structures. Other examples include simple mixtures, alloys, and polymers formed from a mixture of monomers.

(10) Hansch, C.; Fujita, T. "ρ–σ–π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure". *J. Am. Chem. Soc.* **1964**, *86*, 1616-1626.

(11) Potenzone, R. Ph.D. Thesis, Case Western Reserve University, Cleveland, OH, Feb 1979. See also: Potenzone, R.; Hopfinger, A. J. "Structural Correlates of Carcinogenesis and Mutagenesis. A Guide To Testing Priorities". *Proceedings of the 2nd FDA Office of Science Summer Symposium*, 28 August, 1978; U.S. Government Printing Office: Washington, DC, 1978.

(12) Hansch, C.; Leo, A. *Substitution Constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.

(13) Hodes, L. J. "Computer-Aided Selection of Compounds for Antitumor Screening: Validation of a Statistical-Heuristic Method". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 128-132. Hodes, L. J. "Selection of Molecular Fragment Features for Structure-Activity Studies in Antitumor Screening". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 132-136. See also: Hodes, L. J.; Hazard, G. F.; Geran, R. I.; Richman, S. "A Statistical-Heuristic Method for Automated Selection of Drugs for Screening". *J. Med. Chem.* **1977**, *20*, 469-475. Hodes, L. J. "Computer-Aided Selection of Novel Antitumor Drugs for Animal Screening". *ACS Symp. Ser.* **1979**, *112*, 583-603.

(14) Murine lymphocytic leukemia, strain P388, is the preliminary screen against which all compounds have been tested since 1976. The number of compounds for which there are P388 data in the DIS is now in excess of 200 000.

(15) The importance of this check is not so much for data validation (see ref 8) but to guard against the possibility that by misrouting correspondence some confidential information may be transmitted to persons outside the government who are not authorized to see it.

(16) The Acquisitions contractor is not equipped to handle hazardous substances and is therefore instructed not to weigh samples that are received. When the sample weight is provided by the supplier, it can be entered and flagged as "estimated".

(17) The registration is carried out online, in real time. The registration data will not, however, be searchable until the next subsequent inversion of the Chemistry database, which is a weekly event.

(18) From this point on, all movement of this sample is tracked by means of the barcoded label, which must be scanned as it enters or leaves the storage facility. This, coupled with electronic weighing, as is described in a subsequent paper, is crucial to the maintenance of an accurate inventory.

(19) Milne, G. W. A.; Feldman, A.; Miller, J. A.; Daly, G. P. "The NCI Drug Information System. 3. The DIS Chemistry Module". *J. Chem. Inf. Comput. Sci.* **1986**, following paper in this issue.

The NCI Drug Information System. 3. The DIS Chemistry Module

G. W. A. MILNE* and ALFRED FELDMAN

Information Technology Branch, Division of Cancer Treatment, National Cancer Institute, Bethesda, Maryland 20205

J. A. MILLER and G. P. DALY

Fein-Marquart Associates, Baltimore, Maryland 21212

Received April 21, 1986

The Chemistry Module of the Drug Information System (DIS) handles a database of 400 000 structures. New or modified records are created in this database on a daily basis and are merged into the file promptly. The Chemistry database is searchable in a wide variety of ways and provides novel methods for both input and output of chemical structures.

INTRODUCTION

In terms of data content, the Chemistry and Biology databases are regarded as the most important in the National Cancer Institute's Drug Information System (DIS). From the point of view of capability, however, the programs that manipulate these databases differ widely. While searching of the Biology database, which is described in part 5 of this series, is straightforward, the search of the Chemistry database is very complex. Corresponding differences are encountered in file management, in input and output processing, and so on. These differences are reflected in the sheer amount of code required by the Chemistry Module,[1] which makes it by far the largest of the DIS and which has required more developmental effort than any other. The large amounts of software reflect the special needs that pervade every aspect of chemical data processing. Here one finds continual mingling of graphics and alphanumeric material, requirements for exhaustive, accurate, and fast searching, complex searching algorithms, many different modes of input and output of data, a need for frequent updating, and a major data security concern. All of these problems come together in the Chemistry area.

For chemical input and output, special hardware is required, operated by specific, device-dependent software, and this software must be interfaced with the rest of the system. Chemical structures further require specialized software for validation, updating, indexing, searching, and so on. Much of this processing is severely computationally intensive, warranting the introduction of techniques capable of alleviating the system's load, but which add to the complexity of the system. Graphical chemical data undergo only the normal processes of editing, updating, and searching. These functions, seemingly similar to the processes undergone by conventional data, i.e., text and numbers, in actuality are so incompatible that it was deemed best not to let the chemical requirements burden the TDRS—the resident database management system of the DIS. It was decided to maintain a separate database, with separate update, maintenance, and search software for chemical structures, and to interface only the outputs of these two systems. Having to deal not only with the complexities of chemical structure processing but also having to maintain and interface two separate file management systems contributes substantially to the complexity of the Chemistry module.

Systems of the magnitude of the DIS are not created from scratch, but built on experience gained with earlier systems. Thus the DIS benefited considerably from the experience gained with the system that it replaced, a system operated for many years under contract to the NCI by two organizations: Chemical Abstracts Service (CAS), which dealt with the chemical information, and VSE Corp., which handled the biological data.[2] But, partly because that system was not well integrated—its chemistry ran on one computer, its biology on another, and the linkage between the two was never adequate—the NCI decided in 1981 to design and develop the Drug Information System (DIS). The predecessor system and template for the DIS became the Structure and Nomenclature Search System (SANSS), a system which, for a number of years, had been operated jointly by the NIH and the EPA.[3] A summary of the relationships between the SANSS and the

0095-2338/86/1626-0168$01.50/0

DIS is provided in Appendix A.

This parentage allowed the DIS to benefit not only from the experience gained during the earlier operation of the SANSS but also from the adaptation of much of the SANSS software to the DIS. Since the chemistry module possesses the largest amount of code, it benefited the most, but because the success or failure of numerous design features in SANSS had become clear during the 8 years that SANSS has been in general use,[4] much information of strategic value could also be transferred to the new system design. The adapted software included the data preparation programs, specifically, in the chemistry area, programs processing the chemical structure, molecular formula, and the various textual and numerical fields in the basic data record for compounds. The software relating to Boolean logic was usable, and many programs for file maintenance, for processing and validating input, and for generating output also proved applicable. The larger size of the DIS, as well as additional requirements such as the monitoring of shipping and testing, required additions and enhancements. The structure display capabilities of SANSS, in particular, required major enhancements to be acceptable in the DIS environment. Existing display capabilities had to be supplemented because of the increasingly common use of graphic terminals (input) and laser printers (output) by the system. Last but not least, many utilities, both internal and external, had to be created. The major differences between the SANSS and the DIS will be found in Appendix A.

Inevitably, a parentage manifests itself also in the form of idiosyncrasies and constraints whose justification is primarily historic. The DIS, for example, inherited from SANSS the search for atom-centered fragments (FPROBE), which, perhaps because of file context, is one of the least used of the structural screens in the system. Conversely, the CAS/VSE forerunners to the DIS had no capability for chemical name searching. This has been incorporated into the DIS, where it has proved to be an important means of retrieval.

STRUCTURE INPUT

Several disparate requirements in the area of structure input exist in the context of the NCI drug screening program, and the DIS has attempted to meet these. As the database is being built, it is essential to be able to enter chemically accurate and graphically acceptable structures rapidly. Users who are required to accomplish such high-volume input can afford to acquire specialized terminals such as microprocessors. Other users, however, wish merely to enter a single structure for search purposes and need to be able to use a standard computer terminal for this purpose. In the same way, distinct choices present themselves at output time: mingling graphics and text is highly desirable, and the DIS can do this. When an instantaneous output is required, however, the DIS allows recourse to a wide variety of output devices, not all of which can provide high-quality graphics.

Since the earliest days of chemistry, attempts have been made to handle chemical structures like any other data. Nomenclature has been developed to allow referencing such data by means of the spoken and written word, as well as through the use of conventional indexes, and line notations allowing the keypunch to be used to enter structures have appeared. All these attempts to couch chemical structures in a textlike format, however, require a translation. This translation could sometimes be done mentally but, as often as not, it had to be accomplished with pencil and paper. The textlike translation thereby obtained was more difficult to comprehend and thus was prone to errors. In the past decade, graphic terminals have finally made it possible to avoid this time-consuming and unproductive mental effort by allowing structures to be captured in their normal graphic form. The structures, of course, still are translated, but by the computer now. In modern systems, the connection table is obtained as an invisible byproduct of entering a structure. The process is foolproof in that if the structure appears graphically to be correct, then, no matter how it was constructed, the corresponding connection table will be correct as well.

The optimal means of structure entry, it turns out, is not easy to determine. The drawing of structures by means of a touch pen on the face of a computer terminal's screen causes fatigue, the operator's arm having to be raised for a long period of time. To avoid this, the "graphic tablet", the "joy stick", the "mouse", the "thumbwheel", and so on have been developed as alternatives to the touch pen. But with either touch pen or its alternatives, the entry of characters proves cumbersome. Drawing these would represent a return to pre-typewriter days, and so two approaches have developed for dealing with them. Some systems[3] allow the user to pick a character with the touch pen from a menu and to "drag" it to a location on the display. These systems usually allow also frequently occurring structural fragments, such as ring systems and functions, to be picked from menus. Other systems allow keying characters from the keyboard. The first approach is slow. In the latter, touch pen and keyboard do not mix well. Picking up the pen interrupts the cadence of typing. It also distracts the eye, which touch typing—done blindly—had freed to attend to the typescript. From an ergonomic point of view, alternating between keyboard and touch pen is not sound.

Judicious review of the input problem by the DIS led to the adoption of several input methods. The structure input method used by SANSS was adopted for those doing structure and substructure searches in the DIS. In this procedure, structures are represented in a format suitable for display on nongraphic machines. The structure may be distorted, and bonds are represented by arbitrary symbols. The single bond is represented by ***, the double bond by +++, and so on. This convention, for all its deficiencies, has one redeeming feature; it can be used on any computer terminal. For this reason it survives, being used in the NIH-EPA CIS as well as in CAS ONLINE.

For those concerned with database building, the DIS offers a choice of two input methods. The first of these, which is currently in production use, uses a touch pen, while in the second all communication is via a keyboard. The former is very easy to learn, but in the latter case, optimum speed is achieved after only few day's practice. The touch-pen approach employs a Victor 9000; it is popular with users and has been used to enter tens of thousands of structures in an acceptable time per structure.[5] The newer approach uses an IBM PC (XT or AT) and shows promise of being much faster than the touch-pen procedure because it is possible for the user to adopt a typist's rhythm. The number of keystrokes per structure is very low (20–60), and if this is done with the speed of a professional typist, as is possible, the time per structure is greatly reduced.

STRUCTURE OUTPUT

The advantages of providing chemical structures as computer output, rather than as codes or distorted representations, are self-evident. Several problems, however, must be overcome before these advantages can be fully realized. Just as with structure input, different output methods are necessary to accommodate the different hardware available to various users, as well as their differing needs in terms of system response. These are described in this section.

A. Intermingling of Text and Graphics. Chemical structures can be regarded as an open-ended set of ideographs, and in this sense chemistry has something in common with ideographic languages, such as Japanese. Since the development of printing, the keying of texts has been increasingly facilitated by ever more efficient typesetting machines. From their beginning, however, chemical structures, like Japanese characters, had to be printed from hand-drawn "cuts", not unlike block printing, which antedates the invention of movable type.

This difference persisted in computer output devices. Conventional line printers cannot accommodate graphics, and offline graphic devices, particularly plotters, are usually too slow to produce volume output. Graphic terminals were the first devices capable of handling both text and graphics, and, as they became both cheaper and more sophisticated, they became increasingly useful. But these proved incapable of producing "hard" copies of acceptable quality of their displays. Nor could hard copies be produced with adequate speed or at an adequate cost.

The first devices capable of combining text and graphics were the printer/plotters, of which the VERSATEC[6] is the best-known example. These machines are fairly fast and have good graphics, and for these reasons, the NCI used to maintain a printer/plotter in addition to an offline printer. Neither machine could replace the other. The printer/plotters nonetheless have certain drawbacks. Characters within graphs, being generated from vectors, are not well formed. Also, these machines use a wet process, and this led to a considerable maintenance burden.

The recent introduction of laser printers for computer output had a major impact on this situation. This technology is in the process of revolutionizing written communications in Japan,[7] and it is having no lesser effect on the printing of chemical structures in particular and diagrams in general. The Information Technology Branch of NCI acquired a Hewlett-Packard 2680, using it as a remote offline print station to its IBM mainframe. The laser printer is capable not only of mixing text and graphics but also of using different fonts and variable spacing. Examples of the output from this printer are given in parts 2 and 5 of this series of papers. The output is of far better quality than that produced by either the line printer or the printer/plotter, and this machine has replaced them both. It can print letterheads on blank paper, obviating the need for replacing paper stock. It can print letters in different languages and alphabets, such as Japanese. Finally, because many users like the rapid turnaround, they readily abandoned the printer/plotters for the laser printer, even though it reduces all plots to fit on a regular page.

Only for copying chemical structures directly from the face of an online graphic terminal is there, so far, no entirely satisfactory solution. Graphics copiers are still relatively expensive, space-consuming, slow, and incapable of producing high-quality copies.

B. Rendition of Structures. Two methods are commonly used for preparing structures for output, and the DIS takes advantage of both of them. The methods differ in the manner in which the coordinates of vectors and characters of a structure are obtained. In the first approach, they are captured at input and saved for output. This form of output is optimum in final reports and is used heavily. In the second, they are computed from the data in the connectivity table. Such output is far less esthetic and does not reflect, for example, stereochemistry. It is, however, very fast and terminal-independent, and can be invoked with any structure, be it newly entered as a query structure or retrieved from the database.

Retaining coordinates in the structures's record demands additional storage, more actually than the amount required for the connectivity table. This has provided the impetus for the second approach, the development of algorithms to generate structures from what information is available in the connectivity tables. The best known of these is the ASD of CAS.[8] Their advantage, beyond economy of storage, is that graphical representations can be obtained even where only connectivity tables are available. Their disadvantage is that the resulting structures are less esthetic; they are not necessarily rendered in their original appearance, nor even in their conventional appearance, neither of which are known to the generating algorithms. Any steric configuration denoted in the original structure will also be lost because such stereochemical bonds, while retained for structure output, are stored in the connection table as ordinary single bonds.

Such structures can be made terminal-independent, and the DIS uses them for this reason, as mentioned in the second part of this series.[9] The DIS further uses this approach to display the older DIS structures (the first 250 000 that were registered by NCI), the original input having been lost. Since then, however, the DIS has been capturing and storing the coordinates of its chemical input, so that its output structures are exact copies of those originally entered. The format of the connectivity tables, which do not include the coordinates, is shown in Appendix B.

C. Output Formatting. Chemical structures come in a variety of sizes, some filling a page, others fitting within a line of text; most are intermediate in size. They may have to be centered or placed near the margin. They may have to fit in a table or be imbedded in text. Occasionally, to facilitate comparisons, two structures must be shown together. The DIS allows the user to specify all such types of presentation of retrieved structures.

Output format specification is of concern for online terminal displays or for offline hard-copy printouts. Output formats are defined by the user; the availability of a number of commonly used output formats on the DIS facilitates their specification. Formats are available for displaying structures alone or one, two, or four to a page, and for structures with associated data. Normally, the system will automatically reduce the size of the larger structures and surround all structures uniformly with blank space. Formats may not only specify the disposition of structures and data but also select fonts and a letterhead. They are therefore the basis of the DIS letter-writing programs, which were described in the previous paper. In letters to potential submitters of compounds, the DIS determines the addressee's language and writes the letter in it, as was described in the previous paper in this series.

The output commands in the DIS are powerful in that they can direct the output requested to any of several output devices, such as the user's terminal, online impact printers, online laser printers, or merely disk. Then, depending upon the output device that is to be used, the DIS will select the appropriate form of the structure diagram.

QUERY FORMULATION

A. Background. Computer retrieval methods depend on classification. The more precisely the classification categories can be demarcated, the more effective are the searching procedures. Further improvement of retrieval performance depends on the accuracy with which relationships existing among the classification categories can be expressed, for example, as among "man", "dog", and "bite".

In text-based systems, the determination of categories is difficult. As a consequence, the number of categories used in retrieval systems is relatively few, and consistency in their assignment is usually poor. Relationships among categories are often ignored, notably in inverted list systems.

In contrast, clearly demarcated categories abound in chemical structural data. Any fragment, large or small, can define with exceptional precision a corresponding class of compounds. Furthermore, in the atom-by-atom search described below, a method is available for determining even the most complex relationship that may exist among fragments.

It is further possible—and a number of systems have done so—to make the use of categories completely transparent to the user, requiring specification of only the structure or substructure to be searched.

Any chemical can be named, and it follows that a file of chemical compounds can be searched by means of methods applicable to text data. Several commercial systems[10] provide access to large chemical databases upon this basis. But nomenclature imparts a bias because it permits the retrieval of structures for which names are available, while preventing that of the others. Structural relationships, moreover, are ignored, unless, by chance, they are reflected in the nomenclature. In attempts to redress these flaws, a few additional categories, such as number of rings, sizes of rings, element counts, molecular formula, and so on, have been used. Other systems such as CAS ONLINE, SANSS, and DARC[10] have been unwilling to put up with these limitations and have opted to exploit the more extensive and precise retrieval capabilities inherent in chemical structures. There are many chemicals whose structure is not known, but the number of "unstructured" compounds in a database such as the one under consideration here is close to zero.

B. Search Specification. The DIS has invested heavily in such chemical structure search software and has been able to provide its users with an assortment of varied and powerful commands. These, first of all, allow the searcher considerable latitude in the specification of a query. Queries may usually be resolved in more than one way, and internal double checking is thereby facilitated. The DIS commands further enable the user to monitor the progress of a search and to adjust its performance. Whereas in batch systems the logic of a query must necessarily be fully specified before submitting it, in an interactive system such as the DIS, an indication of the number of prospective hits, communicated to the searcher before they are displayed, allows refinement of the query by modification of the search criteria. With each modification, 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. The answers can then be either sampled by the user or printed or displayed in toto.

The interactive commands also enable the searcher to influence the rate of convergence of the search. While, to a considerable extent, the efficiency of the search is a function of software and system design, the experienced DIS searcher, through use of appropriate search strategies, can significantly affect the rate at which overall search objectives are realized.

The arsenal provided by the DIS includes commands for field searching applicable to text or numeric data as well as the special commands for searching chemical structural data. It further provides commands for performing logic on temporary results files. As already mentioned, there are commands for output formatting. And finally, there are utility commands that, for example, allow an index to be perused, the output of a query to be curtailed, or a search strategy to be saved.

1. Field Search Command. The commands usable with text and numeric data are the "field search commands", which use the standard field mnemonic/field value format common to the entire DIS. All the contents of the DIS files other than Chemistry (Biology, Inventory, etc.) consist of text or numeric data stored in fields and subject to field search commands. An important feature of the DIS is that the database need never be specified in a search. Every field mnemonic is unique, and the DIS knows which of the 24 databases is implied. This permits multiple database searching and relieves the user of a significant burden.

Much of the data on the Chemistry module consists of text and numbers. They include molecular weights, boiling points, and other physical constants; atom counts; ring counts; stereo text descriptors; and data relating to the Hodes model:[11] activity, novelty, and model version. Last but not least, there are chemical names. These are available for some 12 000 Selected Active Compounds (compounds of particular importance to NCI) and for about 40 000 others and include systematic chemical names, trivial names, trade names where available, and Wiswesser Line Notations, which are treated as synonyms.

Field search commands have the following syntax:

"Field mnemonic"/"Value"

"Field mnemonic" represents one or more of the data elements on file. The field mnemonics are qualified by their values. Thus, "NSC/740", which references a field mnemonic and a value, is a valid command, implying a search of the index labeled "NSC" for the entry "740". The search is trivial; it gives one hit:

```
OPTION? NSC/740
  One hit was found and stored in file 1 associated with the
CHEMISTRY database
```

"Value" represents the contents of a field. The field mnemonic CNAM references the field containing a compound's name; the value "CORTISONE" specified for it will retrieve all compounds entered as Cortisone.

```
OPTION? CNAM/CORTISONE
  One hit was found and stored in file 2 associated with the
CHEMISTRY database
```

The same field can be examined partially by using the field mnemonic NAMF. Thus "NAMF/CORTISONE" retrieves all compounds with "cortisone" somewhere in their name. This typically will lead to more than one retrieval:

```
OPTION? NAMF/CORTISONE
  8 hits were found and stored in file 3 associated with the
CHEMISTRY database
```

Numeric values are dealt with similarly. Thus, CAS/1247423 retrieves the compound with CAS Registry Number 1247-42-3.

```
OPTION? CAS/1247423
  One hit was found and stored in file 4 associated with the
CHEMISTRY database
```

And MW/437 retrieves all compounds with that molecular weight[12]

```
OPTION? MW/437
  493 hits were found and stored in file 5 associated with the
CHEMISTRY database
```

Values may be requested as ranges, for example: "DNSC/>1-APR-85" or "NSC/374533 TO 374650". With this capability, a user, for example, can specify any numeric value, such as a date or a molecular weight, to be exact, partial, or ranged.

The syntax of key specification accommodates Boolean logic. One may query for NAMF/NITRO AND FLUORO AND NOT CHLORO. The resulting search depends, however, upon the punctuation of the Boolean statement or, if no punctuation is given, the program's parsing of the search statement:

```
OPTION? (NAMF/NITRO AND NAMF/FLUORO)
NOT NAMF/CHLORO
  22 hits were found and stored in file 6 associated with the
CHEMISTRY database
```

The file to be searched is specified implicitly by the field mnemonics used. Field mnemonics are characteristic for each file: Thus, ADDR belongs to the NAMECODES file, CAMT belongs to the CHEMISTRY file, and TSY belongs to the Biology file. Not all values are searchable, but all fields are displayable. The field mnemonics MW, DACQ, and DNSC, for example, are searchable; other field mnemonics, such as CMSL (selection comments) and OPSP (other potential suppliers), are not. The definitions of the DIS fields show whether their values are searchable or not.

2. Structure Search Commands. A file of chemical structures lends itself to two kinds of searching, which in the DIS are called "identity" and "substructure" searches. In response to the first, the system will determine whether an exact match can be obtained between any structure on file and the submitted query structure.[13] In response to the latter, the system will retrieve any structure on file that contains the query structure imbedded within its own structure.

The processing techniques used to execute such searches not only differ from those used for field searching but also are of several types. Some systems, notably that of CAS, use different approaches for identity and substructure searching. A special file is devoted to identity searches, on which all structures are described by "normalized" codes, codes that are always identical for identical structures no matter how drawn. Such structures can be ordered uniquely, facilitating reference to them. The DIS uses precisely this approach. Its "identity search" is in fact a search for a hash-encoded version of the connection table in question; it is quite unrelated to the "substructure search". In substructure searches, however, the attributes, on which the order of the file depends, may not be relevant, as the search may be concerned with other attributes. This makes it necessary to maintain another file for substructure searching.

For substructure searches, two approaches are available. One depends on the retrieval from the database of all structures that contain a structural fragment present in the query structure. This is generally termed a "screen search". The other is based on a traversal of the query structure during which corresponding atoms and bonds on a file structure are matched. The process is repeated for each structure on a file. This is the so-called "atom-by-atom search".

Each of these two approaches has shortcomings. Any given atom-centered fragment will surely retrieve all compounds containing that fragment, but such a search will usually retrieve many false drops. For example, a search for all aromatic bromo compounds will retrieve bromobenzenes as well as bromonaphthalenes, even though only the former may have been sought. The defect manifests itself not through the loss of potential hits, but by an inability to exclude all misses. The atom-by-atom search, on the other hand, is very precise and will not retrieve false drops. Its problem is that it is too voracious of computer time to be considered for searching any sizable files. The earliest system capable of substructure searching[14] achieved both efficiency and precision by combining these two approaches, using fragment screens to reduce the set that was passed on to the substructure search. The efficiency, of course, is relative, and the complexity cannot be avoided. In dealing with chemical structures, at least two, if not three, complex procedures replace the field searching used with text and numbers.

The approach taken by the DIS then is to use screens to reduce the database of a half-million structures down to a manageable number of structures (one thousand or less), all of which possess all of the desirable structural fragments. This small list may then be searched upon an atom-by-atom basis.

a. Full Structure Searching. Once a query structure has been defined, as is described below, the single command IDENT will invoke a search through the Chemistry database for just that structure. The connection table corresponding to the query structure is packed into a hash-encoded format, and the resulting hash code is retrieved from a file containing all the hash codes derived from the Chemistry database. If there is more than one dot-disconnected structure, this process is repeated for each of them. The results files are all intersected, and then an atom-by-atom search is done through the final results file to guard against "hash code collisions". The search is fast and unambiguous and relies upon the fact that the connection table corresponding to the structure in question is in fact unique.

b. Predefined Screens. Search fragments can be picked by a searcher from a list and submitted as a query. The DIS provides its users with a lengthy list of so-called "Structural Features Codes", and they may be entered one at a time into a search program called SPROBE, which operates as below.

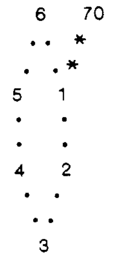
```
OPTION? SPROBE
  Specify Structural Feature Code and permissible multiplicity
limits.
Next SFC = SCN116
  Found 893 compounds having 1 or more occurrences of
SCN116

Next SFC = FG145
  Found 18 333 compounds having 1 or more occurrences of
FG145

Next SFC =
  175 hits were found and stored in file 7 associated with the
CHEMISTRY database
```

The first structural feature code (SFC) entered—SCN116—retrieves the 893 phenothiazines in the database. With the second SFC—FG145—a total of 18 333 tertiary amines are retrieved. At this point the user ended the search by declining to enter a third SFC, and the DIS intersects the 893 and the 18 333 to produce a group of 175 compounds that contain both features. This file is then stored for future use.

The deficiencies of this approach abound. An inexperienced user, for example, finds it difficult to select from a code book fragments that will be optimum for a search, and in any event, no list could be long enough to cover even the most common possibilities—a computer system is capable of handling far more fragments than is practical for inclusion in a code book. The FPROBE search in the DIS (see below) can deal with some 7 million different atom-centered fragments; the list of predefined fragments in the DIS, on the other hand, is less than 1000.

Because of these difficulties and because, furthermore, the atom-by-atom search requires a query in the form of a structure, many systems, including the DIS, have added two more programs, one of which is used to enter a query structure, the other of which algorithmically creates therefrom the necessary fragments.

The input of query structures has been described elsewhere.[15] The structure is assembled with a group of simple commands, such as RING, CHAIN, and so on. Thus the successive commands, RING, AGROUP HO AT 1, and ALTBD 1 2 will create the query structure



which, at search time, will degenerate to phenol because hydrogens are implicit and all unidentified nodes, or atoms, which are denoted as numbers in the diagram, will be assumed by the programs to be carbon.

c. Atom-Centered Fragments. Once a query structure such as that above has been assembled, any atom-centered fragment from it can be used as the basis of a search. An atom-centered fragment is defined as a central atom, with all its first neighbors and the bonds joining it to the neighbors. Thus the atom 1 in the example requires an atom-centered fragment consisting of a carbon carrying two other carbons, to which it is attached by means of aromatic bonds, and one oxygen, the bond in this case being single. The DIS command FPROBE 1 will find all compounds in the Chemistry database which contain exactly that fragment.

```
OPTION? FPROBE 1
Type E to exit from all searches, T to proceed to next fragment
search.

Fragment # 1
           70*****1C. . . . .6C
                    .
                    .
                    .
                   2C
Required occurrences for hit: 1
  78 343 hits were found and stored in file 8 associated with
the CHEMISTRY database
```

Any atom whose identity is not specified will be assumed by default to represent carbon. A node in the structure can be defined to be any element, and it is also possible to allow an atom to be identified as any one of a group of elements. When specifying an atom, with the SATOM command, it is possible simply to enter a list of up to 105 elements.[16] Alternatively, one may use the predefined "superatoms" X (any halogen), M (any element except C, B, Si, N, P, As, Sb, O, S, Se, Te, H, F, Cl, Br, or I), or El (any element except C, H, D, or T).

d. Searching for Rings. The capability of searching for specific ring systems provides the DIS with a third screen in addition to predefined fragments and atom-centered fragments. There are relatively few acyclic structures in the database, and as a result, the ring search, RPROBE, is a very powerful and heavily used option. When a search is requested for the ring or set of rings in a query structure, the DIS discerns a variety of alternative ways of interpreting the question, and the correct one of these is established by querying the user.

The ring that is being sought may be fused to a second ring, thus a search for benzene would retrieve naphthalene. If this is not desired, it can be prohibited by the user. At the second level, one may wish or not to allow non-carbons in the ring beyond what has been specified. Such a relaxation would allow pyridine to be retrieved when the query structure was benzene. The third level concerns the location of specified non-carbons. Pyrimidine may or may not be an acceptable retrieval if the query structure was pyridazine, and the user must announce this. At the fourth level, ring substituents beyond those in the query structure may be allowed or disallowed. Finally, the search program will examine the nature of the first substituent atom and also the bonds between it and the ring atom. Both of these features must match what was specified in the query structure. The query structure can be modified, however, to identify these variables precisely or generically.

In the example shown below, the 2-halopyrazine ring is the object of the search. The user is offered and declines (1) inclusion of that structure in larger ring systems, (2) heteroelements beyond the two specified, and (3) changing either of the two nitrogens to some other non-carbon. At level 4, however, the user opts to allow additional substituents. The program notes that the query structure contains a substituent at position 2, that the substituent is a wild-card (atom 7, X = halogen), and that the bond to the ring was unspecified (A = any). The search is then carried out on these terms; 254 hits are found, and they are stored in temporary results file 9.

```
OPTION? RPROBE

                     7    N
                      ?  ? ?
                       ? ?  ?
                        C    C
                        ?    ?

                        ?    ?
                        C    C
                         ?  ?
                          ? ?
                           N
Multivalued atoms specified
Node Elements

7 X

(1) Allow inclusion in larger ring systems? (N/Y) (N)N
(2) Allow heteroelements at additional positions? (N/Y) (N)N
(3) Allow other element types at the position you show as
    heteroelements? (N/Y) (N)N
(4) Allow substituents at additional positions? (N/Y) (N)Y


                        Conditions of search
           Characteristics to be matched              Type of match

  Type of ring or nucleus                                 EXACT
  Heteroatoms at 1 4                                      EXACT
  Heteroatoms are N N                                     EXACT
  Substituents at 2                                       IMBED
  Substituents are *
  Subst bonds are A
  This ring/nucleus occurs in 254 compounds.

  254 hits were found and stored in file 9 associated with the
CHEMISTRY database
```

It should be noted that in the ring search the type of bonding in the ring is ignored. This decision was made in the early design of RPROBE[17] because it was felt that requiring this additional level in the search complicated it out of proportion to the benefits that were possible. As it is, the correct bonding can be retrieved by means of FPROBE, and any incorrect bonding is easily eliminated in the atom-by-atom search.

e. Atom-by-Atom Searching. The object of a substructure search is to find every compound in the database with a structure that includes imbedded in it the "query structure" of interest. A simple example of such a search would be the task of finding all compounds that contain a 7-hydroxyquinoline moiety. This example is shown below. The FPROBE search retrieves 78 343 phenols from the database, and the RPROBE search retrieves the 200 quinolines. If these two lists are intersected, the resulting list contains 189 compounds, including all the 7-hydroxyquinolines. It will also contain all compounds in which there are both a phenolic hydroxyl group and a fully or partially saturated quinoline ring system. The task of winnowing out the false drops is best done by the atom-by-atom search SUBSS. This will take each of the 189 candidate structures and attempt to map the query structure onto it. If the attempt fails, the candidate is discarded, and when the SUBSS is complete, only 160 7-hydroxyquinolines remain.

```
OPTION? FPROBE 1

Fragment # 1
           70*****1C.....6C
                   .
                   .
                   .
                   2C
Required occurrences for hit: 1
  78 343 hits were found and stored in file 8 associated with
the CHEMISTRY database

OPTION? RPROBE
          C??C
         ?    ?
        ?      ?
   C??C         C**O
  ?    ?       ?
 ?      ?     ?
C        C??C
 ?      ?
  ?    ?
   C??N

(1) Allow inclusion in larger ring sysetms? (N/Y) (N)N
(2) Allow heteroelements at additional positions? (N/Y) (N)N
(3) Allow other element types at the positions you show as
   heteroelements? (N/Y) (N)N
(4) Allow substituents at additional positions? (N/Y) (N)Y

                    Conditions of search
        Characteristics to be matched              Type of match

  Type of ring or nucleus                            EXACT
  Heteroatoms at 7                                   EXACT
  Heteroatoms are N                                  EXACT
  Substituents at 1                                  IMBED
  Substituents are O
  Subst bonds are CS
  This ring/nucleus occurs in 200 compounds.
  200 hits were found and stored in file 11 associated with
the CHEMISTRY database

OPTION? #10 and #11
  189 hits were found and stored in file 12 associated with
the CHEMISTRY database
OPTION? SUBSS 12
  Doing substructure search
  Type E to Exit
  File entry 20, Hits so far 14
  File entry 40, Hits so far 33
  File entry 60, Hits so far 49
  File entry 80, Hits so far 69
  File entry 100, Hits so far 88
  File entry 120, Hits so far 100
  File entry 140, Hits so far 118
  File entry 160, Hits so far 138
  File entry 180, Hits so far 152
  160 hits were found and stored in file 13 associated with
the CHEMISTRY database
```

The atom-by-atom search is authoritative and accurate; its only shortcoming is that it consumes a great deal of computer time—approximately 100 ms per candidate structure. It is this that makes the preliminary screens necessary, because they can usually remove more than 99.9% of the file from consideration in a very short time, so the number of compounds that must be submitted to the final search is thus kept to a reasonable level.

f. Automated Substructure Searching. The substructure search for 7-hydroxyquinolines that was described in the previous section used a specific atom as the object of an FPROBE search and then intersected those hits with the RPROBE retrievals. In this way a small file of candidates for atom-by-atom searching was generated.

This process relies to an extent upon the user's understanding that the hydroxyl-bearing carbon is a distinctive feature of the query structure and that it might prove to be discriminatory in a search. It is a matter of debate just how discriminatory it proved to be in this case when one reflects that imposition of the condition that a hydroxyl be present reduced the number of candidates only from 200 to 189. As a general matter, however, the DIS will, if the users wish, relieve them of the need to make decisions of this sort. The command SSEARCH will, unassisted, examine the query structure and then, if it is appropriate, perform the RPROBE search, allowing additional substitution or not, as the user wishes. Then it goes on to determine if the structure contains atom-centered fragments which could be usefully searched for. In this case, it searches for more than the hydroxyl-bearing carbon because it reduces the 200 RPROBE hits not to 189 as before, but to 160. Finally the substructure search through these 160 hits is carried out, and the same final answer is obtained.

```
OPTION? SSEARCH IMBED
          C??C
         ?    ?
        ?      ?
   C??C         C**O
  ?    ?       ?
 ?      ?     ?
C        C??C
 ?      ?
  ?    ?
   C??N

Ring search complete: 200 hits
Starting fragment search......
Fragment #3      160 hits
Fragment search complete: 160 hits
Starting substructure search......
Search complete
  160 hits were found and stored in file 14 associated with
the CHEMISTRY database
```

The SSEARCH option has two benefits. It relieves the novice user of possibly difficult decisions, and it attempts, with some success, to exploit the most efficient strategy for the query structure in question. It does not search for all atom-centered fragments, but only for those whose frequency of occurrence in the database is low enough that decent economies in the atom-by-atom search will result.

MANIPULATION OF RESULTS FILES

A. Boolean Logic. The temporary files that result from DIS searches may represent final results, in which case they may be printed, as is described in the next section. Frequently, however, the results files must be combined with one another in various ways so as to produce the final results.

Temporary files in the DIS can be combined with Boolean AND, OR, or NOT operators, with commands such as

```
OPTION? #2 AND #14
```

which causes the intersection (logical AND) of temporary results files 2 and 14. The # sign indicates to the DIS that the "2" and the "14" are not search terms.

Boolean expressions may be quite complicated

```
OPTION? ((#3 AND #5) OR #12) NOT #15
```

and, more significantly, Boolean operators may be imbedded in search expressions

```
OPTION? (#7 OR (NSC/>350000 AND CAMT/>1
GM)) AND MW/200 TO 300
```

This provides the user with a powerful additional means for narrowing down the results of a search or focusing quickly upon the correct results of a search.

Boolean operations are also important for cross database retrieval. Different databases in the DIS are keyed according to different identifiers. The CHEMISTRY database is keyed on NSC numbers, which are numbers consecutively assigned to new compounds. The INVENTORY database is keyed on NSC/Sample Numbers, which are assigned serially to different samples of the same compound. The SHIPPING database is keyed on NSC/Sample/Shipping List Numbers, NAMECODES on Name Codes, and so on. Results files retain the order of the databases from which they are obtained. If obtained from the INVENTORY database, the entries will be keyed by sample number. If from CHEMISTRY, they will be keyed by NSC numbers, and in all cases, their origin is indicated when such files are created. When two files with different origins are intersected, a conversion is automatically carried out, the second file to be cited being converted to the format of the first, and the intersection executed. There is also a CONVERT command that does such conversions without intersection.

B. Output Commands. Temporary results files can be viewed by means of either the TYPE or the PRINT commands. The TYPE command causes the data to appear at the user's terminal, while with the PRINT command a remote device, usually a printer, is used. The syntax of both of these commands is

```
TYPE file number/format/entry number(s)
```

Thus the command

```
OPTION?  TYPE 13/CHEMISTRY/5-17
```

will cause the complete Chemistry records for entries 5 through 17 of file 13 to be listed at the terminal. It is common for a user to define a format during the session using the FORMAT command, and an example of this is

```
OPTION?  FORMAT NSC CNAM MOLF MW STRC
```

Once a format like this is defined, the format argument can be omitted from the TYPE or PRINT commands; this group of fields will be output. The TYPE command above becomes

```
OPTION?  TYPE 13//5-17
```

The PRINT command operates in just the same way as the TYPE command except that the printing is done at a remote high-speed printer. Two options offered by the PRINT command are the opportunity to delay the printing until the evening, when much lower charges pertain, and second, to retain a copy of the print file on disk for future use.

Beneath these two simple commands lies some quite elaborate software, whose task is to learn what output device is to be used and then to respond accordingly. Every DIS user is associated with a "user profile", which contains, among other things, the identity of their terminal and of the printer of their choice. Both these parameters can be superseded by the user, but the DIS knows, in one way or another, what devices are involved. Given a TYPE command, the DIS must examine the format currently in force to determine if any graphics are called for. If structure records are to be typed, then the nature of the user's terminal is ascertained and the appropriate structure records are retrieved and made available to the TYPE command. For PRINTing, a similar step is necessary. Here the DIS must learn which high-speed printer is to be used and then must collect the correct graphics data and pass it, with any text, to that printer.

UTILITY PROGRAMS

The DIS makes available to the user a variety of utility commands. Many of these, of course, are used routinely by other systems as well. Among those of interest in the Chemistry area are the following:

An EXPAND command is useful for finding index entries. It shows sections of the index both preceding and following the presented index term.

The EXIST command is provided to allow the retrieval of all records in which there is data of any sort in a specific field. The entry

```
OPTION?  EXIST/STXT
  45 743 hits are found and stored in file 15 associated with
the CHEMISTRY database
```

retrieves the 45 743 compounds that have stereochemical information in their STXT field.

Elision, which allows one to bypass a variable number of characters, is a useful feature in name searches. It is specified by means of a question mark (?) inserted in the attribute. Elision is commonly used to accommodate the suffixes of a root word. For example, the term "meth?" will match methyl, methoxy, methinyl, methenyl, etc. In chemical names in particular, elision is useful in the middle of a term, and the DIS permits such usage. For example, "butyl?aldehyde" will be matched by names such as butylethylacetaldehyde, butylaldehyde, butylphenoxyacetaldehyde, and so on. Here, the "?" is understood by the DIS to stand for a character string of any length, including zero.

As one uses the DIS, temporary results files accumulate, and it can be troublesome to keep track of their origin and content. The FILES command will provide a directory of all one's files at any time

```
OPTION?  FILES

  file      size     sorted      database
   1        1293      YES        CHEMISTRY
   2           6      YES        CHEMISTRY
   3          17      YES        ORDER
   4         275      YES        INVENTORY
   5        1065      YES        SHIPPING
   6           1      YES        BIOTEST
```

and further detail about any specific file can be obtained with the HISTORY command:

```
OPTION?  HISTORY 12
  Created on 14-Nov-85 at 17:46:21 by the command:
  NSC/165743
```

Finally, the RECAP command will play back one's DIS commands, and this may be very helpful in relocating one's place in the overall session:

```
OPTION?  RECAP
  History of this Session
  DACQ/JAN-82
  NSC/>370000
  #1 AND #2
  NSC/165733
  EXIST/ACAT
  RECAP
```

Results files can be deleted, restored, or permanently saved. For files that are permanently saved, a USE command will produce them at a later DIS session. It is also possible to save permanently one's query structure or one or more format statements.

FILE ORGANIZATION

As has been described above, two distinctly different database types are maintained in the DIS Chemistry module. All the nonstructural chemistry data exist as a set of files under TDRS, the database manager that serves the remainder of the DIS. The chemical structural data, on the other hand, are stored as a set of special structure search files derived from the Structure and Nomenclature Search System (SANSS) of the CIS.[3] Because of the additional software required to generate and search them, these special SANSS database files have been retained only where absolutely necessary. Thus, all searchable textual and routine numeric data are maintained in Chemistry's TDRS databases, and all displayable data fields, including graphical structure images and connection tables, are also stored in the TDRS databases.

The dividing line between TDRS and SANSS was not established on merely a structure vs. nonstructure level. Rather, it was arrived at by examining the functional complexity of each of the required search capabilities. A number of the search capabilities that were traditionally considered to be a part of SANSS have thus been turned over to TDRS. In general, the SANSS search software and database structures that have been retained intact are those that process queries which implicitly contain multiple nested criteria. These are the RPROBE, FPROBE, SPROBE, RCOUNT, and MF searches. Molecular formula searching in the DIS is fairly complex because, in addition to full molecular formulas, the molecular formulas of specific addends can be retrieved. Further, any molecular formula may be specified only partially, and element counts can be ranged. In the DIS, these five searches are carried out by traversing inverted and heavily indexed hierarchical database files. This type of file organization, while not novel,[17] is particularly appropriate for chemical structure searching and provides for powerful and rapid retrieval.

Four unique search file types are maintained in the Chemistry module: one each for RPROBE, FPROBE, and SPROBE, and a fourth that contains information for both RCOUNT and MF. Like TDRS databases, each of these special databases consists of one or more separate incremental files. Most of the databases in the DIS consist of a "core" and some number of incremental additions. The increments arise from the frequent updates that are carried out, a procedure which is described in detail in part 6 of this series. During the course of a search, the core and increments which comprise the database are searched consecutively and individually. Each search produces a list of hits, and the final results file is obtained by merging the responding lists from each.

Every searchable file consists of a directory block that defines the structure of the file, several hierarchical levels, and one or more sections containing sorted variable-length lists of accession numbers. Below the directory, entries in each of the hierarchical levels minimally contain some portion of the complete search key and a pointer to an entry in the next lower level which, in turn, contains another part of the key and yet another pointer. At the bottom of the hierarchy, the pointers refer instead to a list of responding accession numbers which can be extracted for inclusion in the search results list.

The hierarchical organization for each of the four special Chemistry databases is as follows:

RPROBE—ring system search
Level 0, index to ring/nucleus types
Level 1, ring/nucleus type
Level 2, heteroatom element types
Level 3, heteroatom positions
Level 4, substituent positions
Level 5, substituent bond types
Level 6, substituent element types

FPROBE—atom-centered fragment search
Level 1, fragment size
Level 2, central atom types
Levels 3, 5, 7, and 9, first, second, third, and fourth neighbor element types
Levels 4, 6, 8, and 10, first, second, third, and fourth neighbor bond types

SPROBE—structural feature codes search
Level 1, code types
Level 2, code qualifiers
Level 3, occurrence counts

RCOUNT and MF—ring count and molecular formula searches
Level 1, key types (rings or molecular formulas)
Level 2, ring sizes and element types
Level 3, occurrence counts

In any of these searches, a complete vertical path through an increment thus defines the set of compounds in the increment that are characterized by a unique combination of features. The "combining" (a Boolean AND operation) is done implicitly and rapidly as the search proceeds from one level to the next. For example, in the RPROBE illustration given earlier for 7-hydroxyquinoline, only those compounds containing two fused 6-membered rings AND a single heteroelement at position 1 AND a nitrogen at that same position AND all three conditions applied simultaneously to the substituent are retrieved. Similar "exact" searches could also be performed by TDRS, but would require six separate searches, each producing unmanageably large results files, which would then have to be intersected in order to complete an equivalent query. To distinguish the two retrieval systems further, the SANSS search algorithms permit multiple simultaneous traversals of the hierarchy in those cases where the query structures have been defined in an ambiguous manner or when the search criteria have been otherwise relaxed. TDRS possesses no similar capability.

INTERFACING HOST AND MICROPROCESSORS

The input of chemical structures by means of microprocessors (PC's) constitutes an interesting application of hierarchically structured distributed processing. The system is distributed because the microprocessor works in conjunction with the host to which it is connected. Each has its function: The host maintains and updates the files to which the PC provides input. The host determines whether input structures are new or not; it maintains the indexes and the organization required to process queries entered from the PC. The function of the PC is to facilitate the input of structures, to render that task as effortless as possible, for both the casual and the full-time user. To achieve this it must manage, in real time, the interactions with the user, issue prompts, process the responses, and be ready with status and error messages. As the input of chemical structures is a complex task, the programs designed to support and facilitate it are rather extensive. The PC may have to capture the movement of a touch pen and trace, virtually simultaneously, a corresponding path on the screen. Even where the structure is entered solely by keystrokes, as in the adaptation described above, of an IBM PC for input, the effort to produce as much of the structure as possible with the least amount of input, as quickly as possible, often triggering a sizable display with a single keystroke, involves a considerable amount of internal processing. Thus both the IBM PC and the Victor 9000 require auxiliary boards and other accessories, as well as all the internal memory they are capable of carrying.

Thus, in developing programs to support chemical input, it is not only the extensive programming of the PC that must be taken into consideration. A judicious choice must be made as to what processes are to be handled by the host and what processes are to be handled by the PC. It is clear that the PC should perform display management and that the host should maintain the database. But which system is to correct the connectivity tables for tautomerism and resonance? The DIS chose to let the host perform such significant operations. Having two separate input devices (the Victor and the IBM PC), should their outputs be identical or device-dependent? In the DIS, it is the latter. Finally, should the development of the PC programs, an extensive task, be done on the PC or on the host? The DIS opted for the former.

The distributed system encompasses not only host and microprocessor but also the user. Although more difficult to formulate, human factors must be considered. This was done by inviting experts in this field to examine the input operation and further by extensively testing the system under operational conditions, making written records of all complaints or suggestions voiced by the operators and following up on each one of these. In this manner, we believe to have ensured efficient cooperation among host, PC, and user, allowing each to do what he, she, or it does best. That is what makes this application of distributed processing so interesting.

APPENDIX A

The Chemistry module of the DIS is a direct descendant of the Structure and Nomenclature Search System (SANSS), which is the central compound index of the NIH–EPA Chemical Information System (CIS). While SANSS is a very powerful search system that meets the needs of the CIS, it was necessary to modify it for use in the DIS. The major modifications which have been made are described below.

1. The DIS has the capability for the online registration of new compounds—something that is not possible in SANSS. This uses a modified version of the IDENT search software from the SANSS but had to be encompassed within a general update capability which, in a dialogue with the user, solicits for structure and associated information.

2. Because of the need for frequent database updating, a set of semiautomatic file updating programs was developed and installed in the DIS. This is described in more detail in part 6 of this series.

3. The substructure search techniques used by the SANSS were expanded, both to simplify their use and to increase their specificity. The ring search, RPROBE, used to characterize ring compounds in the database on the basis of node connections within rings, the position of heteroatoms in rings, the specific heteroatoms permitted at those positions, and the position of substituents to the ring systems, was significantly enhanced. In the DIS, in addition to these four levels of specification, RPROBE now tests the nature of ring substituents and the bonding to those substituents. Charge and abnormal mass and multiple atom, or "wildcard", designations are now permitted.

4. Automatic substructure searching is allowed in the DIS. The user can now trade-off the efficiency of specifying search strategy for the convenience of merely submitting the fragment or fragments to be used in a substructure search. The automatic search seeks the optimal route to a solution by reviewing the likely statistics of different search paths.

5. The routines for atom-by-atom searching were streamlined and speeded up by approximately an order of magnitude. Various other routines in SANSS were similarly improved as they were incorporated into the DIS. Boolean intersections, for example, were speeded up, also by an order of magnitude.

6. In SANSS, output diagrams for chemical structures were generated from the connection table. These diagrams were not always acceptable to the chemist, one reason being that any implied stereochemistry would be lost. In the DIS, the coordinates obtained during the input of the structure are retained with the connection table, so that the output from the system reproduces the appearance of the structures at input.

7. The full structure search used in the DIS differs from that in SANSS in that proton counts, required by SANSS, are no longer required. The proton counts are used by SANSS to guard against accidental collisions during the IDENT search. The same result is achieved in the DIS by an atom-by-atom comparison of the query structure and all potential hits. This is a reliable approach and no longer requires the user to undertake the troublesome task of ascertaining the proton count of the query structure.

8. The presence in the DIS of other major databases, such as Biology, which are linked to Chemistry, required a great deal of work in the area of database interfacing.

9. Query structures in the DIS may be specified and searched for at a greater level of detail: charged atoms, abnormal mass, textual stereo descriptors, coordination (covalent) bonds, and isotopes are all countenanced by the DIS. Furthermore, although not searchable, stereo bonds are shown, as appropriate, at both input and output.

APPENDIX B

The format for data transmission of chemical structures from the microcomputer consists of three record types. These types are the header record, the trailer record, and the data record. All data are transmitted by using the communications protocols developed by Fein-Marquart Associates. All data are transmitted by using standard ASCII format. The format for the header record is as follows:

HEADER, the header value is always 0.

NNODES, the number of nodes which will follow. This value does not include the header or the trailer record.

There are five types of data records. These records can describe an atom, a nonlocalized charge, a "D" type structure ("D" structures are those with an undetermined bond site), a type "M" structure (a compound with an undetermined structure), and dot-disconnect marker positioning. Each of these types will be described.

The most common case is the atom record. This record will contain the element, the bond, the charge, and the associated hydrogens for the node. It will also contain the picture information and the connections so that the structure can be reproduced exactly. The following is the standard format for transmitting atoms:

NODE, the ordinal number for the node currently being transferred.

TYPE, the type flag for the current node. In most cases, this field contains the element symbol for the current node. Cyclical carbons are represented by lower case "c". These carbons are represented graphically as Luhn dots. The field may also contain "M" structure indicators. The "M" structure indicates an incompletely defined fragment or structure. These types have an "M" in the first position of the field followed by the number of the "M" type structure. For example, if a compound has three "M" fragments, the third would be identified by "M3" in the type field. The molecular formulas for the "M" fragments are found in the trailer record. This field may also contain "D" structure indicators. The "D" structure fragment indicates that the bonding site for the fragment is undetermined. Like the "M" structure, the "D" also has a number associated with it. All components of the "D" structure are contained in other data records. There are minor variations in the "D" structure records. See the "D" record format for a complete explanation. Those atoms that are possible sites of attachments for "D" fragments are identified by the information contained in their CHG field.

XCOOR, the *x* coordinate of the atom. This and YCOOR are used so that the graphic depiction of the structure can be reproduced.

YCOOR, the *y* coordinate of the current atom.

CHG, the charge of the current atom. If this node is the potential site of attachment, the value of this field will be 100 plus the number of the "D" fragment for which it is a possible point of bonding. For example, if the current node is a potential bond site for fragment "D1", the value of this field would be 101.

RELCHGP, the relative placement of the charge or the potential attachment marker to the atom. The position field is used for display purposes only. The codes for the position are the same as those used with RLHYDP (below).

NHYDS, the number of hydrogens connected with the current atom.

RLHYDP, the position of the hydrogens relative to the atom. This value is only used for display purposes. The code values for the position of the hydrogens are

1, above the atom
2, up and to the right of the atom
3, to the right of the atom
4, below the atom to the right
5, below the atom
6, to the lower left of the atom
7, to the left of the atom
8, above the atom to the left

MASS, the mass of the current atom. The variable is used to specify mass when the atom has an abnormal mass.

NCON, the number of connections for the current atom. For each connection (NCON) the sequence CA, the node to which the current atom is connected, CB, normal bond type for the connection, applies. The bond types are

1, single
2, double
3, triple
4, stereo up (represented by a dotted line)
5, stereo with wedge toward connected atom
6, stereo with wedge away from connected atom
7, stereo unknown

"D" format records indicate that the bonding site for the following fragment cannot be determined. The format for this record is

NODE, the ordinal number of the node currently being transferred.

TYPE, the letter "D" followed by the number of the "D" fragment.

XCOOR, the *X* coordinate for the start of the "D" fragment.

YCOOR, the *Y* coordinate for the start of the "D" fragment.

NCON, the number of connections for the current "D" fragment. For each connection (NCON) the sequence CA, the node to which the current "D" is connected, CB, normal bond type for the connection, applies. The bond types are

1, single
2, double
3, triple
4, stereo up (represented by a dotted line)
5, stereo with wedge toward connected atom
6, stereo with wedge away from connected atom
7, stereo unknown

Nonlocalized charges are transmitted as their own record. These records can be identified by either a "+" or a "−" in the first byte of the field following the NODE number field. The NCHG field contains the "+" or "−" followed by the value of the charge. The nonlocal charge record will also contain the *X* and *Y* coordinates of the charge. The format of the NONLOCAL CHARGE records is

NODE, the ordinal number of the current record.
NCHG, the nonlocalized charge value.
XCOOR, the *X* coordinate of the charge.
YCOOR, the *Y* coordinate of the charge.

Dot-disconnected position records contain information concerning the *x* and *y* coordinates of the dot-disconnect marker (*) and any multiplier value pertaining to that structure.
NNODE, the ordinal number of the current record.
TYPE, the character "*" followed by a blank.
XCOOR, the *X* coordinate of the "*".
YCOOR, the *Y* coordinate of the "*".
MULT, any integer multiplier other than 1, or if the multiplier is fractional, the numerator of the multiplier.
DIVI, if the multiplier is fractional, the divisor of the multiplier.

The final record type is the trailer record. The trailer record may contain "M" definition formulas. The format of the trailer record is

NNODE, the value of NNODE for a trailer record is always −1.
MLEN(*i*), for *i* = 1–5, the length of the next type "M" molecular formula. If there is no "M", for the current *i* the MLEN(*i*) field will be set to 0. All MLEN(*i*) fields are delimited by commas.
MMFORM(*i*), the molecular formula of the next "M" fragment. Embedded within this field and delimited by vertical bars are the *x* and *y* coordinates used for the printing of the "M" fragment. All MMFORM(*i*) fields are delimited by commas.

The following is an example of data transmission:

structure		data
		⟨soh⟩0,10,⟨cd⟩⟨lf⟩
*23/9HCl		⟨soh⟩1,* ,34,8,23,9,⟨cd⟩⟨lf⟩
	−2	⟨soh⟩2,C ,30,12,0,0,2,7,0,1,2,2,⟨cd⟩⟨lf⟩
“OH		⟨soh⟩3,C ,32,12,0,0,0,0,0,1,3,1,4,1,4,1,⟨cd⟩⟨lf⟩
		⟨soh⟩4,0 ,32,10,101,7,1,3,0,0,⟨cd⟩⟨lf⟩
M1—H2C═C—H		⟨soh⟩5,H ,34,12,0,0,0,0,0,0,⟨cd⟩⟨lf⟩
		⟨soh⟩6,M1,29,12,0,0,0,0,0,⟨cd⟩⟨lf⟩
		⟨soh⟩7,C1,36,8,0,0,1,7,0,0,⟨cd⟩⟨lf⟩
	*2M1:N4C2	⟨soh⟩8,D1,44,16,1,1,9,1,⟨cd⟩⟨lf⟩
	D1–NH3	⟨soh⟩9,N ,47,16,2,6,3,3,0,0,⟨cd⟩⟨lf⟩
	+2	⟨soh⟩10,−2,47,9,⟨cd⟩⟨lf⟩
		⟨soh⟩−1,15,44|15|2|M1:N4C2,0,,0,,0,,0,,⟨cd⟩⟨lf⟩

REFERENCES AND NOTES

(1) The DIS source code carries a minimum commentary level of 100%. Thus in 100 lines of source code, at least 50 lines are comment lines.

The number of lines of program code in the Chemistry module, excluding comments, is on the order of 150 000.

(2) Chemical Abstracts Service (Columbus, OH) and VSE Corp. (Alexandria, VA) participated in separate contracts with NCI. The contract with CAS ended in 1985; VSE continues to provide support to NCI in the biological data processing area.

(3) Heller, S. R.; Milne, G. W. A.; Feldmann, R. J. "A Computer-Based Chemical Information System". *Science (Washington, D.C.)* **1977,** *195,* 253–259. Milne, G. W. A.; Heller, S. R.; Fein, A. E.; Frees, E.; Marquart, R.; McGill, J. A.; Miller, J. A.; Spiers, D. S. "The NIH–EPA Structure and Nomenclature Search System". *J. Chem. Inf. Comput. Sci.* **1978,** *18,* 181–186. In 1984, the NIH–EPA Chemical Information System, which includes SANSS, was transferred into the private sector. It is currently offered by two vendors (CIS, Inc., Baltimore, MD, and Information Consultants Inc., Washington, DC).

(4) The first operational version of SANSS was completed in 1976 and made available for public use as a part of the NIH–EPA Chemical Information System. It has been used continuously since then by well over 1000 fee-paying users.

(5) The question of time per structure is discussed in part 2 of this series of papers. With the touch-pen method, a typical structure can be entered in under 3 min. The keyboard-entry program reduces this to less than 1 min.

(6) Versatec, Inc., a Xerox Co., is located at 2805 Bowers Avenue, Santa Clara, CA 95051.

(7) Friedman, N. K. "Japanese Word Processing: Interfacing with the Inscrutable". *Abacus* **1986,** *3,* 34–42.

(8) Dittmar, P. G.; Mockus, P.; Couvreur, K. M. "An Algorithmic Computer Graphics Program for Generating Chemical Structure Diagrams". *J. Chem. Inf. Comput. Sci.* **1977,** *17,* 186–192.

(9) Milne, G. W. A.; Feldman, A.; Miller, J. A.; Daly, G. P.; Hammel, M. J. "The NCI Drug Information System. 2. The DIS Pre-Registry". *J. Chem. Inf. Comput. Sci.,* preceding paper in this issue.

(10) Systems offering nomenclature searching with no structure searching include Lockheed's DIALOG, NLM's CHEMLINE, and Burrough's SDC/ORBIT and BRS. Both nomenclature and structure searching are supported by SANSS and by CAS ONLINE. The service provided by DARC/QUESTEL is primarily a structure search and does not include nomenclature searching.

(11) Richman, S.; Hazard, G. F.; Kailkow, A. K. "The Drug Research and Development Chemical Information System of NCI's Developmental Therapeutics Program". In *Retrieval of Medicinal Chemical Information*; Howe, W. J., Milne, M. M., Pennell, A. F., Eds.; ACS Symposium Series 84; American Chemical Society: Washington, DC, 1978; pp 200–221.

(12) The DIS allows retrievals based upon the full molecular weight (MW), which contains all the addend molecular weights. Retrievals are also allowed on the molecular weight of any particular addend (MWAD).

(13) A full and precise match at the addend level is required in the IDENT search. Thus aniline will retrieve aniline, aniline hydrochloride, and all other salts of aniline. An IDENT search for aniline butyrate will retrieve only aniline butyrate, but if the search is done with just the butyrate, all butyrates, including aniline butyrate, will be retrieved. The version of IDENT used in the DIS Pre-Registry is more flexible and handles addend discrepancies slightly differently. This is described in more detail in the preceding paper in this series.

(14) Ray, L. C.; Kirsch, R. A. "Finding Chemical Records by Digital Computers". *Science (Washington, D.C.)* **1957,** *126,* 814–819.

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

(16) Seventy-eight different elements are represented at least once in the NCI database. The elements that are not represented are the five noble gases, one lanthanide (promethium), and thirteen actinides (actinium, protactinium, and all elements beyond uranium). This reflects the eclectic approach that has been adopted in the search for anticancer drugs.

(17) Feldmann, R. J.; Heller, S. R. "An Application of Interactive Graphics—the Nested Retrieval of Chemical Structures". *J. Chem. Doc.* **1972,** *12,* 48–54.

The NCI Drug Information System. 4. Inventory and Shipping Modules

G. W. A. MILNE*

Information Technology Branch, Division of Cancer Treatment, National Cancer Institute, Bethesda, Maryland 20205

J. A. MILLER and J. R. HOOVER

Fein-Marquart Associates, Baltimore, Maryland 21212

Received April 21, 1986

The Inventory/Shipping package of the NCI Drug Information System (DIS) is designed to support all inventory and shipping operations associated with the testing by the NCI of large numbers of chemicals for anticancer activity. Two major databases, an Inventory database and a Shipping History database, contain all of the data associated with these operations. Software that supports the operations in an online interactive manner also provides for the accessing and updating of these databases as necessary. Special hardware in the form of barcode reader/printers and digital balances is also interfaced to the system to improve the efficiency of the operations.

INTRODUCTION

In the first part of this series, the general structure of the NCI Developmental Therapeutics Program (DTP), whereby large numbers of chemicals are tested for anticancer activity, was described. As part of this program, a physical inventory of those chemicals is maintained,[1] and shipments of the chemicals are made to screening laboratories[2] and other recipients[3] as necessary in order to perform specified series of tests on the chemicals.

The NCI Drug Information System (DIS), also described in general in the first part of this series, is designed to support the various inventory and shipping operations performed in the course of day-to-day DTP project activities. Support of these operations is primarily interactive, and this support extends to automated laboratory stations, where material is subdivided, weighed, and packaged for shipment. Two major databases are also supported as part of the DIS Inventory/Shipping module, and these provide Inventory and Shipping History information, respectively, on the chemicals tested under the DTP project.

OVERVIEW

The inventory and shipping operations performed in the course of DTP project activities, all of which are directly supported by the DIS, are illustrated schematically in Figure 1. The major steps in these operations are as follows.

Receipt of Chemicals. All chemicals ordered for screening are received by the NCI acquisitions contractor, who indicates the receipt of the material to the DIS and registers the material with respect to the Chemistry database. New compounds are assigned new NSC Numbers, and additional materials for existing compounds are assigned new sample numbers under those NSC Numbers. This contractor also orders and receives refills of chemicals, as requested by NCI. All received ma-

0095-2338/86/1626-0179$01.50/0 © 1986 American Chemical Society