Pure & Appl. Chem., Vol. 49, pp. 1845-1853. Pergamon Press, 1977. Printed in Great Britain.

SYSTEM FOR CHEMICAL RETRIEVAL

Janet Ash and E. Hyde

Imperial Chemical Industries Limited, Mereside, Alderley Park, Macclesfield, Cheshire.

Abstract - This paper reviews the traditional and modern computer methods for recording chemical compounds. Details are also given of the atom connectivity method of recording compounds, and also line formula notations. The value and limitations of both methods are described, and finally examples of systems based on both atom connectivity and notations are given. The ICI CROSSBOW system is used as an example system based on both notation and connectivity records, and the CSSS developed at NIH, Bethesda is given as a method of an interactive computer system which is orientated towards use by the scientist.

The volume of information in the published literature and that generated internally in research organisations has led to the development of large data banks, and the need for computer systems for retrieval purposes. The information contained in these data banks can be divided into four main classes:

(a) Textual
(b) Results of Compound Testing
(c) Physical Constants
(d) Structural

The first type, textual, has been dealt with earlier. In this area of textual information chemistry differs little from other disciplines. The words used to describe compound names, details of compound syntheses, uses and experimental conditions can be retrieved by conventional information retrieval techniques. Extraction of words corresponding to a user's interests can lead to retrieval of useful information, but the imprecise nature of language means that text searching is unsuitable for retrieval of specific data.

The second type of information in chemistry is that of experimental observation where chemical compounds are submitted to a series of tests, for example, for toxicology, biological activity or clinical trials. The results of these tests are given as a series of opinions regarding what has taken place. Test results such as these obviously do not fit into a conventional text system and although these results are the key to the development of a useful product, there are no major publicly-available systems for handling such information.

The recording and retrieving of physical constants, like the results of compound testing, can not be handled in the same way as text, and although the nature of the data is precise and easy to record, there is a lack of publicly-available systems in this area also. One reason for this may be the limited usefulness of such a system. Retrieval of constants is a straightforward look-up procedure and the overall advantage of a sophisticated computer system over manual methods is only slight.

However, the move to systems based on structure analysis is changing this situation, and data bases of physical constants are being assembled.

It is in the area of structural representation that the use of computer systems for creating, searching and manipulating the data bases can be of particular value to the chemist.

The four principle methods of representing chemical structures in the literature are as follows:

(a) Nomenclature
(b) Fragmentation
(c) Linear Notations
(d) Connection Tables

The method chosen by the producer of the data bank greatly influences the design of the computer system for handling the structures. Before considering further the requirements of systems and the facilities provided it is necessary to consider the four methods of structure representation in some detail.

METHODS OF STRUCTURAL REPRESENTATION AND THEIR USE IN COMPUTER SYSTEMS

Nomenclature

The two-dimensional structural diagram is the most explicit method of describing chemical structures, but because of the problems of communicating diagrams in any other than a visual media, numerous other methods of structural representation have been developed. The original complement to the structure diagram, nomenclature, proved satisfactory until an individual chemist became unable to keep track of all the compounds produced. Systematic names were devised but the rules for producing a unique name for all compounds being prepared at the present time are now so complex that it is almost impossible for one person to be familiar with the rules. This leads to inconsistencies in naming compounds and consequent loss of information when searching for specific compounds. In addition, nomenclature is not suitable for use in computer systems which require logical arrangement and conciseness. For this reason, nomenclature as a form of structural representation in computer systems will not be considered further.

Fragmentation

With the increase in the number of compounds, the demand for specific compounds decreased and questions were directed to classes of compounds and partial structures. Searching for structural fragments using nomenclature in a manual or computer-based system is very difficult and the documentalist turned to methods of fragmenting a structure and assigning several 'names' to one compound. Many different fragment codes have been developed, but in each case a list of pre-determined structural units is devised and the compound is coded by recording all the structural features which are present from the total list of fragments. The structural units can be described in terms of names, such as phenyl, steroid, sulphate, or in numerical terms where each number corresponds to a different structural feature, but whichever method is used to record the units, they can be retrieved using the same techniques employed in a text-based system. Fragment codes were thus widely adopted as a method of storing chemical structures as they could be incorporated into existing systems. However, in a fragment code only the occurrence of structural features is recorded and not their relative positions and it is not possible to reconstruct the total structure from a knowledge of the coded fragments. A fragment code is thus only a partial structural representation and cannot be used for searches for specific structures or for substructures that do not form part of the code itself. The greatest value of the fragment code lies in its use as a screen to enable a large part of a file to be eliminated prior to a detailed search. The molecular formula of a compound can be used as a screen when manually searching an index for a specific structure, but the fragment code is able to provide a much more discriminating screen prior to a substructure search which may contain only the most frequently occurring elements.

A file of compounds coded by fragmentation methods can be used in a manual or computer-based system. Originally, fragment code systems used punched cards, optical coincidence cards, edge-notched cards or other similar devices. Using punched cards, each hole in the card corresponded to a particular position and the cards were searched for a given fragment or combination of fragments present in the same compound. The cards retrieved usually contained a record of the two-dimensional structural diagram, the name of the structure and full bibliographical details. These punched card systems could be directly converted to a computer-based system and searched for the presence, or absence, of specified fragments.

Search programs in fragment code systems need to provide for Boolean logic in the expression of the search question. For example, consider a search for:

Fragment A AND (Fragment B OR Fragment C)

but NOT Fragment D

Such a search would eliminate all compounds which contained fragment D and only select those compounds which contained both fragment A and fragment B and those which contained fragment A and fragment C. Some fragement codes also allow for the specification of the number of fragments of a given type present in the compound, for example, the search question may include the fact that the required structures must contain at least two NO_2 groups.

Because it is not possible in any way to reconstruct the structural diagram from a knowledge of the fragments, output from a fragment search can only be in terms of the information recorded on the file, which is normally the full bibliographic details and the chemical name.

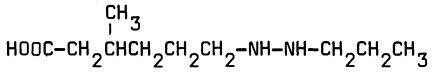
Notations

Since their inception, fragment codes have become more sophisticated and certain codes, such as the GREMAS code used at IDC[1], have been enhanced by the provision of links between

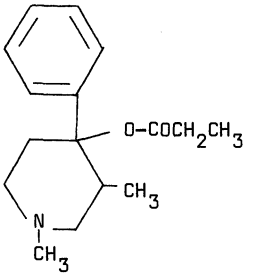
fragments. However, fragment codes remain a method of pre-classification of the structures. Even these more elaborate fragment systems do not preclude the need for a total structural representation for a large number of substructure searches and for any detailed analysis of files of structures where post-classification techniques are required.

Specialised chemical notations have been developed to describe structures both uniquely and unambiguously in a concise string of symbols. Each symbol represents an atom and its associated bonds or a group of chemically-related atoms and bonds. Ordering rules are used to ensure that a unique notation is obtained for each structure. Notations, although less meaningful to a chemist than nomenclature, can be used in a manual permuted index and are an efficient method of input of structures into a computer-based system.

One of the original notations was developed by Dr. G. M. Dyson (1,2). The rules used for coding compounds were similar to those used for assigning systematic names and seniority was given to rings and longest chains. Examples of the Dyson notations for two compounds are shown below:



Dyson notation: $C_6C,3.Q_1:6/N_2:2/C_3$

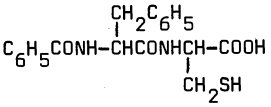


Dyson notation:

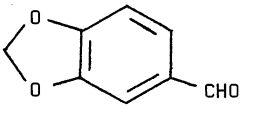
A6ZN,C,1,3.(XC_3)4:4/B6

It was the Dyson notation that was adopted by IUPAC (3), but at present the only main user of this notation is Shell Research, Sittingbourne (4). The most widely used notation is the WLN designed by W. J. Wiswesser (5). In this notation prominence is given to all non-benzene ring systems and the compound is coded by taking a path through the molecule, the direction of which is governed by ordering rules, and symbols are assigned to each functional group in turn.

Examples of the Wiswesser Line-Formula Notation are shown below:



WLN: SH1YVQMVY1R&MVR



WLN: T56 BO DO CHJ GVH

When using notations for input to a system, human errors in coding and key-punching are inevitable. A satisfactory system must therefore provide for validation of the input notations (6). Simple syntactical checks on a notation can be made, but validation is usually performed by a molecular formula checker program in which an input molecular formula is compared with one generated from the notation. Such a checker program will detect most errors, but cannot detect errors in coding of position of substituents on a ring, for example.

Notations are a total structural representation, which contain a degree of chemical significance in the symbols used and there is therefore considerable scope for searching and manipulating a file of notations. The symbols themselves are fragments with the additional advantage that they are linked according to the relationship within the molecule. Slightly modified conventional KWIC programs can be used to obtain computer listings of notations. When the symbols of the notation are used as index points related compounds appear grouped together in the same manner as an inverted file of chemical fragments. These permuted indexes for manual consultation are valuable for substructure searching if the notation files are not too large, but when the number of compounds on the file exceeds about 100,000 the sheer volume of output from a permuted listing limits its usefulness and presents

problems with updating.

File screening and many substructure searches can be performed by a string search of notations. String search programs, which are available as standard software from many of the computer manufacturers, include the usual AND, OR and NOT logic, with additional facilities such as FOLLOWED BY and IGNORE. The IGNORE logic is particularly useful when searching the WLN in which the position of substituents on a ring is coded by a locant letter preceded by a space. Many of the letters have alternative meanings and to avoid false drops it is necessary to exclude any letters preceded by a space.

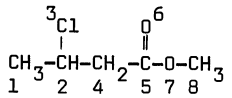
A notation is frequently unintelligible to a chemist, and a list of notations is unacceptable as output from a notation search. Structural diagrams can be provided as output in one of two ways. The Co-ordinates of the 2-dimensional diagram can be stored for each structure, thus enabling a structure which was input by a chemical typewriter or a VDU to be reproduced directly prior to output. Alternatively, the structural diagram can be generated automatically from a notation. The WLN is particularly suitable for generation structure display by computer program because the largest ring in the structure appears first in the notation and overprinting is reduced to a minimum (7). Overlapping of substituents on rings can be avoided by a look-ahead procedure and by extending bonds where necessary.

Connection tables

Although notations can be searched for both specific structures and partial structures, and fragments can be generated from notations, it is not always possible to trace the required path through a notation for a substructure search. For a detailed atom-by-atom search a more explicit representation of the structure is necessary in which all the atoms and their inter-connections can be immediately identified, as in a Connection Table. The simplest form of connection table, the redundant connection table, is created by listing each atom, excluding hydrogen atoms, in the structure in turn together with all its connections. This form of connection table can be used to input structures to a computer and for atom-by-atom searching, but because each atom-bond-atom pair is recorded twice, once for each atom in the group, the redundant connection table must be compacted for storage and converted to a unique representation for registration. Rules for creating a unique compact connection table were devised at CAS (8), details of which will not be given here, but the two forms of connection table are shown below:

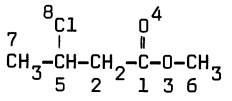
Redundant connection table:

1	C	2	1				
2	C	1	1	3	1	4	1
3	Cl	2	1				
4	C	2	1	5	1		
5	C	4	1	6	2	7	1
6	O	5	2				
7	O	5	1	8	1		
8	C	7	1				



Unique compact connection table:

C	-	-
C	1	1
O	1	1
O	1	2
C	2	1
C	3	1
C	5	1
Cl	5	1



A complete atom-by-atom search is very time-consuming, and must be preceded by an efficient screen. Fragments for screening can be generated efficiently from the compact connection table although they are in no way related to the chemical groups present in the molecule. In connection tables such as those shown above, all chemical significance of the elements is lost. Carbon is recorded as C regardless of whether it is singly or doubly bonded, in a ring or a chain, part of an alkyl group or an acid group.

A connection table can be used for structure input, for atom-by-atom searching and for simple fragment generation, but it is essentially a computer form of representation and has little or no use outside a computer system.

Chemical Structure Systems

In devising systems to handle chemical compounds, forms of structure representation are required which will fulfil the needs of three major functions:

(a) Compound registration and the retrieval of single compounds.

(b) Reduction of file size by substructure screening.

(c) Detailed structure manipulation, which is required for the following purposes:

1) Atom network searching.

2) Definition of substructures for structure/property analysis.

3) Structure display in both the two-dimensional form for information and the more refined three-dimensional form for structure evaluation.

It is necessary to state at this stage that no one method of structure representation exists which fulfils all three of these major needs. In fact, of the systems in use today none would claim to be both economic and all embracing. A few are effective, but with varying degrees of efficiency in the two functions of compound registration and retrospective retrieval, and do provide a representation which is suitable for some aspects of structure manipulation. These may include both structure/data analysis and two-dimensional structure display. A smaller number of systems have structure representation suitable for structure evaluation and three-dimensional modelling, but as yet these facilities have not been included in the large commercially-available data base systems.

The three main functions will now be considered in relation to the various data bases and systems available. Each system will be mentioned only briefly, but details of the systems can be found in the appendix or in the references.

Compound Registration

Compound registration is the process by which compounds are compared with an ordered file of structures and are only inserted in the file if they are not already present, thus avoiding the occurrence of duplicates on the file.

The major compound registration system is that of Chemical Abstracts Service which include all structures reported and indexed in Chemical Abstracts (9). The compounds are represented in the form of a unique connection table, by which they are registered. The IDC system, to be described separately, also uses a topological, or connection table, representation for registration (1).

At the Institute of Scientific Information, ISI, all novel compounds appearing in the publication Index Chemicus are coded into WLN (10). Registration is now performed by ISI, and some in-house company systems incorporate this data base into their own compound registering system. In the same way, in-house systems use the Excerpta Medica data base (11) a specialised service covering the medical field, in which the WLN is included in the bibliographic tapes produced.

Many other small specialised tape files based on the WLN are also available, such as the Aldrich Chemical Company data base (12), and the Hansch partition coefficient data file (13), both of which use the notation to register the compounds.

Services based on fragment codes can obviously not provide registration facilities.

Reduction of file size by substructure screening

The reduction of file size by substructure screening is a necessary preliminary step to detailed structure manipulation discussed below. However, many systems only provide facilities for file screening and this alone can be of value, particularly in conjunction with bibliographic text searching.

Any of the three representations, fragment codes, notations or connection tables, can be used for reduction of file size. Fragmentation is used in the Derwent Patent Services (14). Search programs are supplied by Derwent, but because the fragment code is not a total structural representation, it is not possible to carry out structure manipulation on the selected compounds. Many similar in-house systems have been set-up using fragmentation codes, for example, the Ring Code, devised by the Pharma Documentation Ring, at SKF (15).

String searches of notations, discussed earlier, can be carried out on any notation data bases. ISI supply a search program called RADIICAL for use with their WLN tapes, and at Shell Research, Sittingbourne, string search programs have been developed for use with the

IUPAC notation (4). Many computer manufacturers provide software for string searching, such as ICL's FIND2 package.

Both the fragment code and string search systems use screening techniques which are related to the chemical groups present in the molecule. When considering screens generated from connection tables, however, the chemical significance is usually lost. Most of the work on connection table screens has been carried out at Sheffield University under Professor M. F. Lynch (16) and his methods have been examined by UKCIS at Nottingham.

Structure manipulation techniques

Systems involving structure manipulation must be based on a total structure representation, i.e. either a notation or a connection table. Many systems have been set-up to deal with specific problems, but most of these systems contain a relatively small number of compounds and do not require the registration and file screening techniques discussed above. A notable example of such a system is Corey and Wipke's work on the design of synthetic pathways (17). A number of sophisticated systems have been devised, for example, the DARC system (18), Lederberg's Dendral system (19), and the Lefkovitz Mechanical Chemical Code (20)

To illustrate the breadth of services possible in systems using structure manipulation techniques, details will now be given of two major systems, ICI's CROSSBOW system and the Chemical Information System (CIS). The latter was written at N.I.H., Bethesda and has been applied to the Cambridge X-Ray Crystal File. The ICI CROSSBOW system is in use in a number of computer installations in Europe and North America. The N.I.H. system is under examination at Manchester University by Dr. O. S. Mills, using the Cambridge file, and also adopted for use with the Mass Spectral Search System, MSSS, which is accessible internationally in the Cyphernetics network. These two systems together cover the problems of retrieval from large data banks and the need to manipulate structures at the three-dimensional level for compound conformation and evaluation. A third major system, the IDC system, will be described separately.

A compound retrieval system which is suitable for structure manipulation and analysis purposes must be a multi-part system. The system will require a structure to be represented in different ways for the various manipulation techniques which must be applied. The three systems to be described, CIS, CROSSBOW and IDC, fulfil these conditions.

The ICI CROSSBOW System

Input to CROSSBOW is by use of WLN, and after input the coded compound is first checked for validity of coding using a checker program which checks molecular formula and some notation syntax. The WLN of the compound is then examined for novelty in the file and if novel this record is submitted to a fragmentation program which automatically assigns the fragments from a pre-defined set. Both the fragments and the WLN are then stored on a disc file for on-line use. This file can be searched for individual compounds and linked to property files to retrieve both the compound information and associated property data. The system provides substructure retrieval facilities and the three levels of search are available when answering users questions. The first level involves a fragment search and the second uses string searching of the WLN. These levels have been designed to be complementary to each other, and their main purpose is to reduce the file size efficiently. The third level of search is an atom-by-atom search carried out on the sub-file produced by the two screening levels of search. Before compounds can be searched at the atom-by-atom level, the sub-file is converted from the WLN form of representation to the CROSSBOW connection table form. This connection table is automatically derived by program from the WLN. Unlike a normal atom connection table it consists of chemical units which are symbols describing both the atom and its associated bonds.

The use of chemical units greatly increases the speed of identification of specific atom arrangements when compared with the normal connection table. In the CROSSBOW connection table connectivity between adjacent atoms is assumed unless a statement to the contrary occurs. This not only reduces the size of stored records, but has the advantage that these statements signal the terminal and branching atoms. Finally, the CROSSBOW connection table stores the ring atoms of each ring together, and preserves the ring junction information given in the WLN. Thus, the resulting connection table has important information recorded which facilitates the compound manipulation necessary at the third level of search. At this level of search the atom network of a molecule can be manipulated both for substructure retrieval and structure analysis purposes. The connection table records produced for the third level of search are also used to generate a two-dimensional structure display. This is used at output when retrieved structures are printed onto 8" x 5" index cards along with their associated property data.

The CROSSBOW system has been used for substructure retrieval purposes on files of half a million compounds. The system has been employed successfully to literature retrieval, structure property, NMR C^{13} and to reaction analysis problems. Last year using both structure and property files over 6,000 queries were answered by ICI's in-house system.

N.I.H. Chemical Information System Applied to the Cambridge X-Ray File

The N.I.H. system also provides various levels of search. The approach to the problem of substructure searching is entirely different from ICI CROSSBOW. Firstly, the N.I.H. system uses Chemical Abstracts connection tables for input, and secondly, the emphasis is on a self-operation by the enquirer, whereas the ICI CROSSBOW system is primarily set-up to be operated by information specialists. The N.I.H. system achieves this user interface by storing records on compounds in a strict hierarchical order, particularly good for ring compounds. By interrogating on-line under single parameters the enquirer can quickly set-up sub-files of those compounds containing their desired structural features. He can repeat this on other features building up sub-files which he can ultimately match to locate those compounds containing all the desired components. Boolean logic can be applied when setting-up questions. During the process of file interrogation, the user is kept informed of the number of hits at each stage. Setting-up structural features as part of a question is relatively easy, especially on a graphics terminal. Computer commands initiated at the terminal enable the enquirer to build-up complex rings with substituent patterns. A structure display program is available and facilities are provided for the enquirer to display hits sequentially.(21)

Using the Cambridge X-Ray file questions can be formulated which include the use of co-ordinates. Structure diagrams are then given spatial characteristics. It is possible to turn this form of display through an axis to examine spatial relationships and then display in a double image for stereo viewing. Finally the selected viewing angle can be converted into a molecular model form.

The CIS/X-Ray file was chosen as an example of a system based on connection tables because it provides a scientist-based interrogation system with wide facilities. CIS is also being applied to the MSSS Mass Spectra System which holds over 30,000 compounds. The facilities of CIS have also been adapted to the National Cancer Institute (NCI) file of some 300,000 compounds submitted to screening tests. The NCI system uses Chemical Abstracts as a bureau to register and hold the file. On-line retrieval of compounds and substructure enquiries is possible and also a file of structures is held which provides structure diagram output on a graphics terminal.

The development work carried out at NIH and applied to connection table files produced by Chemical Abstracts provides an attractive way forward because it results in an on-line system which can be handled directly by the scientist. For notation-based systems to be equally attractive to users it will be necessary to provide a similar interrogation system to that developed at NIH.

CONCLUSION

The full potential of structure searching and manipulation has not been exploited by the commercial organizations, and commercial systems in use today do not claim to completely cover the three main functions required of a system. It is possible that the need for such a system does not exist universally and even if the need were established, it is doubtful whether universal agreement could be obtained on the methods to adopt in the system. There is a simple explanation for this situation. The global needs have never been evaluated and nearly all progress in the field of structure systems has been made by individuals fulfilling local needs, one of which is often the need to make a profit. Most of the techniques do exist to enable a comprehensive system to be built up which would meet the diverse needs of scientists. The major factors missing, however, are the assembly of structural data in a unified form and the universal adoption of a registry system.

REFERENCES

(1) 'Chemical Information Systems' Eds. Ash, J.E. and Hyde E. Ellis Horwood Ltd. Chichester 1975 Chapter 13: The IDC System

(2) Dyson, G.M. A Notation for Organic Compounds. Nature 154 114 (1944)

(3) Rules for the IUPAC Notation of Organic Compounds Longmans 1961

(4) Dammers, H.F. and Polton, D.J. Use of IUPAC Notation in Computer Processing of Information on Chemical Structures J. Chem. Doc. 8 (3) 150-160 (1968)

(5) The Wiswesser Line-Formula Chemical Notation. 3rd Edition. Smith & Baker. CIMI, P.O. Box 2740, Cherry Hill, New Jersey 08003.

(6) Dow Checker program

(7) Thomson, L.H., Hyde, E. and Matthews, F.W. Organic search and display using a connectivity matrix derived from the Wiswesser Notation. J. Chem. Doc. 7 (4) 204-207 (1967).

(8) Morgan, H.L. The generation of a unique machine description for chemical structures - a technique developed at Chemical Abstracts Service. J. Chem. Doc. 5 (2) 107-113 (1965)

(9) Dittmar, P.G., Stobaugh, R.E., & Watson, C.E. Chemical Abstracts Service, Chemical Registry System. I. General Design. J. Chem. Inf. Comput. Sci. 16, 111 (1976)

(10) Garfield, E. et al. Index Chemicus Registry System. Pragmatic approach to substructure chemical retrieval. J. Chem. Doc. 10 (1) 54-58 (1970)

(11) Blanken, R.R. & Stern, B.T. Planning and Design of on-line systems for the ultimate user of biomedical information. Information, Processing & Management, 11 207-227 (1975)

(12) Aldrich Chemical Co. Catalogue of chemical compounds available on magnetic tape.

(13) Hansch, C. et al. Partition Coefficients and their Uses. Chem. Rev. 71 525-616 (1971)

(14) Hyams, M. Chemical patents information. Chemistry in Britain 6 (10) 416-420 (1970)

(15) Craig, P.N. and Ebert, H.M. Eleven years of structure retrieval using the SK & F Fragment Codes. J. Chem. Doc. 9 141-146 (1969)

(16) Lynch, M.F. et al. Evaluation of a substructure search screen system based on bond-centred fragments. J. Chem. Doc. 1974, 14 (1), 44-8.

(17) Corey, E.J., and Wipke, W.T. Computer-assisted design of complex organic syntheses. Science 166 178-192 (1969)

(18) Computer Representation and Manipulation of Chemical Information. Eds. Wipke, W.T., Heller, S.R., Feldmann, R.J. and Hyde, E. Wiley-Interscience 1974. Chapter 10: DARC System in Chemistry.

(19) Ibid. Chapter 12: Heuristic Dendral Analysis of Molecular Structure.

(20) Lefkovitz, D. A chemical notation and code for computer manipulation. J. Chem. Doc. 1 (4) 186-192 (1967)

(21) Ibid. Chapter 3. Interactive Graphic Chemical Structure Searching.

APPENDIX

Aldrich Chemical Company data base

A magnetic tape data-base is produced by the Aldrich Chemical Company to cover commercially-available compounds in the main Aldrich Catalogue and in the Alfred Bader Special Library. Compounds are coded in WLN and are recorded together with their catalogue number. An additional tape of names of compounds, molecular formula and the price of the compound is also available. No services are provided with the tape, but the tape has been linked to the CROSSBOW system at ICI, through which substructure searches are made within ICI.

Derwent Publications Ltd

Derwent Publications Ltd specialise in the documentation of patents in chemical and related fields, and produce the publication Central Patents Index. This is divided into 12 sections and 3 of these sections, known as Farmdoc, Agdoc and Chemdoc, which deal mainly with non-polymeric chemical compounds, form the basis of a magnetic tape file of compounds. All compounds appearing in these three sections are coded using a General Chemical Code: a fragment code with over 300 features. The code gives a broad classification of the compound type and details the types of rings, coding both specific and more general fragments, and codes functional groups. A pre-classification system such as this is not suitable for inclusion of new concepts. However, the large number of fragment types recorded enables searches for specific fragments or for a variety of fragment combinations to be performed. Search programs are supplied by Derwent for IBM series machines, but it is obviously not possible to carry out searches for specific structures or to undertake a detailed atom-by-atom search on the tapes.

Excerpta Medica Foundation

A specialised system covering 3500 journals in the medical field is provided by Excerpta Medica. The tape services, Drugdoc, includes the full chemical name of all drugs reported, together with the WLN, the generic and trade names, the bibliographic reference, medical indexing terms, adverse effects, etc. An alerting service is offered, based on the name of the drug, but no structure searching is undertaken.

Institute of Scientific Information: Index Chemicus Registry System

All novel compounds appearing in the publication Index Chemicus are coded into WLN and are recorded on a file with the corresponding molecular formula, and the abstract and compound numbers. The tape is produced monthly and is provided in conjunction with a bibliographic tape. The monthly tape contains about 10,000 compounds and a cumulative file for retrospective searching is available which contains over 1 million compounds. A string search program, RADIICAL, is available for use with these tapes.

Hansch Partition Coefficient Data File

At Pomona College, Hansch and his co-workers produce a tape file of structures, coded in WLN, together with physical properties such as partition coefficient and PKA values. The file contains about 6,000 compounds, updated six-monthly, the data for which is obtained from the literature or directly from research workers. Structures are registered using the WLN and this tape has also been linked to the CROSSBOW system.