Pure & Appl. Chem., Vol. 49, pp. 1807 - 1816. Pergamon Press, 1977. Printed in Great Britain.

COMPUTER BASED SYSTEMS FOR THE RETRIEVAL OF DATA: CRYSTALLOGRAPHY

Olga Kennard*, F. H. Allen, M. D. Brice, T.W.A. Hummelink, W.D.S. Motherwell, J. R. Rodgers, D. G. Watson

University Chemical Laboratory, Lensfield Road, Cambridge, England.

Abstract - The article reviews four major computerised crystallographic databases: the Powder Diffraction File; Crystal Data; the Protein Data Bank and the Cambridge Structural Database. The Structural Database is described in some detail. Four retrieval systems: the Feldmann Search System; the DARC System; the TOOL-IR System and the Cambridge System, which have been applied to the Structural Database are discussed with illustrations of some typical search questions using the Cambridge System.

INTRODUCTION

The word data covers a multitude of sins. Its Oxford Dictionary definition is "things known or granted; assumptions or premises from which inferences can be drawn; facts of any kind". The most workable definition, for those of us in the information field, is the second - "assumptions or premises from which inferences can be drawn". This definition covers not only the contents of information systems but also the motivation, which has led scientists to get involved with such systems - the implicit hope that from the units of facts, the bricks of data in a given scientific field or combination of fields, new inferences and new predictions can be made. This, after all, is the very basis of science from the Darwinian hypothesis to the latest theories of cosmology.

CRYSTALLOGRAPHIC DATA BASES

Data in crystallography used to mean measurements on the external shape of crystals, from which much of our knowledge of crystal symmetry and, indeed, most of systematic mineralogy were derived.

Since the discovery of the nature of X-rays data generally refers to some aspect of the diffraction pattern which results from exposing crystalline material to X-rays or neutron beams. Broadly speaking there are two types of patterns - those given by crystalline powders, and those given by single crystals, and they serve two distinct purposes.

Powder patterns are mainly used for identification of chemical compounds or mixtures and a very large database, the Powder Diffraction File, covering some 25,000 powder patterns, has been developed by the Joint Committee for Powder Diffraction Standards in the U.S.A. The Powder Diffraction File and associated retrieval programs is similar to systems developed for mass spectrometry and infrared spectroscopy. The database contains primary experimental data on the distribution and intensity of diffraction patterns and the retrieval of information and identification are essentially problems of pattern matching. The Powder Diffraction File has great industrial importance and is a widely used and economically viable data system.

Some single crystal data can also be used for identification purposes, and a database of this type is being maintained by the Materials Division of the National Bureau of Standards, U.S.A. The contents of the database have been published[1] as part of the National Standard Reference Data System, and recently part of the database has been released on magnetic tape, with appropriate retrieval programs.

The crystallographic database most relevant to the topic of this meeting contains information not about the diffraction pattern but the numeric data derived from that pattern, principally the positions of atoms in crystals. It is thus not a databank of primary experimental measurements but derived data based on mathematical models. The retrieval of these types of

*External Staff, Medical Research Council

data is mainly of interest to academic scientists who wish to use them for further fundamental research. The atomic positions and a knowledge of expected covalent bond-lengths define the connectivity of the molecule and further derived data, on molecular geometry and intermolecular contacts, can be obtained by quite simple calculations.

Single crystal data are available from two computerised databases. The Protein Data Bank, maintained by the Brookhaven National Laboratory (U.S.A.) currently contains 60 data sets on macromolecules, principally proteins. Tapes are distributed by the Brookhaven National Laboratory in the U.S.A.,by the Computer Centre, University of Tokyo in Japan and by the Cambridge Crystallographic Data Centre elsewhere.

Data on organic compounds and organometallic complexes, excluding proteins and polymers, are compiled by the Cambridge Crystallographic Data Centre[2]. The database contains information on some 16,000 compounds (November 1976) with about 2,500 - 3,000 new entries added annually. Although this is a small number when compared, say, with the CAS database, the file has a number of unusual features which make it a splendid candidate for testing various chemical retrieval systems.

In the next section the main features of the database will be outlined, followed by a description of four retrieval systems which have been used in conjunction with the file.

THE STRUCTURAL DATABASE

The Structural Database contains information on crystal structures of organic compounds and complexes determined by X-ray or neutron diffraction methods and is retrospective to 1935. Historically the database has been divided into three separate files: the Bibliographic File, the Chemical Connectivity File and the Numeric Data File.

Each distinct crystallographic study constitutes a database entry and is identified by an alphanumeric reference code. Each study has an entry in each of the three main files and the reference code provides the link between the files.

The Bibliographic File

The main elements of the Bibliographic File are:

* Compound Name(s)
* Qualifier(s)
* Molecular Formula
* Authors' Names
* Literature Reference
* Chemical Classification
* Reference Code

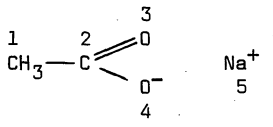
These elements serve as search fields in the different retrieval programs, and much thought has gone into the content and formulation of these fields. The title of a publication is replaced by the compound name with, if necessary, a suitable qualifier. The qualifier gives information mainly about experimental details e.g. "absolute configuration determined", "neutron study" etc. The qualifiers are important retrieval elements for specific queries. The compound name itself, however, is not considered to be a high precision search field and no attempt was made to assign systematic IUPAC names.

The database is organised with respect to 86 broad chemical classes and each entry is allocated to a main class and possibly to several subsidiary classes. With few exceptions, such as antibiotics, the classes refer to chemical structure rather than to function, and act as broad screens for searching the bibliographic file, often in conjunction with other data elements. A detailed description of the file is given by Kennard, Watson and Town[3].

The Connectivity File

The chemical connectivity file is the main tool for chemical structure retrieval and the decision to adopt a connectivity representation was taken after a pilot experiment using the WLN notation. Since the file contains a high percentage of complex structures for which new notation rules would have to be devised it was decided to code the structures in terms of connectivity tables, with conventions very similar to those used by CAS. A typical connectivity table is shown in Figure 1.

FIG 1: CODING OF A SIMPLE CONNECTIVITY TABLE



	EL	NCA	NH	NCH		I	J	BT
AT1	C	1	3	0	BØ	1	2	1
AT2	C	3	0	0	BØ	2	3	2
AT3	Ø	1	0	0	BØ	2	4	1
AT4	Ø	1	0	-1				
AT5	NA	0	0	+1				

NCA = number of connected atoms, excluding terminal hydrogens
NH = number of attached terminal hydrogen atoms
NCH = net charge on the atom
BT = bond type

Six different bond type codes are available: single (±1), double (±2), triple (±3), aromatic (-5), delocalised double (±7) and π (+9), where positive and negative signs denote acyclic and cyclic bonds, as detected by a ring analysis program. The assignment of bond types can sometimes be difficult, particularly in the case of novel structures, and searches involving bond type 7, tautomers etc. need special care.

The Numeric Data File

The main elements of this file are:

* Unit Cell Data
* Atomic Coordinates
* Published Bond Lengths
* Crystal Connectivity
* Remarks
* Summary Flags
* Reference Code

Details of this file are mainly of interest to subject specialists and are fully described by Allen et al[4]. The numeric data are checked for internal consistency, primarily by recalculating the bond lengths from the coordinates on file and comparing with the author's published values[5]. In this process transcription errors, printers' errors and possible inconsistencies in the original publication are located and corrected. Considerable effort has gone into the checking of the numeric data which, however, is becoming an increasingly difficult task since the database has now a doubling period of approximately 4 years while the staff of the Data Centre remains constant. The integrity of the numeric data, however, is of prime importance since it forms the information reservoir of the database. Unlike many other databases, which yield search 'hits' only in the form of bibliographic references, the bibliographic and connectivity searches of the Structural Database lead the users directly to the numeric data, without, as far as possible, needing to consult the original publications. The numeric data retrieved in this way are then used for a variety of further calculations and the production of graphic illustrations.

RETRIEVAL SYSTEMS

The Structural Database was developed over a period of 10 years by crystallographers working in a university chemical department and in close touch with potential users. The aim was to set up a well-defined file structure with adequate search fields to serve future needs but not specifically tied to any particular retrieval system. Currently four different search systems have been implemented.

The Feldmann Interactive Graphic System

The main features of the system are:

* Fully interactive
* Uses a PDP 10 computer
* Uses inverted files
* Searches are based on tree structures
* Connectivity searches are based only on the crystallographic connectivity
* Graphical interactions between the user and program

The system was developed by Richard J. Feldmann[6] at the U.S. National Institutes of Health initially for a subset of the CAS Structure Registry File and later extended to the Structural Database. It is an interactive system which makes use of a large time-shared computer, direct access storage devices and graphic display terminals. It has been designed to allow maximum interaction between the user and the system, with facilities to refine input queries, very rapid substructure searches and display and communication with the system in graphic terms.

The rapid searches are achieved by the use of inverted files but the preparation of these involves substantial computing time.

At present the connectivity searches are restricted to the crystallographic connectivity tables included in the Numeric Data File. These do not carry information on bond type and certain atom properties. A further limitation of this restriction is that compounds for which no atomic coordinates have been published e.g. preliminary communications, conference reports etc. cannot be retrieved. Such publications form about 30% of the database. It is hoped that during 1977 the programs will be implemented to search the Chemical Connectivity File of the database.

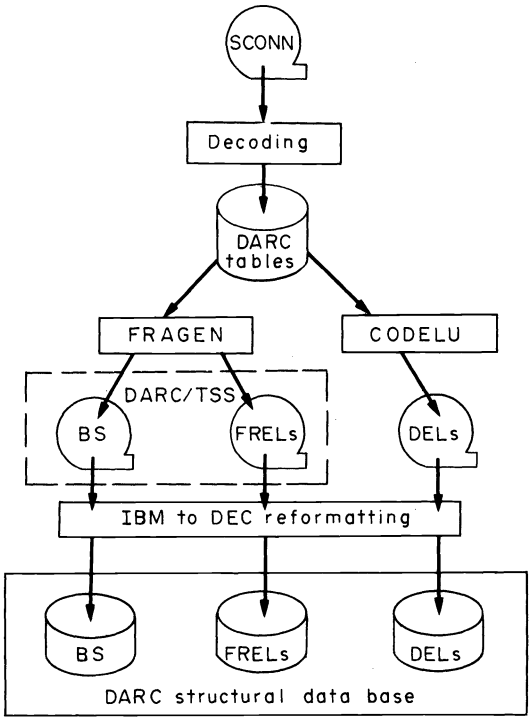
During the past year the system has been extensively tested at the Brookhaven National Laboratories in the U.S.A. and at the Rutherford Laboratories and Manchester University in the U.K. It is now available to users in the U.S.A. via CYPHERNET and in Scandinavia via SCANNET. In the U.K. it will be available shortly through the Rutherford Laboratories and a demonstration of the system has been arranged at the Symposium by two members of the Laboratory, Dr. M. Elder and Miss P. Machin.

The DARC System

The main features of the system are:

* Fully interactive
* Uses a PDP 11/35 computer
* Input is via a Rand tablet
* Screens
* Node-by-node graph matching
* Searches the Chemical Connectivity File in the DARC code

The DARC system, developed by Professor J. E. Dubois[7] and co-workers has been implemented on the Structural Database during the past few months as shown in flowchart 1.



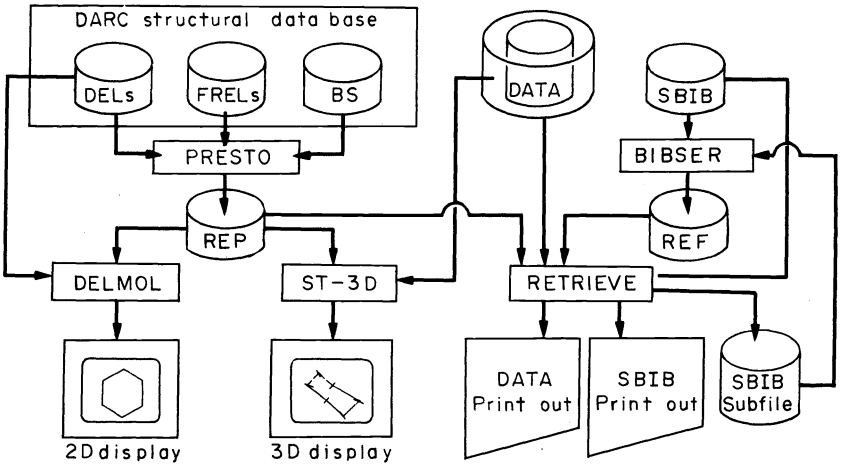
Flowchart 1

GENERATION OF THE DARC STRUCTURAL DATA BASE FROM CAMBRIDGE CONNECTIVITY FILES

The Chemical Connectivity File (SCONN) was converted to the DARC connection tables, which are common to all the DARC input processes. The DARC retrieval system and the software developed by the Cambridge Crystallographic Data Centre - particularly the bibliographic search program BIBSER and the retrieval program RETRIEVE, which uses the reference code to retrieve information from the various files have been combined into one search system. Flowchart 2 illustrates the system and the different types of output, which include 2-dimensional or 3-dimensional display of chemical structures, print-out of numeric data or bibliographic references, creation of subfiles.

Flowchart 2

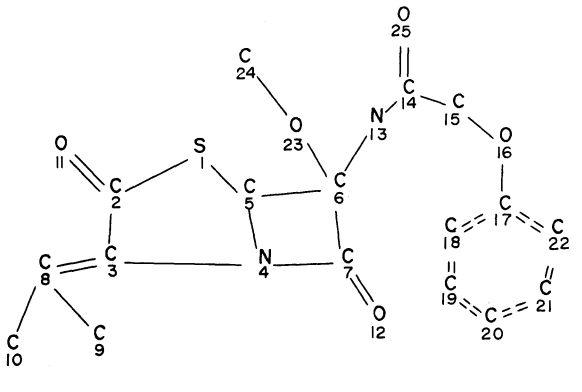
PLURIDATA RETRIEVAL SYSTEM
ON CAMBRIDGE CRYSTALLOGRAPHIC DATA BANK



The structure or substructure sought can be drawn freely on a Rand tablet with alpha-numeric information entered from a keyboard. Fig 2 shows the input when a completely defined structure is sought and the structure found in the file and generated from the DARC code.

FIG 2: RETRIEVAL OF FULLY DEFINED STRUCTURE

Structure Sought



Structure Found

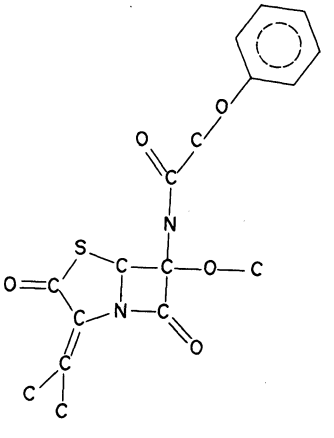


Fig. 3 shows four different three-dimensional views using the Numeric Data File. Several plot options, such as atoms represented by spheres etc., are possible. There are interactive facilities for rotating the molecule to obtain different views. It is also possible, interactively, to display on the screen numeric values for bond lengths, valency angles or dihedral angles.

Substructure searches are initiated via the Rand tablet. The search is carried out in several stages, using the DARC topological screen system followed by a node-by-node search.

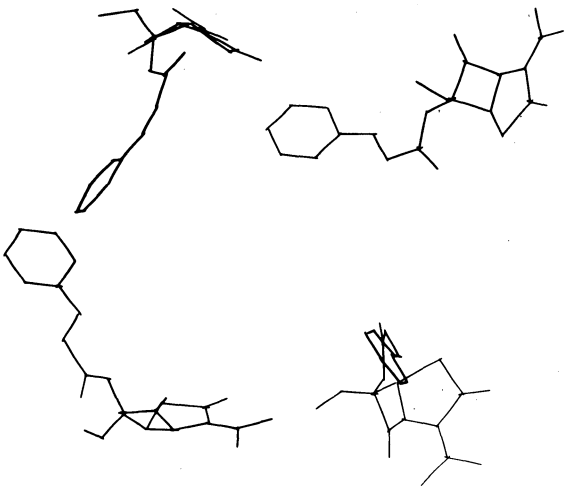


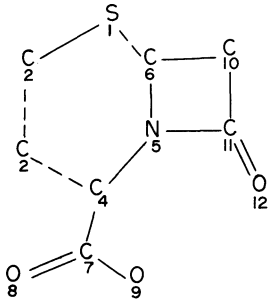
FIGURE 3

DIFFERENT THREE-DIMENSIONAL VIEWS OF STRUCTURE FOUND

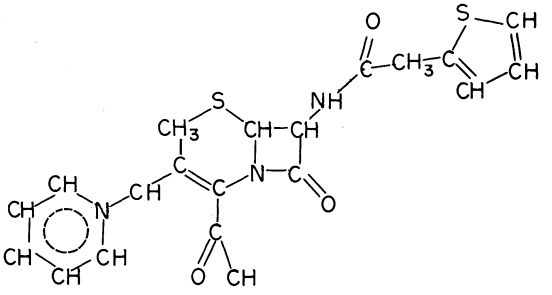
Fig. 4 illustrates the input of fused 4- and 6-membered ring systems. Dotted lines indicate any bond type and atoms marked with an asterisk are those which can be substituted. The search was restricted to structures published since 1973 and resulted in 4 hits.

FIG. 4: SUBSTRUCTURE SEARCH

Substructure sought



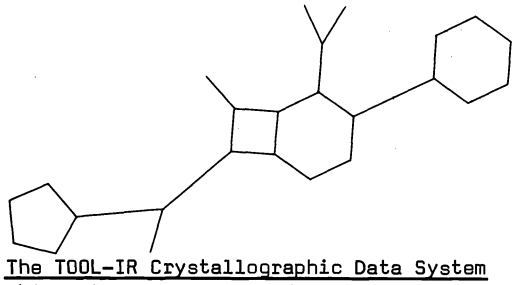
One of the structures retrieved



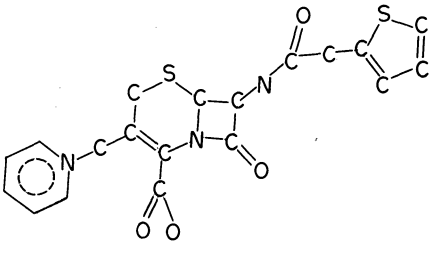
Some of the options available for displaying these hits are illustrated in Fig. 5. Other options for retrieving the bibliographic and numeric data and displaying the structures are as described for the completely defined structure.

FIG. 5: OPTIONS FOR DISPLAYING A 'HIT'

Skeletal



With hydrogen atoms



The TOOL-IR Crystallographic Data System

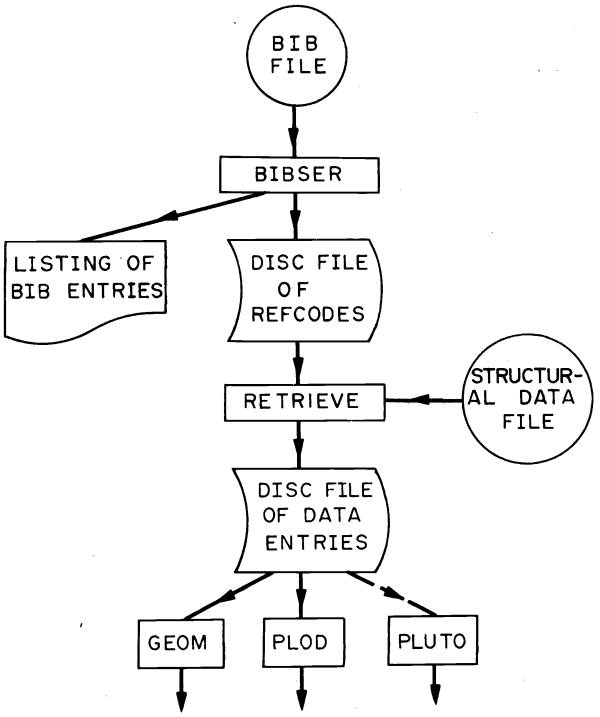
This sytem was developed as a joint project between Professor T. Shimanouchi and Dr. T. Yamamoto at the Computer Centre, University of Tokyo and users in Japanese universities. The system searches the bibliographic file in an inverted form and uses the resultant reference codes to access numeric data for further calculations e.g. molecular geometry and plotting. The files have been placed on-line and can be accessed by over 100 time-sharing terminals and 20 remote batch terminals throughout the country. Fuller details of the retrieval technique used are given in reference 8.

The Cambridge Retrieval System

The fourth retrieval system was developed at the Cambridge Crystallographic Data Centre specifically for the Structure Database by Dr. W.D.S. Motherwell in collaboration with

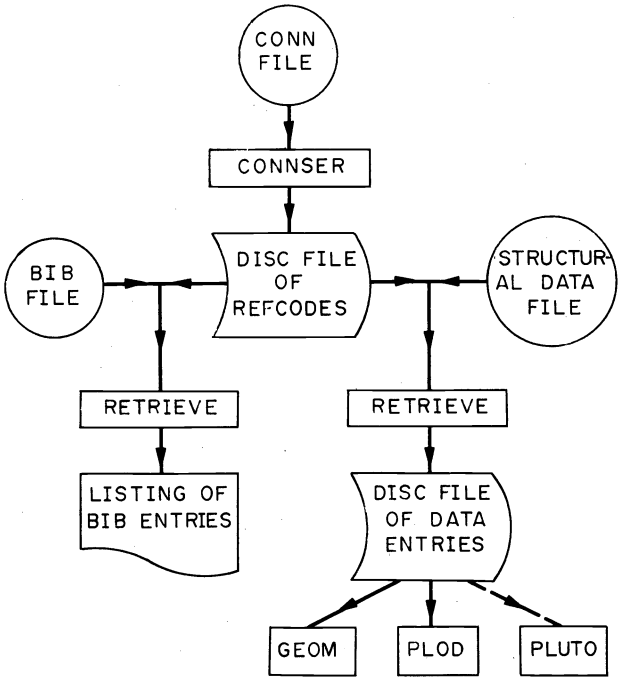
members of the Centre and in conjunction with users providing actual search queries. The programs were modified in the light of practical experience.

The system consists of two main programs, the bibliographic search BIBSER and the connectivity search CONNSER. Numeric data, corresponding to hits from either search, can then be retrieved via the reference code which links files together (using the program RETRIEVE). The programs were written in Fortran IV and have been implemented on the IBM 370/165, CDC and Burroughs Computers. The system functions in batch mode. An overview is given in flowcharts 3 and 4.



Flowchart 3

CAMBRIDGE RETRIEVAL SYSTEM
BIBLIOGRAPHIC SEARCH



Flowchart 4

CAMBRIDGE RETRIEVAL SYSTEM
CONNECTIVITY SEARCH

FIG 6: BIBLIOGRAPHIC SEARCHES

The formulation of a bibliographic search question is indicated in Fig 6.

Bibliographic Search Formulation

```
Q *CLASS '83-86' AND *ELEMENT 'NI' NOT 'HL'
|    |          |   |
|    |          |   └LOGICAL OPERATOR
|    |          └STRING DEFINING A RANGE
|    └SEARCH FIELD
└QUESTION MUST BEGIN WITH Q
```

The search is essentially a character-string matching process. Logical operators AND/OR/NOT can link strings and/or search fields. Ranges of strings, as indicated in the example are possible. A special field for 'ELEMENT' has been introduced and special element symbols for groups of elements, such as HL for halogens, TR for transition metals, can be used.

Searches are in batch mode and the cpu time depends mostly on the number of hits. A search resulting in, say, 200 hits will take 20 seconds cpu whereas one resulting in 2,000 hits will take about 60 seconds cpu on an IBM 370/165.

Bibliographic searches can also be performed manually using a variety of indexes such as a permuted KWIC index of compound names (Allen and Town[9]), molecular formula indexes (ordered by C, H count and permuted on rarer elements) and indexes of authors' names and journal references.

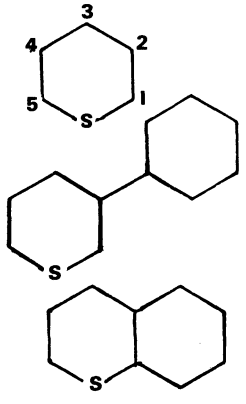
SUBSTRUCTURE SEARCHES

Substructure searches are carried out on the Chemical Connectivity File by a node-by-node graph matching technique and the resultant 'hits' are a list of reference codes which can be used to retrieve both bibliographic and numeric information. The input is a connectivity record of atom and bond properties. Comment records and keywords can also be input.

At present there are no screens in the file though we plan to introduce these shortly. The precision of retrieval is, nevertheless, high mainly because of built-in flexibility which allows the user to formulate his questions at the required level of precision.

Tests may be specified or omitted on the various atom and bond properties. Ranges of element and/or bond types are allowed and it is possible to specify either an exact match or a match within a range. The environment of the search fragment may be controlled by keywords like those shown in Figures 7 and 8.

FIG 7: USE OF KEYWORD NOCS FOR SUBSTRUCTURE SEARCHES

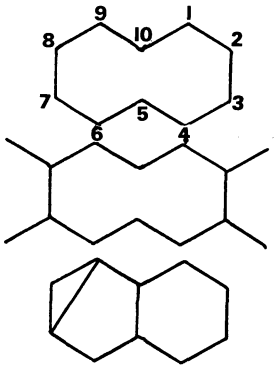


NOCS specifies that no atom in the numbered diagram can have cyclic substitution additional to that already present.

One of the structures retrieved if NOCS is present.

Additional structure retrieved if NOCS is absent.

FIG 8: USE OF KEYWORD NOLN FOR SUBSTRUCTURE SEARCHES



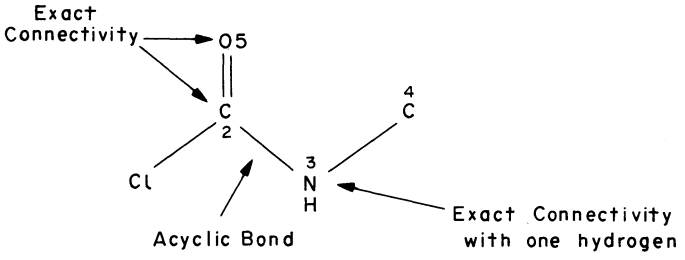
NOLN specifies that no atom in the numbered diagram can be linked to any other atom in the numbered diagram except by the bonds shown.

One of the structures retrieved if NOLN is present.

One of the structures retrieved if NOLN is absent.

Figure 9 illustrates the use of some of these options in an actual search query. It was submitted by Professor A. T. North of Leeds University who wished to investigate the effect of environment on the peptide bond. By formulating the search input as indicated in the Figure 266 entries were retrieved most of which were relevant to the query. At the next stage in this particular investigation the search can be extended to cyclic peptides, by changing the bond type specification.

FIG 9: SEARCH FOR A PEPTIDE BOND FRAGMENT (ACYCLIC) ON AN IBM 370/165 COMPUTER



Formulation of question

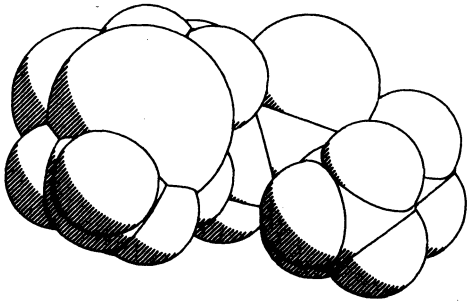
```
Q PEPTIDES (ACYCLIC) ONE HYDROGEN ON NITROGEN
AT1 C 1                    BØ 3 2 1 A
AT2 C 3 O E                BØ 2 5 2 A
AT3 N 2 1 E                BØ 2 1 1 A
AT4 C 1                    BØ 3 4 1 A
AT5 Ø 1 O E                END
```

Search of 14438 compounds gave 266 hits and used 39 secs cpu

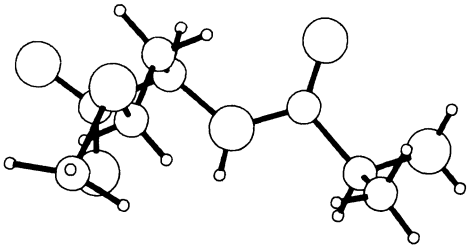
Fig. 10 illustrates two different options for displaying one of the structures retrieved. A variety of other options, some similar to those used by the DARC system are also available. In addition it is possible to obtain tables of bond-lengths, bond angles and torsion angles; the coordinates and symmetry operators which have been retrieved can also be used to calculate intermolecular distances. These are some of the basic data of interest to scientists who are attempting to use the information contained in the Structural Database for the development of new hypotheses.

FIG 10: OPTIONS FOR DISPLAYING RETRIEVED STRUCTURE

Space filling diagram of structure retrieved



Ball-and-stick diagram of structure retrieved



Improvements of both the search programs and the files are planned and the introduction of a fragment screen in particular should greatly speed up retrieval. A sample database of 283 entries together with the corresponding structural diagrams is available to anyone interested in testing a particular retrieval system. Experience to date indicates that each retrieval system has features which make it more or less desirable depending both on the user and computer environment and much can be learned from a comparison of different search strategies applied to the same database.

Acknowledgements

We thank the following organisations:- The Office for Scientific and Technical Information (1965-1974) and the Science Research Council (1974-) for financial help; the Medical Research Council for allowing a member of their staff (O.K.) to participate in this work and the University of Cambridge for the provision of space and computing facilities. We are grateful to Professor Dubois for diagrams illustrating the DARC System.

REFERENCES

1. Crystal Data. Determinative Tables. 3rd Edition (Editors: J.D.H. Donnay and H.M. Ondik). Vol. 1 Organic Compounds, 1972. Vol. 2 Inorganic Compounds, 1973. U.S. Department of Commerce, National Bureau of Standards and the Joint Committee on Powder Diffraction Standards.
2. O. Kennard, D.G. Watson, F.H. Allen, W.D.S. Motherwell, W.G. Town and J.R. Rodgers, Chem. Br. 11, 213 (1975).
3. O. Kennard, D.G. Watson and W.G. Town, J. Chem. Doc. 12, 14 (1972)
4. F.H. Allen, O. Kennard, W.D.S. Motherwell, W.G. Town, D.G. Watson, J. Chem. Doc. 13, 119 (1973).
5. F.H. Allen, O. Kennard, W.D.S. Motherwell, W.G. Town, D.G. Watson, T.J. Scott and A.C. Larson, J. Appl. Crystallog. 7, 73 (1974).
6. R.J. Feldmann, Computer Representation and Manipulation of Chemical Information (Editors: W.T. Wipke; S.R. Heller; R.J. Feldmann and E. Hyde) p. 55, John Wiley and Sons, New York, (1974).
7. J.E. Dubois, J. Chem. Doc. 13, 8 (1973)
   J.E. Dubois and A. Panaye, Bull. Soc. Chim. Fr., 1229 (1976)
8. T. Shimanouchi and T. Yamamoto, Crystallographic Data Services in Japan, Proc. 5th International CODATA Conference, Boulder, (1976)
9. F.H. Allen and W.G. Town, J. Chem. Doc and Computer Sci., 17, 9 (1977).