21 January 1977, Volume 195, Number 4275 SCIENCE

A Computer-Based Chemical Information System

Chemical data stored in a central computer can be used internationally in real time and at low cost.

S. R. Heller, G. W. A. Milne, R. J. Feldmann

A considerable improvement in the power of computers in the 1960's prompted work at the National Institutes of Health (NIH) and at the headquarters of the Environmental Protection Agency (EPA) to explore the feasibility of an on-line chemical information system (*1*). Several capabilities were identified as necessary in such a system, including storage capacity, search and retrieval from libraries of chemical data, computer-assisted analysis of chemical data, and retrieval of information from the chemical literature. The types of numerical data under consideration included mass spectra, carbon-13 nuclear magnetic resonance (NMR) spectra, and x-ray diffraction data. Bibliographic data dealing with mass spectrometry and x-ray crystallography have also been incorporated into the system.

In view of the well-known difficulties implicit in the use of batch-oriented computers in information retrieval, it was considered essential that an on-line interactive computer be used for these tasks and so the NIH PDP10 was chosen as the computer to be used in these experiments. A chemist who searches through a data base often possesses a good deal of information, perhaps poorly defined, that is related to the problem at hand. A well-programmed computer can interrogate him as he is interrogating the data base and so prompt him to recall as much of this additional information as is feasible to solve the problem at hand. In such an interactive system, the computer can report to the user that "17,450 citations satisfy the criteria" that were specified and ask "Do you want to alter the criteria?" If so, the computer can interrogate the user about the new criteria. With each modification in criteria, the user can see how the number of hits changes; in this way the question can be "tuned" until it gives a manageable number of answers, which can then be either sampled by the user or printed in toto.

To build such a chemical information system one must first locate or generate a data base that can be used in each component. Then programs for interactive searching of the data bases must be written that will permit the direct interrogation of a disk-stored data base so that information may be retrieved at low cost. The structuring of the files on the disk and the exact details of the disk input-output statements are of prime importance in this connection. Next, a means must be devised by which the components can be linked. A user, having identified several compounds from their mass spectra, may next wish to examine their NMR spectra, and some means should be available by which he can readily do so. Finally, a mechanism must be found by which a working program–data base combination can be disseminated to as large a group of users as possible. That the system attract many users is important, because costs to the individual user can be minimized in this way and also because the users are a major source of new data for the system.

The NIH-EPA Chemical Information System (CIS), built on the general principles described above, consists of a series of PDP10-resident numerical and bibliographic data bases together with a battery of interactive, conversational computer programs that can be used to search for and retrieve information from any of the data bases. In addition, there are interactive programs that will permit analysis of data, either to reduce them to a form in which they can be used in the searching programs or as an end in itself. New components or updates are merged into the CIS by the mechanism shown schematically in Fig. 1.

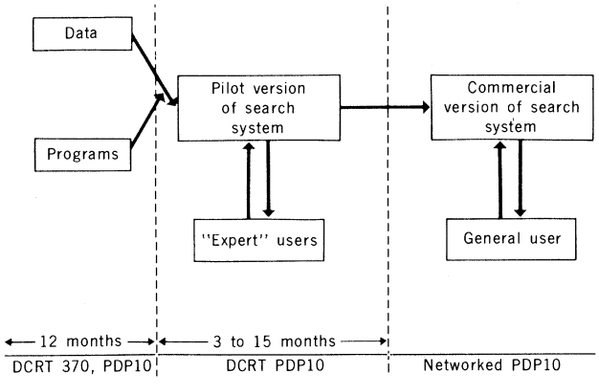
The data bases which comprise the CIS include mass spectra, carbon-13 NMR spectra, x-ray diffraction data for organic and inorganic molecules, and x-ray powder diffraction patterns. All of these files can be searched structurally; that is, they can be examined for the presence of a given chemical structure or substructure. The files are linked by means of Chemical Abstracts Service (CAS) registry numbers, which are unique chemical identifiers. There are also three bibliographic files which contain the literature citations on mass spectrometry, x-ray diffraction of organic molecules, and gas-phase proton affinities.

The analytical programs that are available can accomplish the iterative analysis of complex NMR spectra, general curve-fitting, and linear regression analysis. Other programs can be used to calculate isotopic enrichment from mass spectral data or, for a given species in solution, to find the molecular conformation with the lowest energy.

Computer Networks

In a computer network (*2*), the data bases and all the associated programs are stored in a central computer. This computer center has its own dedicated communications system which permits worldwide, 24-hour access by local telephone call by telephone landlines or sat-

Dr. Heller is a computer specialist in the Management and Information Data Systems Division, Environmental Protection Agency, Washington, D.C. 20460. Dr. Milne is a research chemist in the National Heart, Lung, and Blood Institute, Bethesda, Maryland 20014. Mr. Feldmann is a computer specialist in the Division of Computer Research and Technology, National Institutes of Health, Bethesda, Maryland 20014.



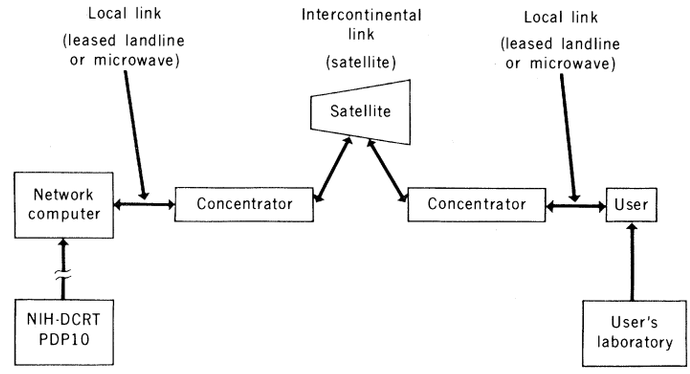


Fig. 1 (left). Addition of a component to the Chemical Information System; DCRT, Division of Computer Research and Technology, National Institutes of Health. Fig. 2 (right). Schematic diagram of a computer network.

ellite links, as shown schematically in Fig. 2. A user in Basel can, using a telephone-coupled terminal, obtain access to the programs and data base on the central computer. Communications are essentially instantaneous; the normal pauses associated with time-sharing are generally more noticeable than the signal transmission times (*3*).

The overall cost of communications is divided among all users on a per hour basis, and the usual cost per user is now about $10 to $15 per hour of connect time. Irrespective of one's location, the cost of contacting the CIS, which may be many thousands of miles away, is nearly negligible, and the speed of computer response is essentially instantaneous.

The only other way to avoid long-distance telephone calls is to maintain several copies of the data bases in different locations. However, this arrangement incurs higher costs for computer storage and also presents formidable problems in connection with the updating of data bases. When a network is available, the data need be stored on a disk only once and the whole system can be updated by simply replacing the old data base, programs, or both, with new ones. Global networking has the further advantage that all communication with users of a particular program can be handled by means of system messages. It is not even necessary to know who a user is to reach him; his signing onto the system identifies him as a user and prompts the system to print out any messages of importance to such individuals.

A disadvantage of networks is that the techniques they employ run counter to the telecommunications policies of a number of governments. The difficulties range from the relatively simple problem that the use of acoustic couplers is not allowed in some countries to the less manageable case of a government demanding 51 percent local ownership of any network nodes in the country. Another problem is that a network involves considerable overhead costs. To ensure that these costs will be covered, some commitment such as an annual subscription fee must be required. This is reasonable for those who plan moderate or heavy use of the system, but it discourages infrequent usage.

Mass Spectral Search System (MSSS)

This is the oldest component of the CIS and has been developed (*4*) from programs written in 1971 by Heller (*5*). The data in this component are unit-resolution electron ionization mass spectra collected over about a decade by a number of groups, such as the Mass Spectrometry Data Centre (MSDC) in England and the NIH and EPA in the United States. The complete MSDC-NIH-EPA collection currently contains some 30,000 mass spectra of different organic compounds, each of which has a CAS registry number, a Wiswesser line notation (WLN) (*6*), and a connection table, which is a mathematical representation of the two-dimensional structure of the molecule.This file is available for lease from the National Bureau of Standards (*7*).

The MSSS is being operated by MSDC via the Cyphernetics Division of ADP Network Services, Inc. (*8*), and to use it one must pay an annual subscription fee of $300 per institution. The user is also subject to charges for connect time and computation.

The complete mass spectra are stored on the computer (*9*), but all the actual searching is carried out on an inverted file of abbreviated spectra; the abbreviated spectra (*10*) consist of the two most intense peaks in every range of 14 mass units in each full spectrum, that is, mass-to-charge (*m/e*) ratio of 6 to 19, 20 to 33, and so on. The file of abbreviated mass spectra is used to generate an inverted file, a list of *m/e* values versus the identification numbers of spectra in which there is an ion at the particular *m/e* value. The intensity of the ion is appended to the identification number. When an *m/e* value is entered as part of a query, the computer reads out from that part of the inverted file the identification numbers of the spectra that contain such an ion, and, while doing this, it checks whether each intensity falls within the range defined by the user as acceptable. This inverted file search (*5*) is very fast, requiring only about 1 to 2 seconds of PDP10 processor time. More importantly, it is largely independent of file size.

In a typical PEAK search through the mass spectral data base (Fig. 3) the user is asked for an *m/e* value and lower and upper intensity limits on the conventional scale (0 to 100 percent). If only one intensity is provided in the response, the program works with a window of ± 30 percent about this intensity. The first entry in the example shown, *m/e* 283, intensity between 10 and 40 percent, produces 329 spectra that satisfy this criterion. The second peak entered, *m/e* 301, produces a certain number of spectra that satisfy that criterion alone but which are not reported to the user; instead, the computer seeks identification numbers that are common to this second list and the original 329 answers. In this case, 40 spectra result. A third peak of *m/e* 245 reduces this list of hits to one, at which point the user types "one" and the answer, pregn-4-ene-3,20-dione, 17-hydroxy-16,alpha.methyl-, number 23445, is printed out, together with the appropriate CAS registry number (REGN), spectrum quality index (QI), molecular weight (MW), and molecular formula (MF).

Typically, the user next retrieves the complete mass spectrum corresponding to number 23445. One may obtain a partial or complete digital listing of *m/e* val-

ues and relative intensities from a computer-generated microfiche copy of the data base or one may use the retrieval option SPEC. If the user has a vector display terminal, he may retrieve the mass spectrum in the conventional bar graph form by using the option PLOT. The mass spectrum can also be plotted and a hard copy of the bar graph may also be obtained either as a photocopy of the screen or by means of an *x-y* plotter, depending upon the equipment available.

This data base can be searched in other ways, for example, for specific molecular weights or complete or partial molecular formulas. The compilation may also be searched by means of structure codes—arbitrary numerical codes used to define a particular structural feature or compound type.

Each of these secondary search modes becomes very much more powerful when combined (Boolean AND) with the PEAK search. One can, for example, search for all compounds of a given molecular weight that have certain mass spectral peaks. The option PMF permits, as an example, identification of all the compounds containing 19 carbon atoms and one sulfur atom that give a base peak in their mass spectra of *m/e* 71. There is just one entry in the file, that for the antidepressant drug thiazesim that satisfies these criteria.

The program KB accepts the complete mass spectrum, and, using the technique developed by Hertz *et al.* (*10*), finds those spectra that are most similar to that of the unknown. This program, which can be run very inexpensively overnight when computer charges are lower, is suited to the user whose mass spectrometer–computer system is in contact, by telephone, with the MSSS.

Other programs in the MSSS are particularly designed for more effective user interaction with the system. The program CRAB may be used to register complaints and also to report errors that have been found in the data, and changes in the system can be disseminated to users by means of the program NEWS. Each search program has associated with it a HELP program. A user, doubtful of the workings of, for example, PMF, may type HELP PMF, and a short explanation of that program will be printed out at the terminal.

The MSSS operates with a series of fixed computer charges which range from $1 to $6. The most expensive search option is KB, which currently costs $6 ($2 in overnight batch) and which involves a sequential search through all the abbreviated spectra in the file. This option is very extensively used, particularly by those whose laboratory systems are on-line to MSSS by way of direct minicomputer interfaces. Interfacing of laboratory minicomputers to MSSS has proved to be one of the more successful experiments in MSSS, and a number of manufacturers (*11*) are now marketing mass spectral data–acquisition systems that can be connected by telephone to the MSSS. The problem that arises when the mass spectrum being investigated has been measured upon a mixture of compounds has been addressed by means of the reverse searching program, PBM (*12*). This procedure is necessary because spectra in the data base are generally measured with pure compounds.

The total number of MSSS transactions per month is currently about 4000. The bulk of the searching is with PEAK and KB. Other searches are used significantly, however, with the possible exception of LOSS, which is a program designed to find all spectra exhibiting a given neutral loss from the molecular ion. Programs that remain idle in this way are ultimately removed from the MSSS. Over 250 laboratories, representing about 150 different organizations, are currently using MSSS, and this number is increasing at the rate of about one per week.

```
TYPE PEAK, MIN INT, MAX INT
CR TO EXIT, 1 FOR ID, REGN, QI, MW, MF AND NAME

USER: 283, 10, 40

    #REFS      M/E PEAKS

     329          283

NEXT REQUEST: 301, 5, 20

    #REFS      M/E PEAKS

      40        283 301

NEXT REQUEST: 245, 20, 90

    #REFS      M/E PEAKS

       1        283 301 245

NEXT REQUEST: 1

 ID#     REGN      QI     MW      MF

23445  2868022    894    344   C22H3203

              NAME

Pregn-4-ene-3,20-dione, 17-hydroxy-16,alpha.
-methyl- (8CI)
```

Fig. 3 (left). The PEAK search option of the MSSS. Fig. 4 (right). The SHIFT search option of the CNMR.

```
OPTION: SH

ENTER SHIFT, DEVIATION, MULT.: 23,.5,Q
  135 MATCHES USING 23.0

L (IST), Q (UIT), OR NEXT SHIFT, DEVIATION,
 MULT.: 43,.5,T   4 MATCHES USING 23.0  43.0

L (IST), Q (UIT), OR NEXT SHIFT, DEVIATION,
 MULT.: L

ID #                NAME
1488   RICKAMYCIN
1498   CHOLESTEROL
1645   TRANS 1,2-DIMETHYL CYCLOPENTANE
2072   2-ENDO-NORBORNANE-METHANOL

WANT THEIR STRUCTURES? (YES OR NO): NO
```

Carbon-13 Nuclear Magnetic Resonance Search System (CNMR)

This relatively recent addition to the CIS is currently operating in a "pilot" version. The data base consists of entries for 4000 spectra representing the same number of compounds. As a result of very vigorous collaboration between scientists in the United States, Germany, Hungary, Switzerland, the Netherlands, and Japan, this file is now being expanded very rapidly and, it is hoped, will contain some 10,000 entries within a year.

Each entry consists of the compound name, molecular formula, and CAS registry number. The proton-decoupled carbon-13 chemical shifts, with the relative intensities (when available), assignments (that is, the atoms responsible for the respective signals), the multiplicity (*S*, singlet; *D*, doublet; *T*, triplet; *Q*, quartet) of the single-frequency, off-resonance decoupled signals, and the experimental conditions under which the measurements were made are appended, as are the WLN for the compound and a compound classification code. A connection table and a numbered structure diagram for the compound are also included.

Once assembled, this data base can be searched in some very powerful ways. In a file of CNMR data, one must be able to identify each carbon atom uniquely and so numbered connection tables for every compound must be available. The display of a chemical structure at a command is simple once the connection table is accessible; more importantly, the presence in the file of the connection table makes it possible to search for structure fragments or substructures, a particularly useful capability in a file of NMR data.

Observed chemical shifts can be used (*13*) to search through the CNMR file to retrieve all the spectra that have the same shift or shifts. A typical search of this sort is shown in Fig. 4. The user is asked to enter a chemical shift in parts per million on the tetramethylsilane scale that is normally used in CNMR spectroscopy and a permissible deviation from

this value. The search is then carried out in exactly the same way as in MSSS, except that no intensity information is used, and the number of hits is reported to the user. At this point, he can terminate the search, inspect the hits that have already been found, or enter another shift, in which case the search is repeated with the second shift, the two lists of hits are intersected, and the number of spectra with both shifts is reported to the user. In addition to the chemical shift, the user may enter the single-frequency, off-resonance decoupled multiplicity of the signal. This information is used as an additional filter on the number of hits.

Once the identification number and the name of an entry of interest have been retrieved from the SHIFT or molecular formula search programs, the entry itself can be printed out or examined on the computer-generated microfiche copy of the data base just as in MSSS. The SPEC option retrieves the name, the structural and molecular formulas of the compound, and the shifts in its CNMR spectrum. Also provided, as available, are the relative intensities of the lines, expressed as a percentage of the most intense line, the single-frequency, off-resonance decoupled multiplicity of each line, and its assignment. An example of the output from the SPEC program is given in Fig. 5. The other options in the CNMR search system are the NEWS, CRAB, and HELP programs, which operate just as in MSSS. The CNMR search system is maintained upon the network by the Netherlands Organization for Chemical Information, which pays the annual disk storage costs for the data base. An annual subscription fee of $100 associated with the use of this system will be instituted in January 1977, and the computer charges for searching are set at $1 or less per search.

X-ray Crystal Structure Retrieval Program

The Cambridge Crystal Structure File is a collection of some 10,000 organic crystal structures that have been reported since 1960 (*14*). The file is leased from Cambridge University by NIH on behalf of the entire United States and currently contains a bibliographic section in addition to the structural data. To this file the NIH and EPA have added registry numbers, standard nomenclature, and connection tables.

Computer programs have been written by Feldmann *et al.* (*15*) to search, display, and manipulate data of the sort that

```
OPTION: SPEC

ID #: 1510

#: 1510

          C2
          *  +
        *     +
      C3       N
      +        *
      +        *
      C4       C5
     *  *     +
    *    * +
 C8+C7     C5

4-VINYL PYRIDINE
C7.H7.N
U.S. BUREAU OF MINES BULLETIN, 649, (1969)
SOLVENT: UNKNOWN
   5 LINES IN THE SPECTRUM.
WANT A LISTING? (YES OR NO): YES

ID # 1510

 SHIFT        MULT        INTENS        ASSIGN

 117.7          T           100            8
 120.2          D           100            3
 140.3          D           100            7
 143.8          S           100            4
 149.7          D           100            2
```

```
TYPE YOUR INITIALS (3 CHARACTERS) XXX = GWA
SELECT FROM FOLLOWING AVAILABLE FILES:

      1 - MASS SPECTOMETRY
      2 - PESTICIDE
      3 - WATER HAZARDS
      4 - OIL AND HAZARDOUS MATERIALS
      5 - EPILEPSY
      6 - CARBON 13 NMR
>6
FILE NOW CARBON 13 NMR
* RING 5
* ABRAN 2 2
* SATOM 1
SPECIFY ELEMENT SYMBOL = N
* ABRAN 6 1
* SBOND 6 8
SELECT BOND STATE    TR=RING TAUT      TC=CHAIN TAUT
RS=RING SINGLE     RD=RING DOUBLE'    RT=RING TRIPLE
CS=CHAIN SINGLE'    CD=CHAIN DOUBLE   CT=CHAIN TRIPLE
RA=RING ALTERNATING      R=RING ONLY     C=CHAIN ONLY
S=ANY SINGLE     D=ANY DOUBLE      T=ANY TRIPLE     A=ANY BOND

>CD
* SATOM 7 8
SPECIFY ELEMENT SYMBOL = O
* D

 7O  8O
  * +
   6
    *
     2***1N
     *   *
     3   5
      * *
       4

* FPROBE 2
E TO EXIT FROM ALL FRAGMENT SEARCHS
T TO EXIT FROM ANY ONE FRAGMENT SEARCH

CENTRAL ATOM C
N      ANY RING
C      ANY CHAIN
C      ANY RING
OCCURRENCE COUNT   1
 FOUND      UNDER      LEVEL      MATCH
   3846         1          1      EXACT
   3430         1          2      EXACT
    584         1          3      EXACT
    174         4          4      IMBED
    108         3          5      EXACT
     59         3          6      IMBED
     59         3          7      EXACT
     58         4          8      IMBED
     58         7          9      IMBED
FILE = 1      MERGED REFERENCES =      53
* RPROBE
    STIT PAT   1
    STIT SEQ   N
SUBSTIT PAT   2
 FOUND      UNDER      LEVEL          MATCH
    258         1      RING  NUC      EXACT
     72         1      STIT  PATT     EXACT
     25         1      STIT  SEQ      EXACT
      7         1      SUBST PATT     EXACT
FILE = 2      MERGED REFERENCES =       7

*INTER 1 2
FILE = 3    INTERSECTED REFERENCES =            1
SOURCE FILES WERE:      1      2
* SSHOW 3
SHOW EVERY NTH REFERENCE     (1) =
STRUCTURE    1 CAS REGISTRY NUMBER       147853
```

```
        O
        %
       C
      * %
N***C    O
*   *
C   C
 * *
  C
```

Fig. 5 (top). The SPEC option of the CNMR. Fig. 6 (bottom). Example of a substructure search.

are in this file. These programs use the device-independent Omnigraph display software developed by Sproull (*16*).

The file may be searched in a variety of ways to find compounds that fulfill specific criteria such as molecular formula, molecular weight, or coordinate data. As with all the other CIS data files, there is in this file a connection table for every compound and so searches may be conducted for a specific structure or substructure as described below. Once an entry has been identified, the molecule can be displayed and the display inspected, rotated, or redisplayed as a stereo pair. Torsion angles, dihedral angles, and interatomic distances can be calculated. This system is available on the network and can be accessed by way of a vector cathode-ray tube terminal or a lower-speed printer terminal, but the latter cannot handle any graphical representation. All the component programs except the substructure search are transaction-priced.

X-ray Crystal Data Retrieval Program

The Crystal Data Determinative Tables produced by the National Bureau of Standards (NBS) contain x-ray diffraction data for some 24,000 single crystals, both organic and inorganic (*17*). Each entry consists of the cell parameters (*A*, *B*, *C*, alpha, beta, and gamma), the number (*Z*) of molecules in the unit cell, the measured and calculated density, the molecular formula, and, depending upon the crystal system, two determinative ratios (for example, *A*/*B* and *A*/*C*); in the near future each entry will also have a CAS registry number.

The programs for searching through this data base, under development at NIH, permit searches for specific space groups, densities, molecular formulas, or unit cells of given characteristics. Any two of these searches can be intersected in a binary AND.

Completion of the development and testing of the software for this component of the CIS is scheduled for 1977. The data base and programs will then be made generally available through the network.

X-ray Powder Diffraction Retrieval Program

A compilation of the powder diffraction patterns of some 27,000 materials has been assembled by the Joint Committee on Powder Diffraction (*18*). The data for each entry consist of the relative

intensity, normalized to that of the most intense line, together with the d spacing of the line. The single problem in searching through such a file lies in the fact that, in practice, powder diffraction patterns are very often measured upon mixtures, whereas the patterns in the data base are derived from relatively pure compounds. For a problem of this sort, the reverse search technique (*19*) upon which the PBM component of the MSSS is based is well suited.

Work is now in progress to adapt a reverse search program to the x-ray powder diffraction file and subsequently to make the resulting system available. It is not expected that this work will be completed before early 1977, and the expected cost to users of this program, which will be managed by the Joint Committee on Powder Diffraction, is at present unknown.

Substructure Searching (SSS)

For the chemist a very desirable method of using the CIS is to search for every occurrence of a complete structural formula or fragment, as opposed to a molecular formula. This procedure is termed substructure searching (*20*) and involves a search through a file of connection tables for the part that has been specified by the user.

A preliminary step in substructure searching is the preparation of a query, and this may be done interactively as shown in Fig. 6. Hydrogen atoms are implicit, and nodes (atoms) that are not further defined are considered to represent carbon. With these programs, even complicated structures can be developed in an elapsed time of under 5 minutes, corresponding to a few seconds of computer time at most.

Once the appropriate connection table has been generated, a substructure search can be initiated. A typical first step would be a fragment probe search, in this case, for the occurrence of atom-centered fragments identical to node 2, the most characteristic node in the substructure. This results in 53 hits. A subsequent search (RPROBE) for all compounds containing a pyrrolidine ring substituted at the 2-position gives seven hits; intersection of these two files of answers results in only one compound, proline, CAS registry number 147853, that satisfies all the criteria that have been defined. Small files can be examined for exact matches by the program SUBSTRUCTURE SEARCH. This program seeks a precise correspondence between the user-generated substructure and structures in the file by an atom-by-atom search of each structure.

```
SUBJECT SEARCH

TYPE FIRST 3 LETTERS OF SUBJECT NAME
OR TO EXIT, 1 FOR LIST OF REFERENCES

SUBJECT: DATA PROCESSING

THE FILE CONTAINS
INDEX #     # OF REFS       SUBJECT
  1           1575          DATA PROCESSING
  2           1161          HIGH RESOLVING POWER DATA
  3          16093          LOW RESOLVING POWER DATA
  4            216          ON LINE DATA PROCESSING
  5            177          PROCESSING OF HRP DATA


ENTER THE INDEX NUMBER CORRESPONDING TO THE SUBJECT YOU WANT
INDEX NUMBERS: 1
    # REFS      SUBJECTS

    1575        DATA PROCESSING

NEXT SUBJECT: BOOK

THE FILE CONTAINS
INDEX #     # OF REFS       SUBJECT
  1            462          BOOK

ENTER THE INDEX NUMBER CORRESPONDING TO THE SUBJECT YOU WANT
INDEX NUMBERS: 1
    # REFS      SUBJECTS

     55         DATA PROCESSING
                BOOK

NEXT SUBJECT: BIOCHEMICAL

THE FILE CONTAINS
INDEX #     # OF REFS       SUBJECT
  1           5221          BIOCHEMICAL


ENTER THE INDEX NUMBER CORRESPONDING TO THE SUBJECT YOU WANT
INDEX NUMBERS: 1
    # REFS      SUBJECTS

      17        DATA PROCESSING
                BOOK
                BIOCHEMICAL


NEXT SUBJECT: ISOTOPIC ANALYSIS

THE FILE CONTAINS
INDEX #     # OF REFS       SUBJECT
  1           4102          ISOTOPE ANALYSIS
  2           1316          ISOTOPE DILUTION
  3            598          ISOTOPE EFFECTS
  4            394          ISOTOPE FRACTIONATION
  5            340          ISOTOPE SEPARATOR
  6            129          NATURAL ISOTOPIC ABUNDANCE


ENTER THE INDEX NUMBER CORRESPONDING TO THE SUBJECT YOU WANT
INDEX NUMBERS: 1
    # REFS      SUBJECTS

      5         DATA PROCESSING
                BOOK
                BIOCHEMICAL
                ISOTOPE ANALYSIS
```

Fig. 7 (left). A SUBJECT search in the *Mass Spectrometry Bulletin*.

```
AUTHOR SEARCH

TYPE FIRST 4 LETTERS OF AUTHOR'S NAME
OR TO EXIT, 1 FOR LIST OF REFERENCES

AUTHOR: NOVOTNY

THE FILE CONTAINS
INDEX #     # OF REFS       AUTHOR
  1            8           NOVOSELOVA A.V.
  2            1           NOVOSELTSEV A.M.
  3            1           NOVOSIBIRSK A.K.S.
  4            1           NOVOSIOLOVA A.V.
  5            1           NOVOTNY J.
  6            2           NOVOTNY L.
  7            6           NOVOTNY M.

ENTER THE INDEX NUMBER(S) CORRESPONDING TO THE NAME(S) YOU WANT
INDEX NUMBERS: 5 6 7
    # REFS      AUTHORS

      9         NOVOTNY J.
                NOVOTNY L.
                NOVOTNY M.

NEXT AUTHOR: JANAK

THE FILE CONTAINS
INDEX #     # OF REFS       AUTHOR
  1            2           JANAK J.

ENTER THE INDEX NUMBER(S) CORRESPONDING TO THE NAME(S) YOU WANT
INDEX NUMBERS: 1
    # REFS      AUTHORS

      2         NOVOTNY J.
                NOVOTNY L.
                NOVOTNY M.
                JANAK J.

NEXT AUTHOR: 1

ALL BULLETIN YEARS? (Y OR N)

Y

  704523    ANALYSIS OF STEROID COMPOUNDS BY GAS CHROMATOGRAPHY Y (CZECH)
              NOVOTNY M.        JANAK J.        CHEM. LISTY V.66 N.7 P.693-726
            1972
  800196    ANALYSIS OF STEROID COMPOUNDS BY GAS CHROMATOGRAPH Y (CZECH)
              NOVOTNY M.        JANAK J.        CHEM. LISTY V.66 P.693-726  197
            3
```

Fig. 8 (right). An AUTHOR search in the *Mass Spectrometry Bulletin*.

An expensive but necessary step in the preparation of each data file in the CIS is to assign connection tables and CAS registry numbers to each compound in the file. As a result, the substructure search (SSS) programs will operate on any of the CIS files and any other file such as CHEMLINE (*21*) that has connection tables. The computer costs for the use of SSS are variable; the average cost of a substructure search is between $5 and $20.

These costs are uncomfortably high for many users, and, since further software optimization appears to be yielding diminishing returns, the use of structure codes as a presearch device to SSS is being investigated. In this approach, a series of about 400 numerical structure codes (known as CIDS codes) are computer-generated from the data base of the connection tables (*22*). The first step in SSS will be a presearch based upon the CIDS codes. If the efficiency of the program is substantially improved by this approach, the structure code search will perhaps be merged into SSS as an obligatory presearch procedure.

Mass Spectrometry Bulletin Search (BULL)

The *Mass Spectrometry Bulletin*, a publication of the United Kingdom Atomic Weapons Research Establishment, consists of about 56,000 abstracts of papers dealing with mass spectrometry published since the mid-1960's. Each citation is reduced to a number of subject key words or codes; in addition, the author's name, the journal reference, the relevant MSDC codes, and elements are retained. The resulting files, which comprise the *Mass Spectrometry Bulletin*, are used to generate a disk-resident file of citations which can be searched on the basis of the above features. This interactive search system is now part of MSSS, all search options being transaction-priced at $3.

This file can now be searched on the basis of subject, subject code, MSDC code, element, author's name, and the general index of the *Bulletin*. Boolean AND and NOT operators can be applied to these searches, and one can, for example, locate all the papers dealing with tungsten except those that also treat rhenium.

Because the author, subject, and general indexes of the *Bulletin* use nonsystematic terms (author's name, for example), misspelling of queries is common. The conversational programs deal with this problem by conducting the search with the first few letters of the input. The different answers that are obtained are listed for the user as shown in the example of the subject search given in Fig. 7. The first query retrieves 1575 references in which data-processing is discussed. The second query limits the answer list to the 55 book references in which data-processing is discussed. Only 17 of these also deal with biochemistry, and only five with isotope analysis. In an author search, as shown in Fig. 8, a request is made for all papers published by Novotny, whose initials are uncertain, and Janak. The two retrieved citations can then be listed as in Fig. 8, at which point the user has the option of limiting the citations listed to those of specific years or of continuing the search with the name of a third author added. It is also possible to search the *Bulletin* for papers dealing with specific elements or with subjects that appear in the general index.

X-ray Crystal Literature Retrieval Program

A section of the Cambridge Crystal Structure File (*14*) currently contains some 14,000 literature citations to published crystallographic work. This file has been made a part of the CIS, and search programs have been appended to it by H. J. Bernstein of the Brookhaven National Laboratory. The file may be searched for a given structure or substructure, author, molecular formula, or molecular weight.

Proton Affinity Retrieval Program

The gas-phase proton affinity determines the behavior and utility of a molecule in chemical ionization mass spectrometry. H. M. Rosenstock and his coworkers at NBS have produced a file of some 500 measured proton affinities together with an annotated bibliography of the appropriate literature citations. The proton affinity data are being merged into MSSS, and the bibliographic component of this file will be appended to the BULL component.

Graphical Interactive NMR Analysis Program (GINA)

A problem that frequently arises in NMR spectroscopy is that a spectrum is too complex to yield to first-order analysis. The program GINA (*23*) is designed to deal with this problem by using estimated values for the various coupling constants and chemical shifts and calculating the expected NMR spectrum. The user can compare this spectrum with the observed spectrum, then alter one or more of the variables, and compare the new calculated spectrum with the observed spectrum. In this way, an iterative approach is made to the true coupling constants and chemical shifts with the user acting as the transducer in the feedback loop. The program, which can be used with graphics terminals or teletypes, is running at NIH and is currently being merged into the CIS.

Mathematical Modeling System (MLAB)

There are many scientists who could use the mathematical power of computers but who are dissuaded by the need for programs. A number of interactive program packages have been designed to overcome this difficulty, and one of the more powerful of these, MLAB, developed at NIH (*24*), has been incorporated into the CIS.

This program is designed to accept a set of data from the user and to perform, upon command, any of a wide variety of mathematical manipulations upon these data. These include linear, nonlinear, and multiple regression; scalar and matrix computation; differential calculus; initial and boundary value problems; root-finding; and minimization.

Isotopic Label Incorporation Determination (LABDET)

As a result of the difficulties surrounding the use of radioisotopes in medical research, stable isotopes are playing an increasingly larger role in this area. At lower levels of isotopic incorporation, however, there arises in mass spectrometric detection and quantitation the difficulty that naturally occurring carbon consists of a large amount of carbon-12 mixed with a small amount of carbon-13 (about 1 percent). Stable isotopes of other elements, for example, deuterium, nitrogen-15, and oxygen-18, occur naturally in very small amounts, mixed with their major isotope. In the mass spectrum, fragment ions formed by the loss of a hydrogen atom from the molecular ion alter the observed ion intensities. It is thus not entirely straightforward to derive the incorporation levels from the mass spectral data, and the LABDET program has been written (*25*) to deal with this problem.

The program accepts the mass spectra

of the unlabeled and the labeled compound. From these, an estimate is made of the level of isotope incorporation in the labeled compound. A theoretical spectrum for this level of isotope is then calculated and compared with the experimental spectrum. An iterative process to fit the estimated isotope level to the spectrum of the labeled compound culminates after a specified number of cycles in the calculation of a correlation coefficient. This program, which handles a tedious calculation very rapidly, has been merged into the CIS and is transaction-priced at $2.

Conformational Analysis of Molecules in Solution (CAMSEQ)

The conformation of a molecule in solution is related to but not necessarily the same as that in the crystal state. The major purpose of the program package CAMSEQ, written by Weintraub and Hopfinger (*26*), is to calculate by empirical and quantum mechanical techniques the molecular conformation of a particular molecule that has the lowest free energy in solution.

The program can work with coordinate data supplied by the user, or it can generate coordinate data from a molecular structure. It then systematically alters the torsion angles in the molecule to generate new conformations, for each of which statistical thermodynamic probabilities are calculated, based on the use of potential (steric, electrostatic, and torsional) functions and terms for the free energy associated with hydrogen-bonding, molecule-solvent, and molecule-dipole interactions.

This program package is currently running on the NIH PDP10 where the software is being optimized. In its present form, the core demand of the programs is very high (about 42,000 36-bit words), and, although there are economic questions concerning the feasibility of making the program immediately available through a network, it is hoped that this will be accomplished in 1977.

CIS Management

Each component of the CIS, when it leaves the U.S. government computer and enters the private sector, does so under the auspices of a non–United States government sponsor. This step is taken in conformity with the Office of Management and Budget circular A76 (*27*) and further ensures that the particular program is then subject to the normal free market forces. If it consistently loses money, that is, if it generates insufficient revenues in subscription fees to defray the costs of disk storage, the sponsor, who pays the disk storage charges, is free to decline further sponsorship and the program, at this point, becomes at least temporarily a dead letter.

A number of details in the CIS structure deserve some further discussion because they represent interesting questions that remain open, and these are treated separately below.

User Manuals

A great deal of effort has been expended to ensure that CIS components are sufficiently alike that a user can switch from one component to another without any extensive reeducation process. Nevertheless, two further steps are taken to assist the user. First, every program in the system has associated with it a HELP file which can be accessed at any time by a user in difficulty. Second, there is written for each component an extensive manual. These manuals, which often exceed 50 pages, are written with some care and cover every aspect of the program they describe. The writing and the printing of these manuals is expensive, and this problem has not been solved yet except by the unsatisfactory approach of selling copies of the manual to users.

Future Directions

As the CIS grows and revenues accruing from the use of the system increase, it seems likely that a dedicated PDP10 could be leased or purchased jointly by the sponsors of CIS components. This would, to a considerable extent, reduce the computer costs discussed above. Increased use of the CIS will assure that the current subscription fees and computation transaction prices will be maximum values and not bases from which prices will increase. Such a prediction is based upon the expectation of a level of use that has not yet been demonstrated with the few components that are currently available to the scientific community.

References and Notes

1. R. J. Feldmann, S. R. Heller, K. P. Shapiro, R. S. Heller, *J. Chem. Doc.* **12**, 41 (1972); S. R. Heller, in *Computer Representation and Manipulation of Chemical Information*, W. T. Wipke, S. R. Heller, R. J. Feldmann, E. Hyde, Eds. (Wiley, New York, 1974), pp. 175–202.
2. S. R. Kimbleton and G. M. Schneider, *Comput. Surv.* 7, 129 (1975).
3. The major portion of the distance involved in satellite transmissions is in fact the distance to and from the satellite. Thus, although North America and England are only about 4800 kilometers apart, a telephone signal from one to the other travels some 80,000 kilometers, by way of a synchronous satellite at an altitude of about 40,000 kilometers, and therefore takes 269 milliseconds.
4. S. R. Heller, H. M. Fales, G. W. A. Milne, *Org. Mass Spectrom.* **7**, 107 (1973); S. R. Heller, D. A. Koniver, H. M. Fales, G. W. A. Milne, *Anal. Chem.* **46**, 947 (1974); S. R. Heller, R. J. Feldmann, H. M. Fales, G. W. A. Milne, *J. Chem. Doc.* **13**, 130 (1973); R. S. Heller, G. W. A. Milne, R. J. Feldmann, S. R. Heller, *J. Chem. Inform. Computer Sci.* **16**, 176 (1976).
5. S. R. Heller, *Anal. Chem.* **44**, 1951 (1972).
6. The WLN are generated from connection tables by means of a program developed by H. Gelernter, Department of Computer Science, State University of New York, Stony Brook 11794 [unpublished work].
7. For details, contact Dr. D. L. Lide, Jr., Office of Standard Reference Data, National Bureau of Standards, Washington, D.C. 20234.
8. ADP Network Services, Inc., Cyphernetics Division, Ann Arbor, Mich. 48106.
9. The combined data base of 30,000 mass spectra and 50,000 literature citations, together with all the programs, occupy some 70 million characters (bytes) of disk storage.
10. H. S. Hertz, R. A. Hites, K. Biemann, *Anal. Chem.* **43**, 681 (1971).
11. These include Finnigan, Hewlett-Packard, INCOS, Varian, and V-G Data Systems.
12. G. M. Pesyna, R. Venkataraghavan, H. E. Dayringer, F. W. McLafferty, *Anal. Chem.* **48**, 1362 (1976).
13. B. A. Jezl and D. Dalrymple, *ibid.* **47**, 203 (1975).
14. O. Kennard, D. G. Watson, W. G. Town, *J. Chem. Doc.* **12**, 14 (1972).
15. R. J. Feldmann, S. R. Heller, C. R. T. Bacon, *ibid.*, p. 234.
16. R. F. Sproull, publication CSL 73-4 (available from Xerox, Inc., Palo Alto, Calif., 1973).
17. These data are available through the National Technical Information Service, Springfield, Va. 22151, as NBS tape 9.
18. G. McCarthy and G. G. Johnson, paper C3 presented as part of the Proceedings of the American Crystallographic Association meeting, State College, Pa., 1974.
19. F. P. Abramson, *Anal. Chem.* **47**, 45 (1975).
20. R. J. Feldmann, in *Computer Representation and Manipulation of Chemical Information*, W. T. Wipke, S. R. Heller, R. J. Feldmann, E. Hyde, Eds. (Wiley, New York, 1974), pp. 55–81.
21. B. Vasta, *J. Chem. Inform. Computer Sci.*, in press.
22. *Handbook of CIDS Chemical Search Keys* (Fein-Marquart Associates, Inc., Baltimore, November 1973).
23. S. R. Heller and A. E. Jacobson, *Anal. Chem.* **44**, 2219 (1972); R. B. Johannesen, J. A. Ferretti, R. K. Harris, *J. Magn. Reson.* **3**, 84 (1970).
24. G. D. Knott and R. I. Shrager, *Assoc. Comput. Mach. SIGGRAPH Not.* **6**, 138 (1972).
25. C. F. Hammer, Department of Chemistry, Georgetown University, unpublished work.
26. H. J. R. Weintraub and A. J. Hopfinger, *Intl. J. Quantum Chem.* **9**, 203 (1975).
27. "Circular A76" (Office of Management and Budget, Washington, D.C., August 1967).