS SO FAR
SEARCHED   4Ø COMPOUNDS,    19 HITS SO FAR
SEARCHED   5Ø COMPOUNDS,    22 HITS SO FAR
SEARCHED   6Ø COMPOUNDS,    25 HITS SO FAR
SEARCHED   7Ø COMPOUNDS,    25 HITS SO FAR
FILE     4,       27 COMPOUNDS HAVING FRAGMENT:  :STEROL
SPECIFY FRAGMENT (CR TO EXIT):

OPTION? NOT 3 4

FILE =  5     RESULTING REFERENCES =    5Ø
SOURCE FILES WERE:   3   4
OPTION?
```

Figure 6. Example of RPROBE, PULL, and NPROBE search for a left-truncated name.

specified files, as has been explained above. If a certain file has been selected, that fact is used by the program during a search, and substances that meet the search criteria are retrieved only if they are in the specified file. It is therefore impossible to impose a file selection after a search has been done, but there is another means of achieving the same result. If a search has been carried out through some number of files, or even through the whole Unified Database, and it is then realized that only substances in one particular file are of interest, it is possible, using the PULL command, to extract the Registry Numbers of that entire file and set them up in a temporary file. Once that is done, an INTERsect between that file and the earlier file of answers to a search will reduce the latter to only those Registry Numbers retrieved in the search that are in the file of interest.

An example of this procedure, in conjunction with NPROBE is shown in Figure 6. Here, an RPROBE search is conducted for all substances in the Unified Database having the basic steroid skeleton. This results in file 1, which contains 598 compounds. In a second command, the Merck Index file (No. 7) is "PULLed", and all 8959 substances are retained in temporary file 2. INTERsection of files 1 and 2 gives file 3, which contains 77 compounds which possess the steroid skeleton and which are also in the Merck Index.

This file can now be searched for all compounds with names ending in ". . .STEROL". The fragment entered, :STEROL, is a left-truncated name, and so the program asks for a temporary file to be identified. The user asks that file 3 be searched and the search begins. After the names for ten substances have been inspected, the program reports this fact and also notes the number of times a name ending in STEROL was found. Like SUBSSS, this search can be cleanly terminated at any time by the user.

A total of 27 of the 77 substances are found in this search and these are all stored in file 4. This file can be combined

```
OPTION? SSHOW-4

HOW MANY STRUCTURES (E TO EXIT) ? 3

TYPE E TO TERMINATE DISPLAY
STRUCTURE    1 CAS REGISTRY NUMBER 57-83-Ø
Pregn-4-ene-3,2Ø-dione (9CI)
Progesterone (8CI)
.delta.4-Pregnene-3,2Ø-dione
Agolutin
Bio-luton
Corlutin
Corlutina
Corluvite
Corporin
Corpus luteum hormone
Flavolutan
Fologenon
Gesterol
Gestone
Gestormone
---------
STRUCTURE    2 CAS REGISTRY NUMBER 57-87-4
Ergosta-5,7,22-trien-3-ol, (3.beta.,22E)- (9CI)
Ergosterol (8CI)
Ergosterin
Provitamin D
Provitamin D2
---------
STRUCTURE    3 CAS REGISTRY NUMBER 57-88-5
Cholest-5-en-3-ol (3.beta.)- (9CI)
Cholesterol (8CI)
(-)-Cholesterol
Cholest-5-en-3.beta.-ol
Cholesterin
Cholesterol base H
Cholesteryl alcohol
Cordulan
Dusoline
Dusoran
Dythol
Hydrocerin
Kathro
Lanol
Nimco cholesterol base H
---------
```

Figure 7. Steroids with the partial name ". . .sterol".

```
OPTION? SSHOW 5

HOW MANY STRUCTURES (E TO EXIT) ? 2

TYPE E TO TERMINATE DISPLAY
STRUCTURE    1 CAS REGISTRY NUMBER 53-1Ø-1
Pregnane-3,2Ø-dione, 21-(3-carboxy-1-oxopropoxy)-, sodium salt, (5.bet
    a.)- (9CI)
5.beta.-Pregnane-3,2Ø-dione, 21-hydroxy-, hydrogen succinate sodium sa
    lt (8CI)
 RØØ1-626Ø
Hydroxydione hemisuccinate
Hydroxydione sodium
Hydroxydione sodium hemisuccinate
Hydroxydione sodium succinate
P 55
Presuren
Sodium 21-hydroxypregnane-3,2Ø-dione succinate
Viadril
21-Hydroxypregnane-3,2Ø-dione sodium hemisuccinate
21-Hydroxypregnane-3,2Ø-dione sodium succinate
---------
STRUCTURE    2 CAS REGISTRY NUMBER 53-41-8
Androstan-17-one, 3-hydroxy-, (3.alpha.,5.alpha.)- (9CI)
Androsterone (8CI)
Androkinine
Androtine
Atromide ICI
3.alpha.-Hydroxy-17-androstanone
3.alpha.-Hydroxyetioallocholan-17-one
3-Epihydroxyetioallocholan-17-one
5.alpha.-Androsterone
---------
HOW MANY STRUCTURES (E TO EXIT) ?
```

Figure 8. Steroids lacking the partial name ". . .sterol".

in a Boolean NOT sense with file 3, to create file 5, which contains the 50 substances which do not have the fragment STEROL in their name, although, since they were in file 3, they do have the steroid skeleton and appear in the Merck Index.

The first three substances in file 4 are shown in Figure 7, and the first two from file 5 are shown in Figure 8. File 4 does, in fact, contain substances such as gesterol, ergosterol, and cholesterol base H, in contrast to the compounds in file 5 whose names are devoid of the name fragment STEROL.

4. Identity Search. A common need of SANSS users is to find and retrieve a structure identical with the query structure. Thus, while it may be of interest to some users to invoke FPROB to find all chlorophenyl compounds, other users want simply to locate chlorobenzene itself, usually to learn its CAS Registry Number, which can be used to locate the compound in other CIS files or retrieve literature citations to the compound from *Chemical Abstracts*.

In the earlier versions of the SSS program, it was not possible to do this directly, and the indirect methods, e.g., FPROBE in conjunction with the molecular formula search, were expensive and inefficient. In view of the growing need

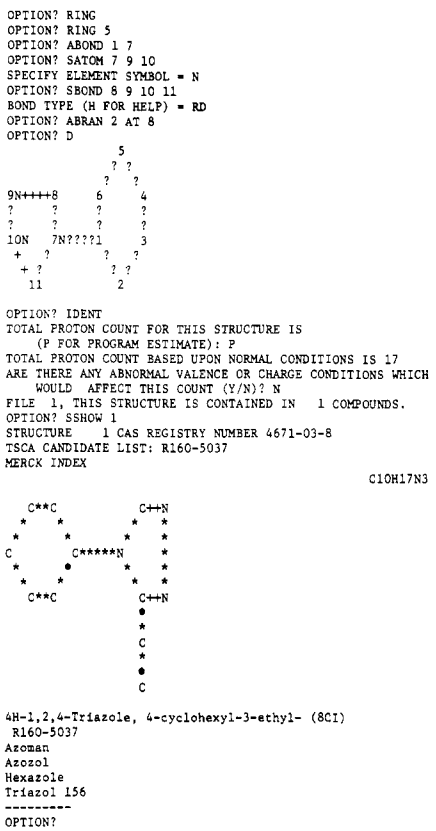

Figure 9. IDENT search for hexazole.

for such an identity match, particularly in connection with registration of chemicals under the Toxic Substances Control Act, a program of this sort has been developed and incorporated into the SANSS.

When the program is invoked, it examines the query structure and asks the user to provide the correct number of hydrogen atoms for it. This is done to resolve some otherwise ambiguous situations resulting from the treatment of tautomeric bonding conditions in the structure of the compound. At this stage, the user has two options: the number of hydrogens can be entered; or the program can be asked to calculate and report back the correct number of hydrogens corresponding to the query structure. If the latter option is exercised, the program asks the user if there is any reason (such as nonneutral atoms) to doubt the program's proton count. If not, the identity search begins.

The number of protons in the query structure is first merged into the connection table corresponding to the query structure and the resulting set of numbers is hash-encoded. This hash-encoded information is used to search through an inverted file of similarly hash-encoded entries corresponding to each structure in the database. This is a very fast look-up type of search and locates the correct compound in a very short time. As a result, the identity search is probably the fastest and most economical method for searching through the full database.

An example of an IDENTity search is shown in Figure 9. The command RING results in the generation of a six-membered ring, whose connection table is stored by the computer This is followed by the command RING 5, which adds a five-membered ring to the structure. The rings are joined with the command ABOND 1 7, which creates a bond between atom 1 (in the six-membered ring) and atom 7 (in the five-membered ring). Atoms 7, 9, and 10 are defined as nitrogen, using the SATOM command, the bonds between atoms 8 and 9 and 10 and 11 are defined as ring-double bonds, and finally, a two-atom branch at atom 8 is added. All atoms except those defined as nitrogen are assumed by the program to be carbon. The result is the query structure that is displayed in response to the command D (for display).

If the program IDENT is now invoked, it asks for a count of the number of hydrogen atoms in the structure. Rather than count them, the user responds by typing P, which causes the program to estimate the number of hydrogens at 17. This estimate is based upon the assumption that all bonds in the molecule, unless otherwise specified, are single bonds. After ascertaining that there is no reason to question this estimate, the search is accomplished and the program reports to the user that one compound has been found and has been stored in file 1. Inspection of this file, using the SSHOW option, reveals that the exact compound, hexazole, is in two files, and has the structure and names shown.

Using a little under 20K words of DEC System 10 core, an identity search requires about 2 seconds of cpu time which means that a typical commercial cost for such a search is about $1.00, plus the cost of the structure generation.

SUMMARY

The SANSS occupies a central position in the NIH–EPA Chemical Information System, and consequently, its continuing development is in two areas. First, the SANSS Unified Database and software must be continually updated and maintained and, second, a variety of links between SANSS and other CIS components must be installed.

The design of the SANSS described here is felt to be generally satisfactory and the mechanisms for file-updating are now working well. As a result, this aspect of the system is approaching full development and now requires mainly a maintenance effort.

Links between SANSS and other CIS components have not yet been completed, however, and future work will focus upon this problem. Currently, one can learn that a particular compound is represented in the CIS Mass Spectral Search System, but retrieval of the mass spectrum can still be done only indirectly, using the CAS Registry Number. Direct retrieval of numeric data from within the SANSS is now possible with programs that are under test, and it is hoped that such capabilities can be made more generally available during the coming year.

REFERENCES AND NOTES

(1) R. J. Feldmann, G. W. A. Milne, S. R. Heller, A. Fein, J. A. Miller, and B. Koch, *J. Chem. Inf. Comput. Sci.*, **17**, 157 (1977).
(2) S. R. Heller, G. W. A. Milne, and R. J. Feldmann, *Science*, **195**, 253 (1977).
(3) R. J. Feldmann and S. R. Heller, *J. Chem. Doc.*, **12**, 48 (1972).
(4) P. D. Dittmar, R. E. Stobaugh, and C. E. Watson, *J. Chem. Inf. Comput. Sci.*, **16**, 111 (1976).
(5) G. W. A. Milne and S. R. Heller, *J. Chem. Inf. Comput. Sci.*, **16**, 232 (1976).
(6) M. Bracken, J. Dorigan, J. Hushon, and J. Overbey, Chemical Substances Information Network, MITRE Technical Report, MTR-7558 (prepared under Contract CEQ7A010 for the Council on Environmental Quality), June 1977.
(7) Those wishing copies of any of these files should contact H. J. Bernstein, CIS Project, Brookhaven National Laboratory, Upton, N.Y. 11973; telephone (516) 345-4379 or FTS: 666-4379 or 800-645-1132.
(8) S. R. Heller, *Anal. Chem.*, **44**, 1951 (1972).