An Application of Interactive Computing—A Chemical Information System

RICHARD J. FELDMANN, STEPHEN R. HELLER,* and KENNETH P. SHAPIRO
Computer Center Branch and Heuristics Laboratory, Division of Computer Research and Technology, National Institutes of Health, Department of Health, Education and Welfare, Bethesda, Md. 20014
RACHELLE S. HELLER
Department of Computer Science, University of Maryland, College Park, Md. 20742

Received October 5, 1971

A collection of computer programs to do chemical information retrieval are described in terms of the interactions between the computer and the chemist. The medium of interaction is the language of the chemist—the structural diagram. In interacting with the computer, the chemist can graphically specify two-dimensional structures as queries and view structures as search results.

To interest and enable the chemist to use computers, collection of programs are being developed which have two main features. They are highly interactive/conversational programs, and when feasible, they are graphical.

The collection of programs are being built into a Chemical Information System (CIS). The goal of the Chemical Information System is to supplement the sources that a chemist uses when searching for information,[1,2] which may be either references or specific chemical or physical data and properties. One feature distinct to a chemist's question is the use of the structural diagram. Thus, our initial assumption was that a useful system must be highly graphical in nature.

It is better that the computer come to the user rather than the user go to the computer. With a terminal located in the user's office or building, the system becomes a routine tool. These assumptions led to our desire to experiment with a dial-up chemical information system.

Figure 1 shows the general outline of the experimental system, divided into three main project components.

Each component in the CIS is, in reality, one or more FORTRAN and Machine Language programs. The user of the system exercises only the retrieval program, but there are several support programs for file generation and file maintenance forming the system infrastructure.

One can enter the CIS through the individual programs or the system administrator program which gives an overview of the system. Figure 2 shows the system flag and tells the user what to do next. The display always identifies the program in execution and the action which will cause the program to go to its next state.

When the user decides which of the components of the CIS he wishes to enter, he simply depresses the tablet stylus over the line naming the desired component.

* Author to whom correspondence should be addressed.

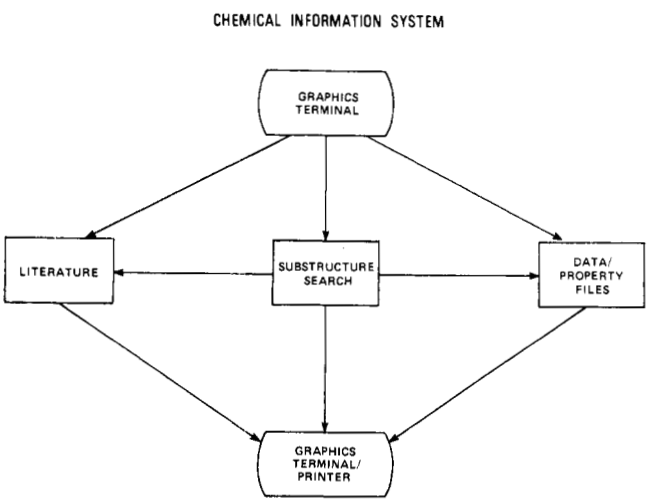
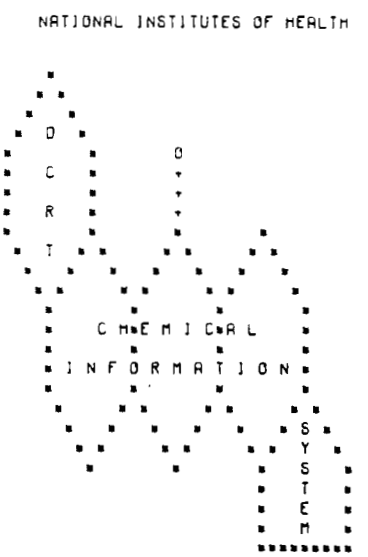


Figure 1. Outline of the DCRT/CIS

SUBSTRUCTURE SEARCH COMPONENT

Substructure searching (SSS) is a method of finding chemical groups, fragments, or substructures imbedded in a chemical structure. Chemical structures can be represented in many ways. We have used both Wiswesser Line Notation (WLN) and the two-dimensional notation of the Chemical Abstracts Service, the CAS connection tables.

Computer programs have been written which allow a user to draw a chemical structure (for an appropriately large class of chemical compounds) and obtain the WIN.[3-5] In addition, he can also search for a substructure[6] by either



SYSTEM ADMINISTRATOR

DEPRESS STYLUS TO PROCEED

Figure 2. System administrator program and system flag

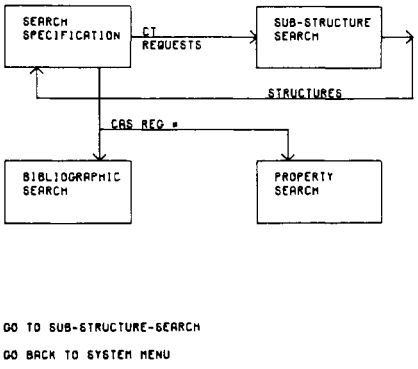


Figure 3. Substructure search (SSS) component

drawing the chemical fragment or inputting a WLN string.[7] The main advantage of WLN is that it has mnemonic content and can be used without a computer. However, the need to learn another nomenclature does not make it attractive for a user-oriented, computer-based system.

We are also experimenting with the Chemical Abstracts Service connection tables. Associated with the CAS connection tables is the REGN, the CAS Registry Number. It is a unique, 9-digit, clerically assigned number. Every compound that has gone in the printed version of CA since 1965 has been put into the file, which is currently about 1.8 million compounds and growing by about 250,000 per year.[8]

Although the connection tables have no associated meaning or semantic content, they are well suited for internal computer manipulation, with which the user need not be concerned. In addition, although having slightly larger storage requirements, the connection tables have the capability of retrieving a significantly larger class of chemical structures owing to the added information in the atom-by-atom connection table, as opposed to a notation such as WLN.

The main area of our current research in structure searching involves the use of the Chemical Abstracts Service data files; in particular, a subset of the 1.8 million chemical structure that *Chemical Abstracts* has registered, called the Common Data Base. This is a 20,000 compound chemical file produced for and being used by the Food & Drug Administration and the National Library of Medicine.

The two groups of computer programs of concern in this paper are the SSS and the Rapid programs. Both programs use the CAS Common Data Base.

The SSS programs allow the user to sit at a graphics terminal, draw chemical fragments, impose a variety of atom and bond restrictions on the structures (e.g., the ring must be aromatic, there must be a halogen beta to the nitrogen, etc.), and interactively search a file in a few minutes to find those structures which contain the fragment imbedded in them.

The search is two part, the first being a screen to reject the majority of compounds in the file that do not meet even such a broad restriction as molecular weight, number of carbon atoms, etc. The second part, called the iterative search phase, involves an exact atom-by-atom, bond-by-bond check of the user's request *vs.* the structures in the file.

By using a PDP-10 time-sharing computer on a small file of 5,000–10,000 compounds, the user can formulate the question in his natural language—the chemical structure—and within a few minutes, see the results of his query. Since this is an interactive program, good queries can be readily answered, and bad queries readily reformulated. The result of the search is the query fragment structure along with the fragment shown imbedded in the full chemical structure, with its CAS Registry Number (REGN).

The use of the SSS component, which consists of two programs, is typically a three step process:

1. Search specification
2. Sequential search through the structure file
3. Analysis of structures found in the search

The SSS cycle permits the user to specify structures and structural properties, to do a structure search rapidly, and to analyze the results. The program can be terminated whenever the user sees that his questions are either producing too few structures (too restrictive a specification) or too many structures (too broad a specification). The result of the cycle is a subfile of structures and the corresponding CAS REGN which in the future can be used to enter the bibliographic search component or the property search component (Figure 3).

The SSS program requires the CAS connection tables as well as the search screens. The search screens rapidly eliminate structures which do not meet the specifications of the search queries.

Interactive Specification of SSS Queries. The interactive specification of SSS queries is based on a number of structure tranformations. The transformations are activated by the following actions with the Rand Tablet stylus (Figure 4):

1. Depressing the stylus on one spot
2. Drawing a line
3. Drawing an arc
4. Drawing a closed loop
5. Depressing the stylus over a key on the right side of the screen (shown in Figure 5)

Bond Specification. The bond specification process has three steps. This process must be repeated for each bond which is to be specified.

1. Depress the stylus over the "BND CLASS"
2. Depress the stylus on two sites for which the bond specification is to be made
3. A list of bond specification types then appears on the screen. Depress the stylus over the desired bond type

The user is prevented from making a bond specification for two sites which are not connected. Bond specification types are shown in Figure 6.

Atom Specification. When a structure is drawn as part of a query specification the atom sites are denoted by a "*" which means that the search will accept any atom type. An atom site is made specific by

1. Depressing the tablet stylus over the atom keys (see Figure 5)
2. Depressing the stylus over the atom site

Infrequently used atom types can be selected by depressing the stylus over the "X" key. This causes all 103 atom types to appear on the screen. The selection is made by depressing the stylus over the desired atom type. If the stylus is then depressed over the structure site, the specification is accomplished.

By repeating the site specification process for different atom types a structure site is given a class specification. The atom class specification is initially set to be an inclusion class. When the stylus is depressed over the "AT CLASS" key, the atom specification class is toggled from inclusion ("IN") to exclusion ("EX"). An atom exclusion class prevents the acceptance of its member atom types during the SSS. Specification of an exclusion class at a site which already has an inclusion class specification causes both classes to be eliminated. Either an inclusion or exclusion class can be deleted at a site by depressing the stylus over the "*" key and then over the site.

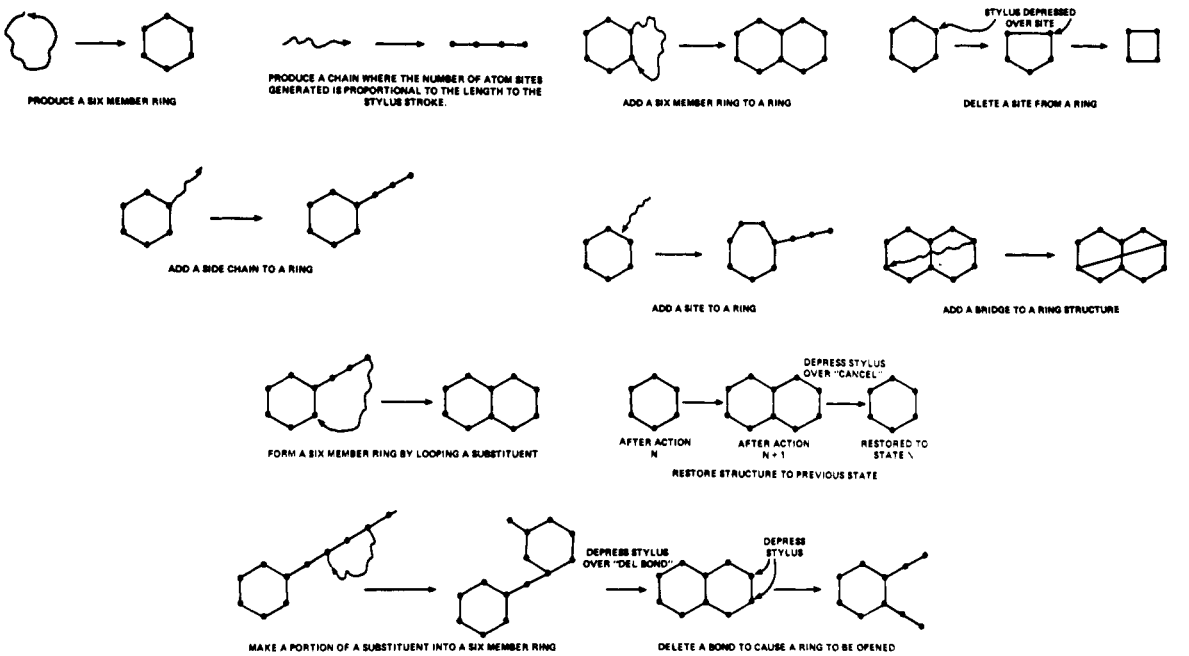


Figure 4. Possible structure transformations

To finish a query specification the stylus is depressed over the "END SPEC" key. The user is then asked if the atoms which are still unspecified ("*") are to be searched as carbon (C) or as any atom type.

Search queries are transmitted between the specification program and the search program by means of an intermediate file. Figure 7 shows what the computer-produced coded questions to four search queries for nitrogen mustards look like. Each line of the connection table is an atom number, atom type, and 1 to 4 neighbors. A "+" in the atom type column indicates an inclusion class. The members of the inclusion are on the line following the "+". Query 1 is the most specific and produces the fewest hits, while query 4 is the least specific and produces a greater number of hits for a given file.

The SSS Program sequentially searches a file composed of search screens and structural data. The over-all characteristics of the search screen are given on the right side of Figure 8. A search screen is a mechanism for rapidly eliminating from consideration those structures which could not possibly answer a structure query.

The screen phase generates a screen for each query. Figure 9 is a typical search screen. The Query search screen has the same form as the search screens of each structure in the file being examined. If a query asks for three rings and a given structure has only two rings, this structure cannot possibly satisfy the screen query, so it can be eliminated.

The hit summary on the left side of Figure 8 gives the user a dynamic indication of the state of the SSS program.

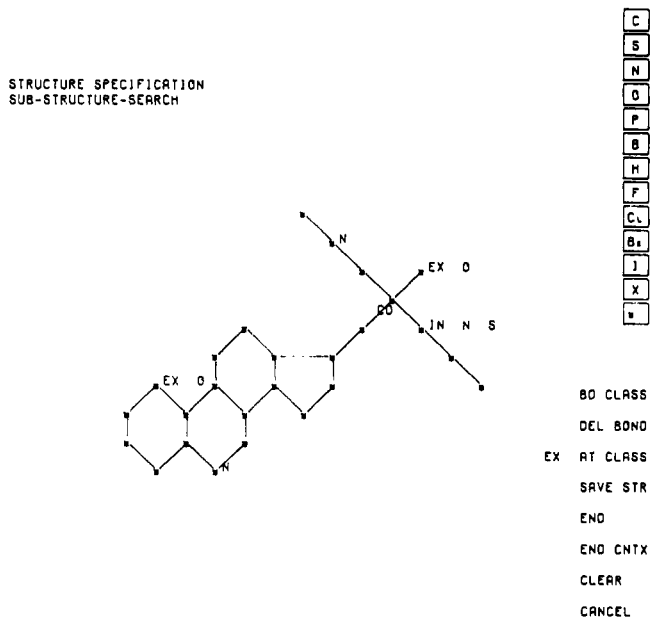


Figure 5. Display screen with function keys

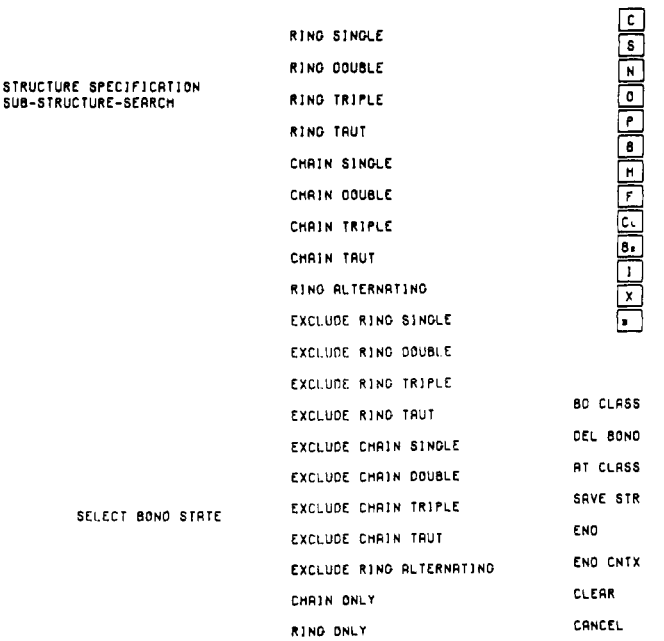


Figure 6. Bond specification

```
 1 N          2
 2 C          1  3
 3 C          2  4
 4 CL         3
   **

 1 N          2
 2 *          1  3
 3 *          2  4
 4 CL         3
   **

 1 N          2
 2 C          1  3
 3 C          2  4
 4 +          3
CLBR
   **

 1 N          2
 2 *          1  3
 3 *          2  4
 4 +          3
CLBR
   **
```

Figure 7. Four coded queries for nitrogen mustards

```
HIT SUMMARY

NUMBER OF STRUCTURES SEARCHED
   50
QUES   SCREEN   ITERATIVE
HITS    HITS
  1     15        11
  2      0         0
  3     18 *      10 *
```

```
COMPONENTS OF THE
SUB-STRUCTURE-SEARCH PROGRAM

1. SCREEN PHASE

NUMBER OF ATOMS
TOTAL WEIGHT
MOLECULAR FORMULA
BOND COUNTS
TOTAL NUMBER OF RINGS
RING COUNTS BY SIZE
NUMBER OF AUGMENTED ATOMS
   AT CONNECTIVITY 1,2,3,4
AUGMENTED ATOMS AT CONNECTIVITY 4
AUGMENTED ATOMS AT CONNECTIVITY 3
AUGMENTED ATOMS AT CONNECTIVITY 2
AUGMENTED ATOMS AT CONNECTIVITY 1
   AUGMENTED ATOM IS
   NUMBER OF OCCURRENCES
   TYPE OF CENTRAL ATOM
   TYPE OF NEIGHBORING ATOMS
   BOND TYPE OF NEIGHBORING ATOMS


2. ITERATIVE PHASE

THIS IS AN ATOM BY ATOM, BOND BY BOND
COMPARISON OF THE TWO STRUCTURES
```

```
SUB-STRUCTURE-SEARCH        TYPE S TO SUPPRESS SCREEN AND ITERATIVE
                            TYPE I TO SUPPRESS ITERATIVE
                            *   PHASE IS SUPPRESSED
```

Figure 8. The search

```
BOND
 0 35 35 35
ABCH
 0
SAF
70  1  1  0  0  0  1  1
MAF
 0
HYDC
17 16
GRAPH
 0  4  1  1  1
NODE
60142142142
REGNO
 0  01145 141
TAUT
 4  2  4  3
TEXT
213226
 ABVAL
 0

REGISTRY NUMBER       50011   TEXT NS
   1    15      0   0   2   3   4   0   0   0
   2    56      0   0   1   0   0   0   0   0
   3    56      0   0   1   0   0   0   0   0
   4    56      0   0   1   0   0   0   0   0
   5    20      0   0   0   0   0   0   0   0

STRUCTURE NUMBER        2

NUMBER OF ATOMS    5 TOTAL WEIGHT    89
ELEMENT COUNTS AS ATOM TYPE NUMBER, NUMBER OF OCCURRENCES
  15   1
  20   1
  56   3
COUNT
CSC      0
CDC      0
CTC      0
CTAUTC      3
RSC      0
RDC      0
RTC      0
RTAUTC      0
RALTC      0
RNGC      0

RING SIZE, RING COUNT
NUMBER OF AUGMENTED ATOMS AT CONNECTIVITY   1,2,3,4
 2  1  1  0

AUGMENTED ATOMS AT CONNECTIVITY   3
    1   15   56   56   56
                 8    8    8
AUGMENTED ATOMS AT CONNECTIVITY   2
    1   15   56   56
                 8    8
AUGMENTED ATOMS AT CONNECTIVITY   1
    1   15   56
                 8
    3   56   15
                 8
LENGTH OF GENERATED SCREEN     57

GENERATED SCREEN

  16   0   0  57   5   0  89   7  15   1
  20   1  56   3   0   0   0   3   0   0
   0   0   0   0   1   2   1   1   0  10
   3   1  15  56  56  56   8   8   8   8
   2   1  15  56  56   8   8  10   1   1
  15  56   8   3  56  15   8
```

Figure 9. Sample search screen for structure from common data base

The screen hit count gives the number of structures which have passed the screen phase. The interactive hit count gives the number of structures which have passed the screen and iterative phases. The iterative phase is an atom-by-atom, bond-by-bond comparison of the query structure and a structure from the file being searched. All possible matchings are tried. To evaluate all combinations so as to arrive at a positive or negative decision as rapidly as possible, the connection table is ordered so that low occurrence atoms are matched first. Then oxygen atoms which have a fair number of occurrences are matched. Next, nitrogen atoms which occur more frequently than oxygen atoms are matched. Finally carbon atoms, which compose the bulk of organic structures, are matched. Only when every other atom type has been matched are the carbon atoms matched. This points to the types of queries which make most efficient searches.

The user can terminate the SSS program at any time if the search is producing too many hits. If the user types "S" (for suppress) followed by the query number, then the query will have both screen and iterative phases suppressed. If the user types "I" followed by the query number, the iterative phase of that query will be suppressed.

The rate at which the SSS program operates is determined by a number of factors. The most important factor affecting real time response rate is the loading of the time sharing PDP-10. At peak loading the elapsed time required to search a file can be degenerated by a factor of four. The specificity of the search question is the next most important factor governing compute and real time response. The search screens perform well only when there are noncarbon atoms and/or specific atom contexts.

Search Screens. Screens eliminate from consideration structures which do not have the properties required by the query. Existing screens for connection table files—e.g., those of the CAS[9] and Walter Reed Systems[10] have the following characteristics:

1. Fixed screen length regardless of the size or complexity of the structure

2. Fixed usage of each bit in the screen

The fixed screen size means that small or simple structures do not completely utilize the screen. Fixed usage of each screen bit means that screen bits must be set aside for infrequently used fragments. In the CAS implementation, 2400 bits are used for each screen. One summary[11] indicated that, on the average, in a file of 150,000 structures, between 110 and 160 screen bits were set, indicating that the greater portion of the screen bits are underutilized. This results in a very large file of structures and screens.

In implementing a screen for the SSS component, an effort has been made to make the size of the screen and the resolving capability of the screen a function of the size and complexity of the structure. The DCRT screens have the following characteristics:

1. Total screen length is variable
2. Individual components of the screen are variable length
3. Time required to compute the screen depends on the size and complexity of the structure

The variable size of the screen is due to variability of the molecular formula, ring count by size, and the augmented atoms. When it reaches the end of the file being searched or the user types "E", the SSS program is overlaid by the structure specification and analysis program. The user then answers "D" to the query in Figure 7 to display the results of the SSS. The atoms of the query are numbered, and the atoms of the satisfying structure are correspondingly numbered (see Figures 10 and 11).

The user can readily determine if his query selected the structures intended. The user can quit the analysis step at any time and refine the query, as he thinks best. The cyclic process of structure specification, SSS, and structure retrieval analysis is very rapid. Instead of the long cycle time using the batch SSS, the user can achieve a few minutes cycle for small files.

More than simply cycle time there is now no need for coding forms (or any intermediaries to code questions, keypunch questions, verify keypunching, etc.). The interactive display and tablet becomes the coding form. In addition, SSS answers are displayed as two dimensional structures.

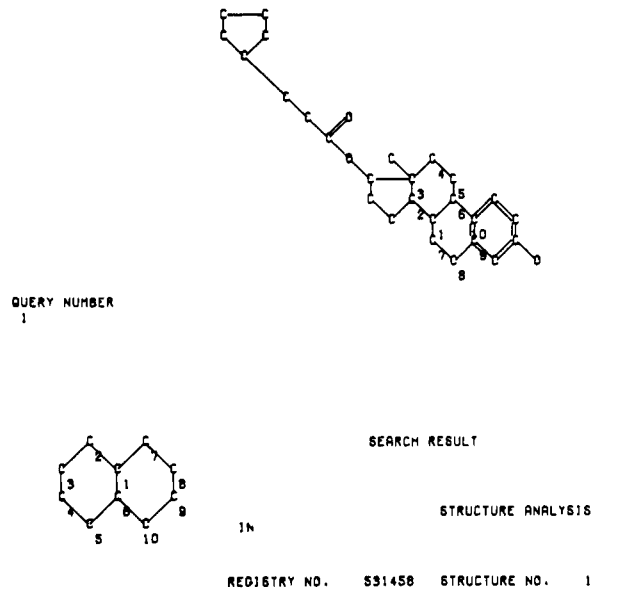


Figure 10. Search query (lower left) and search result (upper right) Example 1

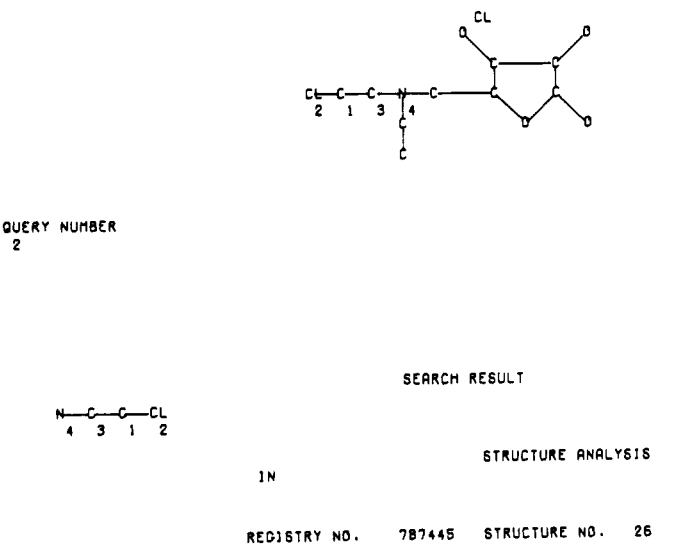


Figure 11. Search query (lower left) and search result (upper right)

LITERATURE SEARCH COMPONENT

The Literature Search component is currently focused on on the CAS CBAC journal. As a prototype of the format in which *Chemical Abstracts* will be published in the future, the experience gained with CBAC will help us to build an integrated bibliographic search and retrieval capability. The CAS registry number which appears in CBAC can be used to link to the structure and data files. CBAC search testing is being undertaken on both the PDP-10 and IBM-370 computers at NIH. A set of programs implemented on the IBM-370 provides current awareness (SDI) capability. Retrospective CBAC searching is accomplished by applying the Hash retrieval techniques. Retrieval of single terms and the associated references (volume and abstract number) take from 1 to 3 seconds. A term can be any compound name, author name, coden, or registry number. Details of the Literature Search Project can be found elsewhere.[12]

The REGN is really to be used as a means to an end or as an internal pointer within the computer to information being sought. A significant change from the historical method of bibliographic/literature search was introduced when the REGN for a chemical was included in the literature. The abstract journal CBAC (*Chemical Biological Activities*), since its inception in 1965, has had the REGN in the journal, both the printed form and the computer readable form. In addition, the Subject Index to *Chemical Abstracts* has the REGN included starting with Vol. 71 (July–Dec. 1969). One application of SSS which would be useful to explore is structure searching of CBAC and CA.

PROPERTY SEARCH COMPONENT

The Property Search Component shown in Figure 12 is a third alternate for entering the chemical information system. The heart of this component is made of files which contain property data and REGN. The property files are to be entered in two modes—i.e., given a REGN, retrieve the

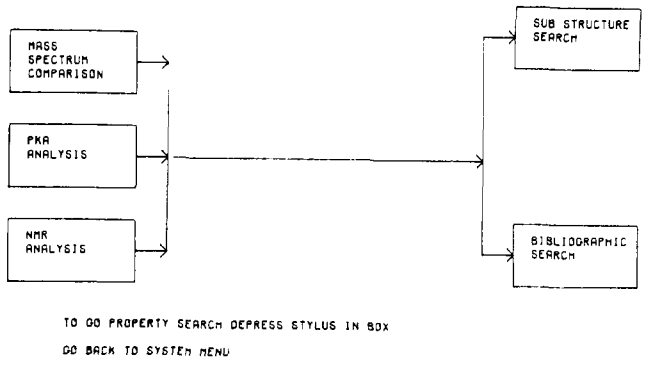


Figure 12. Property search component

associated data element or given a collection of data, search through the appropriate file(s) to find all of the matching or best matching elements and then retrieve the REGN. When the REGN are retrieved, the remaining properties can be retrieved as well as the structure. At present there are no files with REGN.

We are currently investigating the process associating the CAS registry number with existing property data. Further into the future, one might look forward to the automatic capturing of data (machine to machine) as well as the automatic association of REGN with the data. This would be accomplished by having the chemist at his remote graphics terminal specify the structure of the chemical whose data are to be entered. The rapid search and retrieval component could then find the REGN.

Four interactive, conversational search programs for property data components have been developed. Details for the NMR,[13] Mass Spec,[14] pKa,[15] and X-Ray[16] programs can be found elsewhere. For X-ray coordinate data, the current computer programs using the same computer graphics terminal as is used for the SSS are able to retrieve structures, provide for molecular rotation, bond rotation, bond distance and bond angle measurements, clipping, connection table generation/alteration, and plotting. An example of the results of a manipulation of the Ribonuclease-S X-ray data[17] is shown in Figure 13. The interest in X-ray data stems from the fact that chemical structures are in fact three dimensional objects, and their properties are very much a function of their three-dimensional and not their two-dimensional topology. It is evident that the goal of structure searching in the future will be in the area of three dimensional conformational searching.

THE RAPID STRUCTURE RETRIEVAL PROGRAM

In addition to the sequential SSS component, another program is available for the retrieval of a single COMPLETE structure. The SSS component and the rapid search and retrieval components represent the two extremes of completeness *vs.* response trade-off. The SSS produces all structures which satisfy a general query. As presently implemented, a sequential search requires a complete response and requires examination of all the structures in the file. In small files (<10,000 structures), the user can wait for the search to be done. For large files (for example 1.8 million structures), a batch mode computation is necessary because the time required for the search can be hours. Often the only response required is whether or not the query structure is in the file. In a general situation, a user may want to know if a particular compound is in his private file. For this class of queries, it seemed hardly necessary to

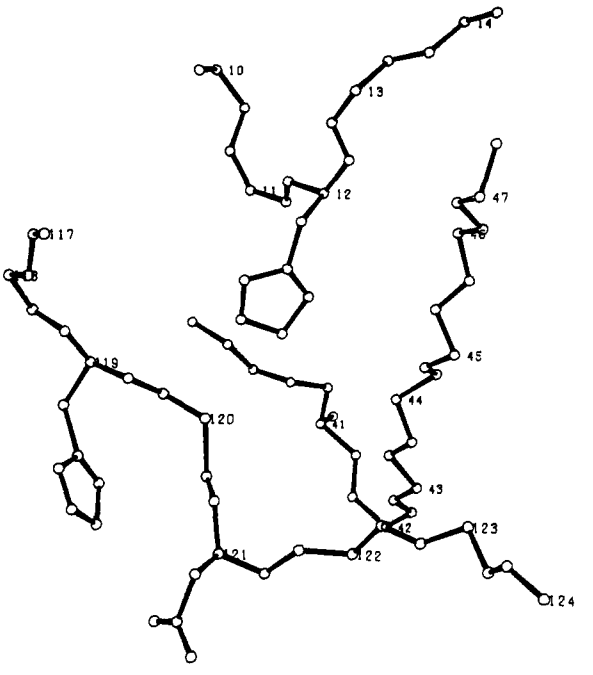


Figure 13. Partial three-dimensional structure of RNase-S with backbone, including HIS 12, 119, LYS 41, and ASP 121

search sequentially through the entire file to detect the presence or absence of a structure. To accomplish the identification of a structure, the methods of scatter storage commonly known as Hash Coding[18] were investigated. The rapid program permits a structure to be identified in seconds. Similarly, if the REGN is known, its structure (an any associated data) can be rapidly retrieved. For a file of 20,000, the response time is about one second. For 1.8 million, the response will go up to a few seconds. The rapid program accesses only one structure or one registry number at a time. Class accessibility has been traded-off for instantaneous response.

The rapid search and retrieval program permits the user to identify or retrieve a structure in less than one second. As applied to the problem of structure identification, a HASH function compresses the representation of the connection table of a query structure into one 36 bit HASH key. The HASH key is used as the starting address for a local search in the HASH file. It is neither necessary nor important that the HASH function give a unique mapping between the connection table representation which is unique and the 36 bit HASH key. The local search in the HASH file looks for the same HASH key. When an equivalent HASH key is found, the associated pointer is used to locate the structure data.

An exact match—the iterative part of the SSS—is then performed to guarantee that the retrieved structure and the query are identical. If it happens that the HASH function has produced a duplicate then the exact match will reject all the duplicates. Thus by performing a simple transformation (the HASH function), the two disk accesses, a local search, and an exact match and identification can be made. The process takes seconds in comparison to hours for a sequential exact match (Figure 14).

Similarly the HASH technique can be used to retrieve the structure associated with a CAS registry number. The registry HASH file is generated with the structure HASH file when the screen file is generated. If other data are associated with the structure, it can be retrieved after either registry or structure identification.

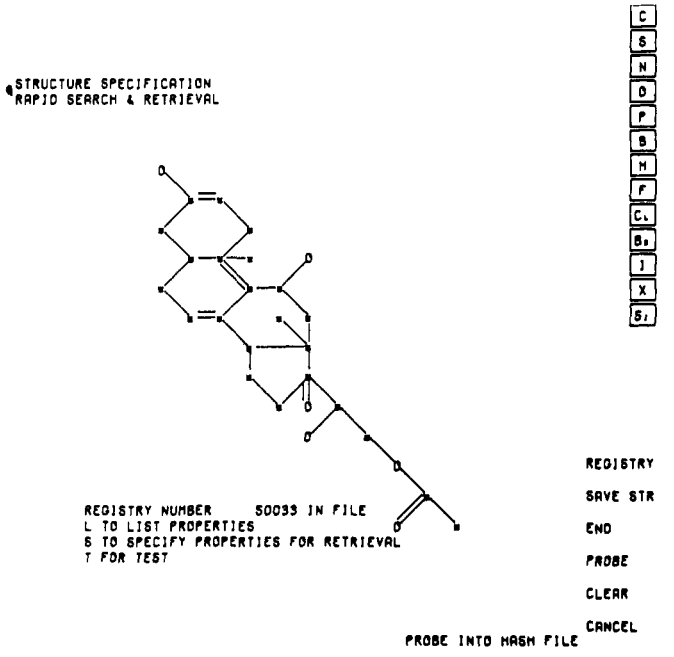


Figure 14. A structure retrieved by inputting the CAS registry number 50033

CONCLUSION

The CIS is in the first stages of experimentation. As the various components are taken from the development stage to the test stage, user reaction will induce changes in the system. The use of graphics for input and output of structural diagrams lays the foundation for interacting with the user. Whereas the rapid search is effective for single queries, the sequential SSS is too inefficient and time consuming to be of practical value. Current plans are to investigate new techniques for quickly accessing large structure files as well is to consider possible methods of three dimensional conformational searching. The REGN will provide the link between files.

LITERATURE CITED

(1) Feldmann, R. J., Heller, S. R., Shapiro, K. P., and Heller, R. S., Abstracts of Papers, "A Dial-Up Graphics Chemical Information System," CHLT 21, 161st Meeting, ACS, Los Angeles, Calif., March 1971.

(2) Heller, S. R., and Feldmann, R. J., "An Application of Interactive Computing—A Chemical Information System," First Workshop of International, Regional and Collaborative Reference Centers WHO-NIMH for Information on Psychotropic Drugs, Plitvice Lakes, Yugoslavia, June 1971.

(3) Farrell, C. D., Chauvenet, A. R., and Koniver, D. A., "Computer Generation of Wiswesser Line Notation," *J. Chem. Doc.* **11**, 52 (1971).

(4) Feldmann, R. J., Chauvenet, A. R., Koniver, D. A., Abstracts of Papers, "Interactive Searching of Chemical Files and Structural Diagram Generation from Wiswesser Line Notation," Abstracts of Papers, CHLT 17, 160th Meeting, ACS, Chicago, Ill., September 1970.

(5) Heller, S. R., and Koniver, D. A., "Computer Generation of WLN. II. Polyfused, Perifused, and Chained Ring Systems," *J. Chem. Doc.* **12**, 55 (1972).

(6) Feldmann, R. J., and Koniver, D. A., "Interactive Searching of Chemical Files and Structural Diagram Generation from Wiswesser Line Notation," *J. Chem. Doc.* **11**, 154 (1971).

(7) Miller, G., "Encoding and Decoding WLN," *J. Chem. Doc.* **12**, 60 (1972).

(8) O'Dette, R. A., Chemical Abstracts Service, Ohio State University, Columbus Ohio, personal communication, 1970.

(9) Wiggington, R. J., "Machine Methods for Accessing Chemical Abstracts Service Information," Proceedings IBM Scientific Computing Symposium on Computers in Chemistry, Yorktown Heights, N. Y., October 1968.

(10) Jacobus, D. P., Davidson, D. E., Feldman, A. P., and Schafer, J. A., "Experience with the Mechanized Chemical and Biological Information Retrieval System," *J. Chem. Doc.* **10**, 135 (1970).

(11) Hazard, M., NCI, NIH, HEW, Bethesda, Md., personal communication, 1971.

(12) Heller, S. R., Shapiro, K. P., and Feldmann, R. J., "An Experimental Computerized CBAC Search Project," *J. Chem. Doc.* **11**, 248 (1971).

(13) Heller, S. R., and Feldmann, R. J., "An Interactive NMR Chemical Shift Search Program," *J. Chem. Ed.* **48**, in press (1972).

(14) Heller, S. R., "An Interactive Mass Spectral Search System," NIH, HEW, Bethesda, Md., unpublished data, 1971.

(15) Heller, S. R., NIH, HEW, Bethesda, Md., unpublished results, 1971.

(16) Heller, S. R., Feldmann, R. J., Bacon, C. T., NIH, HEW, Bethesda, Md., unpublished results, 1971.

(17) Wyckoff, H. W., Tsernoglou, D., Hansen, A. W., Knox, J. R., Lee, B., and Richards, F. M., "The Three-Dimensional Structure of Ribonuclease-S," *J. Biol. Chem.* **245**, 305 (1970).

(18) Morris, R., "Scatter Storage Techniques" *CACM*, **11**, 38 (1968).