The Development of Versions 3 and 4 of the Cambridge Structural Database System†

FRANK H. ALLEN,* JOHN E. DAVIES, JEAN J. GALLOY, OWEN JOHNSON, OLGA KENNARD,* CLARE F. MACRAE, ELEANOR M. MITCHELL, GARY F. MITCHELL, J. MICHAEL SMITH, and DAVID G. WATSON

Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England

Received January 18, 1991

The Cambridge Structural Database (CSD) records bibliographic, 2D chemical, and 3D structural results for organocarbon compounds studied by X-ray and neutron diffraction. In January 1991, the CSD contained 86 026 entries derived from 584 primary sources. The CSD system comprises the database, together with associated software for search, retrieval, analysis, and display of the stored information. Over the past few years, the CSD system has been considerably upgraded to provide efficient, integrated, and user-friendly search facilities. The development of two new systems, Version 3 and Version 4 are described here. Both systems operate from a completely restructured CSD search file (ASER) upgraded to include bit-screen heuristics (Version 3), and further upgraded by a complete set of fully digitized 2D chemical diagram representations of database entries (Version 4). A new program, QUEST, provides integrated search facilities for text, numeric, and 2D chemical information; 2D chemical similarity searching is also included. The alphanumeric query language of Version 3 QUEST is replaced by a fully interactive menu-driven graphical interface in Version 4 in which high-quality 2D chemical diagrams form part of standard system output. Subsequent 3D search and data analysis operations are performed by the program GSTAT, common to both new systems; 3D graphical output is generated by the PLUTO package. A "general" entry-sequential search file permits both versions to operate on a wide variety of hardware platforms with minimal installation problems. Special implementations have also been developed for DEC-VAX/VMS and Silicon Graphics (Unix) environments, to take advantage of more efficient, machine-specific search file organizations. Plans for a further database upgrade and for the integration of extended 3D search capabilities into QUEST (Version 5) are briefly discussed.

INTRODUCTION

Compilation of the Cambridge Structural Database (CSD) began in 1965, at a time when the present widespread use of chemical databases, particularly those relating to 3D structures, could scarcely have been foreseen. In fact, the origins of the CSD date back even further to a Royal Society Conference on Scientific Information held in London in June 1948.[1] Here, the physicist and crystallographer J.D. Bernal stated (page 54 of ref 1) that "... the advance of science is absolutely dependent on the effective satisfaction of scientific workers [for information]...", and that "... the growing abundance of primary scientific publication and the confusion with which it is set out acts as a continuous brake, as an element of friction, to the progress of science". Even with this impetus from one of its leading practitioners, and despite an enviable record of self-documentation,[2] crystallography had to wait another 20 years for the inception of its first computerized database designed to record the complete 3D structural results obtained by X-ray and neutron diffraction.

In 1965, the climate was right, both politically and technically, for the funding and execution of such a project. Government organizations worldwide were becoming aware of the "information explosion", and steps were being taken to allocate various subject areas to different countries. The dramatic advances in computer technology (even then) provided the tools necessary for the compilation and dissemination of these new information resources. Thus, the conjunction of an idea, suitable funding, adequate technology, and a small group of interested subject specialists made it possible to plan a permanent organization—the Cambridge Crystallographic Data Centre (CCDC)—with long-term aims involving: (a) The compilation of a numerical database relating to organocarbon crystal structures; the database was to be fully retrospective, critically evaluated, and widely distributed to the scientific community. (b) The development of software for search, retrieval, analysis, and display of the results contained in the CSD. (c) Fundamental research using the results accumulated in the CSD, employing software developed at the CCDC and elsewhere.

As a database producer, software developer, and research unit, the CCDC differs from most database organizations that concentrate almost entirely on the first function. However, the symbiotic relationship between the three functions has been vitally important in the overall development of the CCDC in the last 25 years. Of necessity, the CCDC was originally staffed by subject specialists: crystallographers who participated in all aspects of the work, assisted by limited clerical backup. This essential staffing core has now expanded to include skilled technical editors, with training in chemistry and information science, and also specialists in computing systems and in software development.

The establishment of a fully retrospective database, maintained on a current basis, was the initial goal for the period 1965–1980. Information was acquired in three logical subdivisions: (i) bibliographic and chemical text, (ii) 2D chemical connection tables, and (iii) 3D numeric structural data. Each of these subdivisions gave rise to a separate file denoted as (i) FBIB, (ii) FCON, and (iii) FDAT in their formatted character forms for distribution and as ABIB, ACON, and ADAT, respectively, in their archive binary forms for searching. Each crystal structure gave rise to an entry in each of these three files; the subentries being linked by a common CSD reference code (six letters identifying the chemical compound, two digits tracing the publication history). During this period, the Centre concentrated on in-house software and procedures for the checking and evaluation[3] of data in each of the three files. Systems were also developed for the dissemination of the bibliographic and chemical information in reference-book

† Dedicated to Professor Michael F. Lynch to commemorate the 25th anniversary of his appointment at the Department of Information Studies, University of Sheffield, U.K.

0095-2338/91/1631-0187$02.50/0

form.[4]

The first search and retrieval software was made available in the late 1970s, together with facilities for data analysis and for 3D search and display of structures. This Version 1 system centered around two separate search programs, BIBSER and CONNSER, acting on ABIB and ACON, respectively; full details are available elsewhere.[5] Version 2, a DEC-VAX/VMS specific upgrade based on indexed-sequential variants of the three-file system, became available in the early 1980s. Version 2 featured significant improvements in the data analysis capabilities through the development of the original Version 1 program GEOM78 into its successor GEOSTAT.[6,7] All programs operated in batch mode with alphanumeric instruction files; bit-screen heuristics were not used in any of the search procedures.

Also in the early 1980s, the in-house activities of the Centre switched to a single integrated file, encompassing all information fields within a directory-controlled, entry-sequential binary archive file (designated ANEW). Check and evaluation procedures were also integrated and work began in 1986 on a phased upgrading of the distributed CSD system. By *CSD system* we mean a master database structured for rapid searching, together with redesigned software showing increasing degrees of integration by comparison with Versions 1 and 2. In particular, the terminology recognizes that upgrades of *both* the database itself *and* of the associated software are necessary to provide major improvements in system functionality and speed of response.

This paper then describes developments in both the database and in the associated software that have led to Version 3 (first released in January 1988) and Version 4 (July 1989) of the CSD system. We conclude by summarizing the further upgrades that are now being undertaken to achieve the next level of integration, to be embodied in Version 5 planned for release early in 1992. We begin, however, with a brief overview of current methods for database building, including details of information content and a short statistical summary.

THE CAMBRIDGE STRUCTURAL DATABASE

Information Content. The CSD stores the primary results of full three-dimensional X-ray and neutron diffraction studies of organics, organometallics, and metal complexes having organic ligands. The database is fully retrospective (earliest reference 1930) and is updated regularly by some 700 new structures (entries) per month. The primary literature is abstracted, together with any associated supplementary (deposited) data. The CSD itself acts as a computerized depository for large-volume numerical results for some 30 journals. A total of 584 primary sources are now referenced in the CSD, of which 74 are regularly scanned in-house to provide ca. 80% of current input. Remaining references are located via a scan of secondary sources, particularly *Chemical Abstracts*.

Each entry in the CSD relates to a specific crystal structure determination of a specific chemical compound. Each entry is identified by a CSD reference code (REFCODE). This consists of eight characters: the first six are alphabetic and identify the chemical compound (initially assigned as a mnemonic of the compound name, now generated automatically for new compounds), the last two characters are digits which trace the publication history and define (a) whether the paper is a republication by the same authors (perhaps reporting an improved coordinate set) or (b) whether the paper is a redetermination by a different set of authors.

The information recorded for each entry may conveniently be categorized according to its "dimensionality", as described below and illustrated in Figure 1.

1D information consists of bibliographic and chemical text strings, together with certain individual numeric items: compound name(s); molecular formula; authors names; journal name, coden, and citation; qualifying phrase(s) (e.g., neutron study, absolute configuration determined, etc.); text comment on the experimental method used (if unusual), errors located in the printed publication, and notes on any disorder in the crystal structure; chemical class assignment(s) (in the range 1–86: e.g., 15 = benzene nitro compounds, 51 = steroids, etc.); indicators of the precision of the diffraction experiment; flags which summarize the CSD data evaluation process.

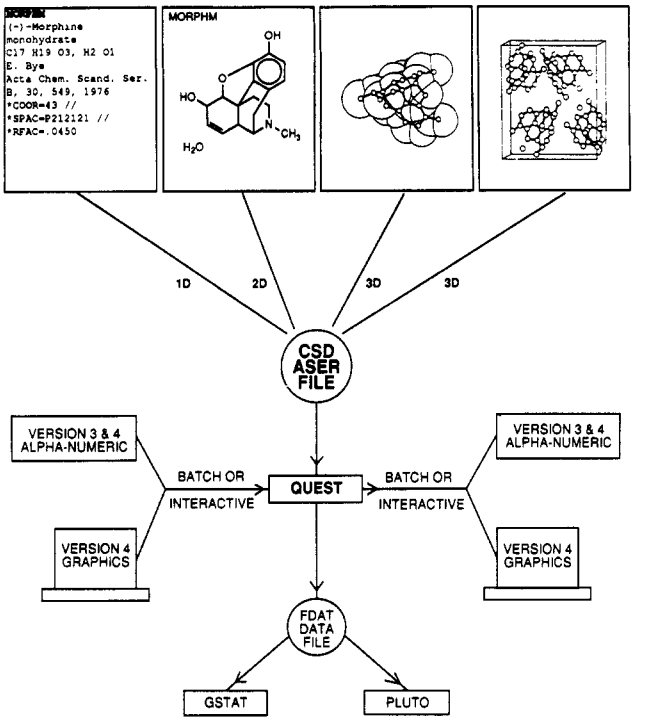


Figure 1. Flowchart for Versions 3 and 4.

2D information consists of the chemical connection table encoded in terms of atom and bond properties. Atom properties are as follows: atom sequence number (n); element symbol (el); number of connected non-H atoms (nca); number of attached terminal H atoms (nh); net charge (ch); coordinates (x,y) for graphical output of the 2D chemical structural diagram. Bond properties are as follows: pair of atom numbers connected by the bond (i,j); bond type (bt) for the bond i–j. Available bond types are single (bt = 1), double (2), triple (3), quadruple (4), aromatic (5), polymeric (6), delocalized double (7), and π (9); bt is coded negative if the bond forms part of a cycle. This wide range of bond types is necessary to cover the broad spectrum of compounds entering the CSD. It is likely that other types describing hydrogen bonds and short nonbonded interactions will be added in the near future. Since 1981, the chemical connection tables have been derived from digitized diagrams entered graphically by the CCDC staff. This provides for direct input of the x,y coordinates noted above as well as the atom and bond property information. Coordinates for connectivity tables entered before 1981 (some 35 000 entries) have been generated either by an algorithm (followed by a graphical edit in many cases) or by redigitization of the diagram. The bond-type cyclicity flags are added by program.

3D information consists of the numerical crystallographic results: unit cell parameters; space group symmetry; atomic coordinates (crystallographic fractional x,y,z) for the bonded crystal chemical unit; covalent radii and crystallographic connectivity established by use of those radii. The crystal chemical unit may consist of one or more molecules or ions, termed residues in the CSD. The coordinate set stored will represent complete chemical moieties, since any symmetry-generated atoms bonded to the crystallographic asymmetric

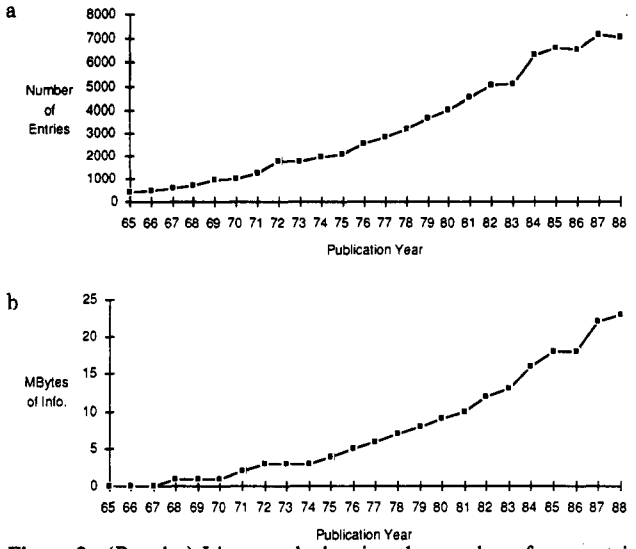


Figure 2. (Panel a) Line graph showing the number of new entries added to the CSD versus publication year. (Panel b) Line graph showing the increasing data content of the CSD using megabytes of information versus publication year.

unit are included in the CSD. The matching of the chemical and crystallographic connectivity representations forms an additional check on the data and on our input procedures; it forms the basis for significant software developments described later in this paper.

Database Building. Raw input is identified and abstracted from each printed source document in three streams that correspond almost exactly with the subdivisions above. The free-format ASCII character files forming the three streams are merged on a refcode basis. The following procedures are then employed in creating the corresponding binary entry in the master CSD archive file, designated ANEW.

Check and Evaluation. These procedures are designed to ensure the accuracy of each CSD entry. This includes visual scans of text fields, together with programmed internal consistency checks. Thus, the related chemical formulation, chemical connectivity tables, unit cell dimensions, and cited density should all be consistent. The unit cell dimensions, atomic coordinates, and symmetry operators should together generate bond lengths which are in agreement with those cited by the authors in the publication, and which are input by the CCDC staff for this evaluation process.[3] The matching of chemical and crystallographic connectivity representations, noted above, forms another powerful evaluation tool. The checks are performed not only to ensure the accuracy of CCDC keyboarded input but also the correctness of the data presented in the relevant publication. Some 15% of all publications contain one or more numerical errors. The raw input file is continually edited and rechecked, and evaluation flags and text comment are added during the process.

Registration. Each new CSD entry is registered against the current master archive file. The crystallographic unit cell dimensions, the chemical formula, and a key derived from the chemical connection table are used to find identities with existing entries. These identities are examined and the initially assigned CSD refcode is altered appropriately before the entry is archived.

Archiving. The check, evaluation, and registration steps result in a considerably upgraded version of the initial raw free-format ASCII file. This data is then converted to binary format, and certain data items are added at this stage (e.g., the bond cyclicity flags, a set of numbers which link the chemical and crystallographic connection tables, etc). The binary entries are then merged with the master archive file. Further processing of this file to the exported search file format is described in a later section.

Table I. Generic Chemical Class Allocation for the Primary Residues in the January 1991 Database Entries

classes	generic coverage	% of database
1–12	simple aliphatics	5.6
13–23	monocyclic hydrocarbons	5.4
24–31	polycyclic hydrocarbons	4.3
32–42	heterocycles	16.4
43–59	natural products	12.9
60–61	molecular complexes, clathrates	2.6
62–70	main group compounds	11.0
71–75	organometallics (Tr-C and π)	16.5
76–86	metal complexes	25.4

Database Statistics. The CSD release of January 1, 1991, contained 86 026 structural entries relating to 76 220 unique chemical compounds. Of these, 87% had 3D atomic coordinates present (the remainder are conference reports, short communications, etc.), and the error rate after evaluation was only 2.4%. The CSD now records 3D coordinates for more than 3.7 million atoms. The rate of increase in size of the CSD is illustrated in two ways in Figure 2. The obvious plot of the number of new entries added by year of publication (Figure 2a) is accompanied (Figure 2b) by a plot of the number of megabytes of information in the database for each publication year. Not only is graph a increasing, but the mean size of each individual entry is also increasing, since more complex structures are now being solved, leading to an enhanced upward slope in graph b. The amount of information processed (in megabytes) during 1987 was approximately equal to the total for the decade 1965–1974, and almost double that processed in 1981.

The range of chemical compounds contained in the January 1991 release of CSD is illustrated in Table I. Here the database is broken down into broad groupings on the basis of the primary chemical class assignments. Over the past few decades there has been a rapid upswing in the numbers of metalloorganic structures determined and these now account for some 45% of file content. Of the remainder, 45% are pure organics and 10% are compounds of the main group elements.

There has been a gradual improvement in the experimental precision of entries in the CSD over the years, as judged by the crystallographic discrepancy index (R). The mean R factor is now ca. 0.05, while some 78% of CSD entries have R below 0.10, and these data would be judged as satisfactory for most applications in molecular modelling.

OVERVIEW OF CSD SYSTEM DEVELOPMENT

Within the context of a small software team, some of whom had other responsibilities within the Centre, a phased approach to system development was a necessity. In this section we summarize the operation of Versions 3 and 4, resulting from the first two development phases. Figure 1 depicts the system flow for both versions.

The central feature of Version 3 is an integration and extension of the functionalities of the earlier BIBSER and CONNSER programs. The new program QUEST (Figure 1) operates on an integrated master file (ASER) to perform searches of the 1D (individual text and numerical items) and 2D (chemical connection tables) information fields summarized in the previous section. ASER is a version of ANEW that has been restructured for rapid searching; bit-screens covering text and chemical connection tables are included in the CSD system for the first time. Version 3 QUEST is driven by an alphanumeric query language and operated initially in a noninteractive manner. The program generates a number of output subfiles corresponding to the hits located in a search. In particular, the FDAT file is preserved from earlier versions as an interface to the 3D graphics program PLUTO and to the 3D search and

data analysis program GSTAT (a further extension of the Version 2 program GEOSTAT). Apart from providing ASER subsets for later in-depth searching, QUEST can also generate the earlier FBIB and FCON files to satisfy the requirements of users with their own software investment in these file structures. Version 3 was first released in January 1988.

Within the phased plan, this initial system upgrade was to be followed as rapidly as possible by a fully interactive version of QUEST which would permit (a) graphical input of substructural, text, and numeric queries and (b) the graphical output of 2D chemical diagrams for the hits. Of equal importance in the development of the new system, designated Version 4, was a major upgrade of the database itself to include the x,y-plot coordinates of atoms in the chemical connection tables (see previous section). These data permit the rapid generation of high-quality chemical diagrams for the very broad spectrum of chemical compounds present in the CSD. All other features of Version 3 are subsumed in Version 4. In the event, the database and software upgrades became available for release in July 1989.

With Versions 3 and 4 well established in the user community, we are now working on the phased upgrade to Version 5. Here, the 3D search capability of GSTAT and the 3D graphics capability of PLUTO are integrated with the 1D and 2D search and display modules of Version 4 QUEST. The numerical and statistical analysis functions of GSTAT will then form a separate program, linked to Version 5 QUEST by a simple interface. Again, the enhanced functionality of Version 5 will be underpinned by a major database upgrade: the matching of the chemical and crystallographic connectivity representations. The importance of this database upgrade, and further details of Version 5, will be presented in later sections of this paper.

VERSION 3: THE ASER FILE AND QUEST PROGRAM

Design Criteria. At the outset of development in 1986, the following criteria for the Version 3 system were identified:

(a) Portability: the file and software should accommodate the varied computing facilities available to CSD users. It was decided to write the new software in (and convert existing modules to) FORTRAN77. This has the merit of being widely understood and contains improved character-handling facilities by comparison with earlier releases of the language.

(b) Flexibility: the software should be able to search (test) all 1D and 2D information items and permit any Boolean logical combination of individual tests (within one level of parentheses) to form the complete search question.

(c) Efficiency: apart from a careful design of the search file to permit rapid reading, there was an obvious need for bit-screen heuristics to enhance search speeds.

(d) Compatibility: the query language for the new QUEST program should, as far as possible, be compatible with that used in earlier systems, particularly in terms of the mnemonics used to identify search fields.

The Search File (ASER) Structure. Criteria a–c above are difficult to satisfy simultaneously in a system that is intended to be general. Direct access, indexed-sequential, and fully inverted file structures can significantly improve efficiency, but all imply some degree of machine specificity. In the case of inverted files,[8] dramatic improvements in search speeds must be traded against increased file preparation time and disc storage requirements. There is also a loss of flexibility: the system can only interrogate those information fields which the designer considers useful as search terms.

These considerations led us to develop an entry-sequential file (ASER, Figure 1), which is machine independent in its basic form. The structure involves three logical records per entry:

SCREEN: a short fixed-length record containing essential numerical integer information that is mandatory for all entries, together with the bit-screens to be described in some detail below.

TEXTCONN: a variable-length record containing the searchable information of primary interest to Versions 3 and 4, namely the bibliographic and chemical text fields and the chemical connection table.

DATA: a second variable-length record containing the bulk of the 3D numerical information for the entry, i.e., symmetry operators, atomic coordinates, etc.

Additionally, a single "file header" record identifies both the file type (version number) and the information contained within the file. The structure is a flexible one: new information items may be added easily (within the limitations of the directories), and the information content can be tailored for a particular application. More importantly, the TEXTCONN and DATA records can be converted to direct-access or indexed-sequential structures to enhance efficiency on specific machines. Implementations for DEC-VAX computers under VMS and for Silicon Graphics Workstations under Unix are described in a later section.

The QUEST Program. The search strategy applied in QUEST is a simple one, common to many chemical databases. The search query is decoded, and bit-screens are generated for any search specifications involving the text or chemical connectivity records. These are added to any specific search requirements based on the mandatory integers to form a SCREEN record for the complete query. This record is then matched against the content of the incoming SCREEN record for each database entry. The TEXTCONN record needs only to be read for those entries where there is a complete SCREEN match; a detailed text and/or substructure match is then attempted. The DATA record need only be read if the detailed search succeeds, so that full information for the hit can be generated. Given that the heuristics (see below) allow us to skip TEXTCONN and DATA in $\gg$95% of cases, then there is a very significant gain in efficiency, even for the entry-sequential file. For the machine-specific implementations, the need to skip these variable-length records disappears completely and efficiency is further increased.

The Query Language. A search within Version 3 is defined by a set of alphanumeric instructions which define a single query. The complete query consists of a series of TESTS, each specifying a particular information field. The search is initiated by a single QUESTION instruction, which defines the logical combination of the individual TESTS. Thus

```
T1 *CLASS  51
T2 *YEAR  1986
T3 *AUTHOR  Smith
T4 *AUTHOR  Jones
QUES  (T1 .AND. T2) .AND. (T3 .OR. T4)
```

would locate those entries in CSD chemical class 51 (steroids), published in 1986 by either Smith or Jones. The paragraphs below give further summary details of the types of TEST that can be constructed with QUEST for various categories of information.

Numerical Tests. There are 38 possible individual numerical search fields. Test values may be specified within the general construct:

```
Tn  *FIELD  .lo.  nval
```

where .lo. is one of the logical operators: .eq. (default), .ne.,

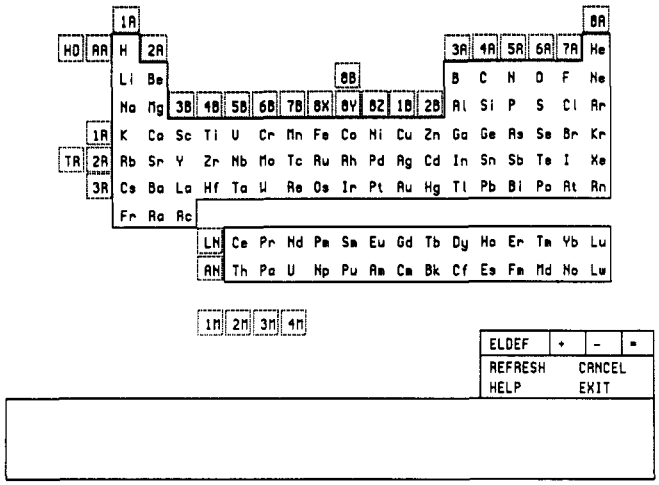


Figure 3. Periodic Table indicating the 30 predefined element group symbols available in the CSD system. The element groups 1M–4M are various subdivisions of the metallic elements.

.lt., .le., .gt., and .ge. Specification of a numerical range is also permitted as

```
Tn  *FIELD   nval1 - nval2
```

Crystallographic Unit Cell Tests. This is a variant of the numerical tests, requiring specification of the crystal system, the six unit cell parameters (a, b, c, α, β, γ), and a tolerance to be applied during cell edge testing.

Text Tests. There are 19 possible text fields that may be searched within the general construct:

```
Tn  *FIELD   string
```

A special feature is provided for searches of compound names and their synonyms. These fields may be searched separately (as *COMP, *SYNO), or together under the composite specification *XNAM. Here the two names are first stripped of all of their nonalphabetic characters (i.e., blanks, punctuation marks, numbers) and then concatenated. This means that the search string cyclobutanedione will hit [1,2]cyclobutanedione *and* cyclobutane[-1,2-]dione (where the characters removed in constructing the *XNAM field are shown in brackets). This feature removes some, but by no means all, of the possible vagaries of compound name searching.

Chemical Formulation Tests. The standard molecular formulae of *either* the discrete bonded residues or ions contained in the crystal structure *or* the formula summed over all residues may be searched for the presence of specified elements (by symbol), or element/count combinations. The test

```
Tn  *RESFORMULA   Os 8 - 10
```

is typical and would locate a bonded unit containing between eight and ten osmium atoms. These searches may be generalized by use of a predefined set of 30 element-group symbols within QUEST. Figure 3 shows the vertical and horizontal subdivisions of the periodic table that constitute these group definitions.

2D Chemical Substructure Tests. A chemical substructure is specified by a "packet" of instructions commencing with Tn *CONN and terminating with an END record. As with other TEST's, multiple use may be made of the Tn *CONN construct, e.g., to locate entries that contain substructure A but not in the presence of substructure B. The system is extremely flexible, both in the definition of the required substructure and in the definition of its environment. The permitted instruction set is summarized in Table II, while a complete example, together with explanation of some specific features, is shown in Figure 4.

Table II. Alphanumeric Instruction Set for 2D Substructure Search Definition within CSD System Versions 3 and 4

(a) Main Keywords and Summary Definitions	
`Tn *CONN`	Start of instruction 'packet'
`C [text]`	Optional user comment on query
`ELDEF`	User-definition of 'element-group' symbol and group composition (see Text)
`ATp`	Definition of atom properties [see (b) below]
`BO`	Definition of bond properties [see (c) below]
`NFRAG [nval]`	Number of occurrences of fragments required
`SAMERES`	Locate fragments in same bonded residue
`DEFB [nval]`	Reset default bond type from single to `[nval]`
`NOCR [nval(s)]`	The defined fragment may not have substituents which form part of a cyclic route (see Text), `[nval(s)]` restricts `NOCR` to operate at the atom(s) specified, otherwise it is global.
`NOLN [nval(s)]`	No direct links shall exist between atoms in the fragment, except those defined by the BOnd records (see Text), `[nval(s)]` restricts `NOLN` to operate at the atoms specified, otherwise it is global.
`ALLBOND [A or C]`	All bonds in fragment are to be acyclic (A) or cyclic (C)
`END`	End of instructions for this test
(b) ATom property definitions	
`Element Symbol`	Individual symbol, pre-set group symbol, user-defined group symbol
`Mca`	Minimum number of connected atoms excluding terminal H atoms. Tests on `Mca` can be avoided by using a value of 99. The 'minimum' criterion can be altered by use of: the 'minimum' criterion can be altered by:
`E`	The `Mca` value must be matched exactly
`Nh`	Number of terminal H atoms to be attached to this atom, can be an 'OR' list as n1,n2, n3, etc.
`Nch`	Integral charge assigned to atom, can be a single +ve or -ve value, an 'OR' list, or special values: 49 (any -ve), 50 (any +ve or -ve), 51 (any +ve).
`Tn`	Total coordination number (see Text)
`C or A`	Atom must be part of a cyclic or acyclic unit.
(c) Bond properties	
`n1 n2`	Defines a connection between AToms n1 and n2, n2 may be an 'OR' list to allow for variable point of attachment of one subfragment to another (see Figure 4).
`Bond type`	May be a single +ve number, an 'OR' list, or set to 99 to allow any bond type. Assumed to be single if not set. Allowed values are : 1 (single), 2 (double), 3 (triple), 4 (quadruple), 5 (aromatic), 6 (polymeric, as in catena-structures), 7 (delocalized double), 9 (π-bond).
`C or A`	Bond must be part of a cyclic or acyclic system.

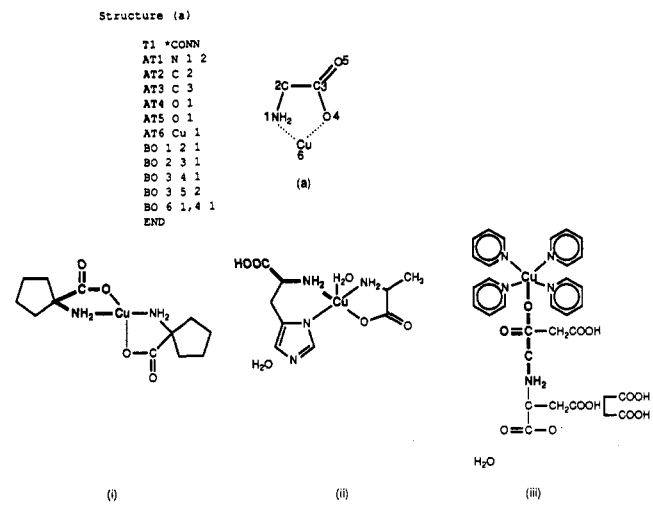


Figure 4. Complete example of a *CONNser: Variable Point of Attachment (VPA) option. The VPA option enables a search structure to be set up where there is uncertainty as to where one of the atoms (or groups) involved in the structure may be bonded. In example (a) shown here, the Cu atom (AT6) may be bonded to either the -NH_2 group (AT1) or the -O atom (AT4) by a single bond. This means that not only structure (i) will be hit but also structures (ii) and (iii).

The basic topology of the fragment is specified by use of ATom and BOnd records, as in Versions 1 and 2 (see Figure 4). Chemical constraints are applied to this topology by use of subkeywords (Table IIA,B). Within Version 3, we have added three new features to improve fragment definition:

(a) A set of 30 predefined element group symbols (Figure 3), together with AA = any non-H atom, may

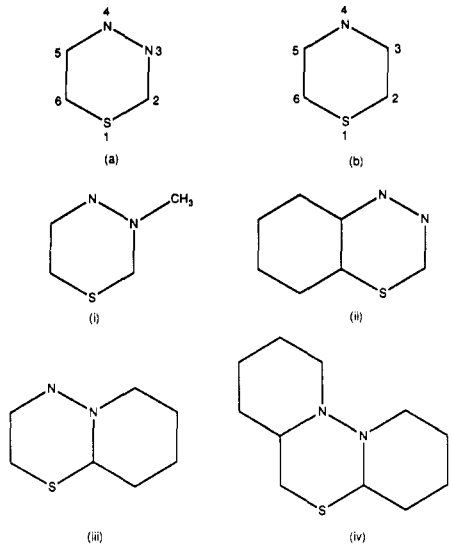


Figure 5. Diagrams to show use of No Cyclic Routes (NOCR) option. The NOCR instruction can be used to restrict the environment of the search fragment. It can take two forms: NOCR requires that no atoms of the search fragment can be connected to atoms outside the fragment by cyclic routes (bonds). NOCR ijk... requires that atoms *i, j, k*... of the search fragment cannot be connected to atoms outside the fragment by cyclic routes (bonds). Consider the search fragment (a) above, with target structures (i)–(iv). NOCR: This will register hits only for (i). NOCR 3: This will register hits for (i) and (ii). NOCR 6: This will register hits for (i), (iii), and (iv). NOCR 4 5: This will register hits for (i) and (iii). Suppose the search fragment was changed to (b). This has topological symmetry. Therefore, if you wish to apply NOCR to atom 2, it must also be applied to atom 6, i.e., NOCR 2 6.

be used in element definition in the ATom records.

(b) The ELDEF record permits the user to define a new element group symbol and the constituent elements of that group. The constituent elements may be standard element symbols, any of the predefined group symbols from Figure 3, or any user-defined group set up by previous use of the ELDEF command. Thus the sequence

```
ELDEF  Pk = N,O,S,7A  (7A = any halogen : Figure 3)
ELDEF  Jk = Pk,Tr     (Tr = any transition metal : Figure 3)
```

establishes Pk,Jk as two user-defined symbols for use in fragment definition.

(c) The Tn subkeyword on ATom specifies (n) as a TOTAL COORDINATION NUMBER, i.e., it invokes a test that the sum of nca + nh (Table IIB) is equal to n. In the case of carbon, oxygen, etc., this is equivalent to the specification of hybridization for C, T4 = $C(sp^3)$; T3 = $C(sp^2)$, etc.

The other keywords of Table IIA are generally self-explanatory. Some are new to Version 3, while the definitions of NOCR and NOLN have been improved over earlier versions. NOCR is used to exclude fusion or bridging of the defined fragment and was previously defined in a global sense; this meaning remains, but numerical qualifiers can now restrict this exclusion so that it applies at specified AToms only (Figure 5). NOLN requires that the links specified on the BOnd records shall be the *only* links permitted between the atoms of the defined fragment. Numerical qualifiers are now allowed for this keyword also to permit ATom specificity (Figure 6).

Global (Main) Keywords. These keywords control the overall operation of QUEST during a given search. They define, for example, the style and information content of the text output for each hit, the subfiles (see Figure 1) to be created, etc. The main keywords also control retrieval of information corresponding to a previously known set of CSD reference codes and, within the later interactive version, access to an on-line HELP document.

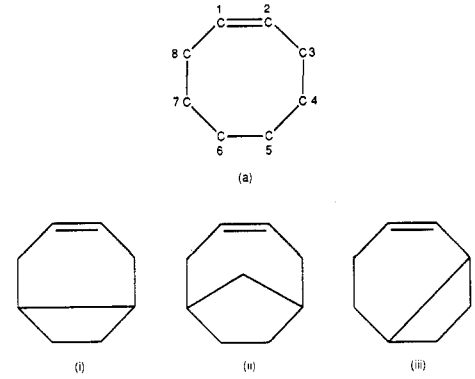


Figure 6. Diagrams to show use of No Direct Links (NOLN) option. The NOLN instruction can be used to place restrictions on the search fragment. It can take 2 forms: NOLN requires that no atoms of the search fragment can be directly linked to any other atoms of the search fragment except by the bonds specified in the bond property records. NOLN ijk... requires that atoms *i, j, k*... cannot be directly linked. Consider the search fragment (a) above, with target structures (i)–(iii). NOLN: This will register hits only for (ii). Atoms 4 and 7 are connected by an indirect link through a bridging C atom. NOLN 4 7: This will register hits for (ii) and (iii). NOLN 3 8: This will register hits for (i), (ii), and (iii). NOLN 3 6 8 5: This will register hits for (i) and (ii). Note that the search fragment has topological symmetry. Thus a link 3–6 would be equivalent to a link 8–5. Therefore, to exclude (iii) we must code both possibilities, i.e., NOLN 3 6 8 5.

THE SCREENING SYSTEM

Bit-screen and other heuristics, included in CSD systems for the first time in Version 3, are added to ASER during its derivation from the in-house archive file ANEW. A single-word numerical key together with bits 0–30 (sign bit avoided) of a further 22 words form a compact knowledge representation for each entry. The heuristics relate to (a) entry content (4 words), (b) compound and synonym names (2 words), (c) author surnames (1 word), (d) elemental composition (1 word, numerical key), (e) element groups (1 word), and (f) 2D chemical structure (14 words).

There is a considerable literature[9] on the application of screens in chemical database systems. In particular, it is desirable that descriptors selected for bit-encoding should be generated algorithmically[10-16] and that these descriptors should be as mutually independent as possible. Further, it is desirable that each bit should be assigned for an equal percentage of entries in a large file.[10-16] In practice, equifrequency of screen assignment is impossible to achieve, and a number of algorithms[14-16] have been designed to yield the best approximation to this criterion. In generating screen sets b–f above, we have occasionally modified these principles in the light of other considerations. These cases are highlighted in the ensuing paragraphs.

Entry-Content Screens. These 124 bits record the presence of specific items of information in a given entry. They include details of (i) any errors located in the original publication; (ii) special features of the experiment, e.g., neutron study etc.; (iii) the precision of the diffraction experiment; (iv) the crystal system, lattice type, etc.; and (v) the presence of nonmandatory information fields in the ensuing TEXTCONN and DATA records. Screens in this set differ radically from those in sets b–f above. They are merely a compact way of recording knowledge that is abstracted directly from the original publication or derived during data processing for each entry. They cannot be derived from the query itself but must be accessed directly by the user via the global keyword SCREEN, or a TEST construct Tn *BTEST, in the QUEST program. In

Table III. Efficiency of Text Screens in CSD System Versions 3 and 4[a]

query string	N_{pass}	S (%)	N_{hits}
(A) Compound Name Screening (62 bit hash code)			
pu	29055	57.3	186
erb	19423	71.5	216
holm	6790	90.0	42
mycin	5146	92.4	221
camphor	3831	94.4	59
butadiene	1327	98.1	205
caryophyllene	358	99.5	12
tetracyanoquino	1670	97.6	321
tetracyanoquinodimethane	563	99.2	168
(B) Author Surname Screening (29 bit hash code, 2 length bits)			
Mo	2648	96.1	26
Sim	1061	98.4	185
Duax	5258	92.6	187
Allen	6583	90.5	93
White	4188	95.4	1060
Watson	6615	90.3	305
Struchkov	3338	96.7	1095
Irngartinger	789	99.0	112

[a] The statistics were generated by using an entry-sequential ASER file containing 68 068 entries. N_{pass} is the number of entries that passed the screening process. S (%) is the percentage of entries eliminated by the screening. N_{hits} is the number of real hits resulting from the detailed character-string matching process.

practice, screens in this category are extremely powerful in restricting searches to, for example, entries with an acceptable level of experimental precision, entries which are error-free in data checking, etc. Thus, specification of a single bit (e.g., SCREEN 88) will immediately eliminate the 26% of CSD entries for which the crystallographic R factor exceeds 0.10.

Text Screening. There are 19 text fields in the CSD, many of which are relatively unstructured and contain free-text comment relating to the crystal structure determination and/or to errors located during data processing. This information is presented to the user if the entry is designated as a hit in a particular search, but is very unlikely to form the basis for a search in itself. Thus, we choose to screen only those text fields that are structured and are likely to be the subject of frequent search TEST's. The compound name and synonym (trivial) name together with the list of author surname(s) fall into this category.

The technique of hashing, commonly used in database search and artificial intelligence applications,[17-19] is applied to letter pairs to generate the text-screen heuristics. Each letter is assigned a numerical value (A,a = 0, ..., Z,z = 25), whence the digram $K = 26i + j + 1$ (i,j are the numerical values for two sequential letters in the text string) represents the key (K) for the letter pair. Nonalphabetic characters in the field are eliminated in this calculation. In this case, a simple hashing function is $H = \mathrm{MOD}(K,x) + 1$, which yields a value for H in the range 1 to x, so H can be taken as a bit position in a bitmap of length x. Obviously if $x = 676$, each value of K is assigned to an individual bit position, the bit map will be sparse and unevenly filled across the database.

After some experimentation, the digrams generated from the compound and synonym fields (taken together) were hashed to a bit vector of length $x = 62$ (2 words). This means that 10 or 11 letter pairs share a bit assignment. Across a database of 86 026 entries, bit-assignment frequencies ranged from 4% (accidental collision of 11 very rare letter pairs) to 90% (the bit to which the letter pair YL is mapped!); 46 of the 62 bits have occupancies in the 25–75% range. Screen efficiency, indicated in Table IIIA for search strings of 2–24 characters in length, exceeds 90% screenout at the 4-character level.

For author names, we initially selected a 31-bit hash vector that showed acceptable efficiency for search strings of 6 characters or higher. For shorter names, screenouts of less than 50% were obtained. The hash vector was reduced to 29 bits (22 or 23 letter pairs per bit), and the last two bits were assigned if the surname was 2 or 3 characters in length (bit 30) or 4 or 5 characters (bit 31). These extra bits compensate for the very low number of bits set by very short surnames and generated the screenout characteristics indicated in Table IIIB. The equifrequency criterion is well approximated: 23 of the 29 bits have occupancies ranging from 35 to 60% across the complete database.

Table IV. Element Frequencies in the CSD[a]

	N_{ent}	freq (%)		N_{ent}	freq (%)
1 C	54566	100.00	17 Pt	1562	2.86
2 H	54070	99.09	18 Si	1312	2.40
3 O	40910	74.97	19 Rh	1084	1.99
4 N	31357	57.47	20 Cr	966	1.77
5 Cl	11285	20.68	21 W	890	1.63
6 S	10835	19.86	22 As	872	1.60
7 P	8778	16.09	23 Mn	866	1.59
8 Br	4009	7.35	24 Ru	868	1.59
9 F	3226	5.91	25 Pd	808	1.48
10 Cu	2936	5.38	26 Sn	777	1.42
11 Fe	2613	4.79	27 K	707	1.30
12 Co	2371	4.35	28 Na	707	1.30
13 B	2032	3.72	29 Hg	676	1.24
14 Ni	1984	3.64	30 Zn	569	1.04
15 I	1928	3.53	31 Os	546	1.00
16 Mo	1861	3.41			

[a] Statistics are generated from a CSD file of 54 566 entries, only the top 31 elements having frequencies greater than 1.0% are tabulated.

Element Screening. Because of the very wide spectrum of chemical compounds covered by the CSD (organics, organometallics, metal complexes), 85 of a possible 104 elements (including deuterium) are represented there. Only the 31 elements of Table IV have occurrences that exceed 1.0%. Element screening is, therefore, highly effective in eliminating unwanted entries for any search (chemical constitution and/or substructure search) that cites an element outside of the top 10 or so occurrence values. For maximum specificity, we could assign each element to a single bit in a vector of length 104, leading again to a sparsely filled bit map with a highly skewed distribution. An attractive solution is to reduce the vector length by allocating increasing numbers of elements to specific bits as we travel down the occurrence table; by this means we approximate the equifrequency criterion. This approach has some drawbacks within the (basic) entry-sequential ASER file structure. A specific search for a rarer element, sharing a single bit with perhaps five others, will still require an expensive read of the TEXTCONN record, albeit for a relatively small number of entries. Thus, we sought an alternative solution, one that allowed as many specific element searches as possible to be satisfied by the content of the SCREEN record alone.

The approach adopted involves prime number packing to a single word. Each element is assigned a prime number value (P_i) that increases with decreasing occurrence of each element, i, in the CSD; carbon and hydrogen are both assigned a value of 1. The prime number key (K_p) is then simply:

$$K_p = \Pi_i P_i$$

Thus, for an elemental composition of C(P_i = 1), H(1), N(3), O(5), Br(17), Os(113), the K_p value is 28 815. To test if N(3) .AND. Os(113) are present, we calculate $K_q = 3 \times 113 = 339$ and test that $\mathrm{MOD}(K_p/K_q)$ is equal to 0. As the number of elements in a given entry increases, and particularly if the number of rarer (higher P_i) elements increases, it is obviously possible for K_p to exceed the maximum single-word integer value. In these cases (only 37 in the current file of 86 026 entries), the key is filled as far as possible and then set negative. Thus, it is only in very rare cases that an element search cannot

be fully satisfied by use of the SCREEN record alone.

The predefined element groups of Figure 3 are an extension of a popular feature for users of Versions 1 and 2, particularly those who are interested in compounds containing metals. No attempt is made here to accommodate the equifrequency criterion: each logical subdivision of the periodic table is mapped by a single bit assignment. Bit 30 (hydrogen or deuterium) satisfies a common user problem of remembering that CSD treats D as a separate element; bit 31 (H) compensates for the assignment of a nonunique prime number in the key described above. In practice, only 8 of the remaining 29 bits are occupied for more than 10% of the entries: those corresponding to groups 5A (N present), 6A (O), 7A (F, Cl, Br), and some of the broader metallic subdivisions. Note that element group bits can sometimes be set for substructure search queries in which ELDEF (see above) has been used to establish a "user-group" definition, provided that the members of the user group are a subset of one of the predefined element groups.

2D Chemical Structure Screens. A total of 428 bits are allocated to descriptors derived by computer analysis of the 2D chemical connection tables of the CSD. Eight bits have been left undefined for possible later additions. The vast bulk of the descriptors refer to atom, bond, or ring-centered fragments, as in the classic work of Lynch et al.[10-13] Attention has focused primarily on fragments involving C, N, or O for obvious reasons. However, six other element "supergroups" are also used in some of the simpler descriptor definitions: X1 (not C, H), X2 (not C, H, transition metal, lanthanide, actinide), X3 (as X2 with N and O also excluded), X4 (as X3 with halogens also excluded), M (any metal), and group 7A (the halogens alone). Similarly, supergroups of bond types are sometimes employed, especially for unsaturated bonds of types 2 (double), 7 (delocalized double), 3 (triple), 4 (quadruple), and 5 (aromatic).

The descriptor generation process must be regarded as semialgorithmic only. It was guided by many preliminary statistical analyses of file content, as well as the following general considerations:

(i) The CSD query language permits substructure definition with a very wide spectrum of specificity, due to its ability to handle variable specifications of element type, bond type, number of connected non-H atoms, number of connected H atoms, etc. The use of AA (for any non-H atom) or 99 (for any bond type) is also common. Thus, the descriptors must also reflect increasing degrees of specificity so that a minimal subset of screens can be set for even the most general search fragment.

(ii) Experience has shown that certain apparently uncommon substructural features are frequent targets for user searches, for example, a search for two metal atoms connected by a quadruple bond is satisfied by only 255 of the 86 026 entries. This is not surprising; crystal structure determinations are often performed specifically to characterize novel or rare chemical features. Most users are aware of this, and searches for the uncommon are commonplace.

(iii) The specificity of a given screen assignment must be balanced against the relative speed of the ensuing atom-by-atom, bond-by-bond subgraph matching process. The routine included in QUEST is known to slow considerably in searches for ring systems as the size of the ring increases. To compensate for this, some highly specific ring screens, extending up to sizes of 18 atoms, are included.

Within these general observations, some sections of the screen generator are genuinely algorithmic, while other sections might be regarded as corresponding to a computerized "manual" selection of special features. No attempt has been made to map the descriptors in conformity with the concept of equifrequency, which of itself implies bit sharing. However, the use of atom-type and bond-type supergroups in descriptor definitions represents an .OR.ing of a number of more specific descriptors, and hence implies that a degree of bit sharing is implicit in the complete process.

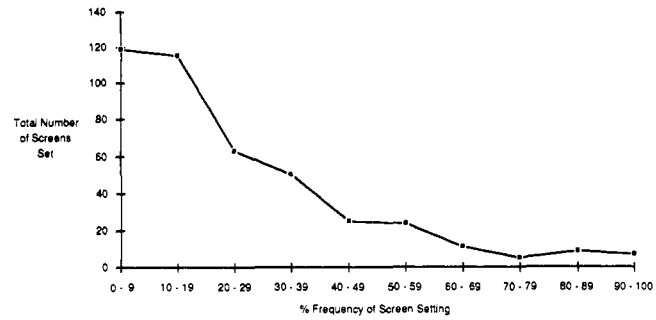


Figure 7. Bit-setting frequencies for the 428 substructure search screens.

Figure 7 shows a histogram of the number of bit settings versus the percentage of CSD entries having those settings. Thus, for example, 70 bits are set for less than 5% of the file content, and 333 (of 428) are set for less than 35% of the file content. The very few high-frequency settings (>90% of entries) correspond to such obviously common descriptors as C-C single bonds, a carbon to non-carbon connection (any bond type), etc. Consideration (i) above leads us to retain these descriptors when they might otherwise be deemed redundant.

The complete screen set can be subdivided into a number of classes, relating to:

(1) The element supergroups identified above.
(2) Simple connections X·Y, of any bond type, where X and Y are C, N, O, X1–4, M, and Hal. A few descriptors, chosen statistically, relate specifically to S, P, B, Si, and individual halogens.
(3) Bond types, both individually and as supergroups noted above.
(4) Bonded pairs, a subset of item 2 with bond type and, in common cases, discrimination on the basis of cyclicity or acyclicity of that bond.
(5) Simple triplets X·Y·Z involving a statistically chosen subset of C, N, O and element supergroups.
(6) Bonded triplets, a subset of item 5 with bond type specified.
(7) Atom-centered fragments around C, N, O, M, and X3 only.
(8) Aromatic–nonaromatic bond sequences indicative of the type of substituent and substituent relationships (ortho, meta, para).
(9) Simple quadruplets W·X·Y·Z involving C, N, and O only.
(10) Larger bond-centered fragments, covering bond connectivities, substitution patterns, bond-type patterns, etc.
(11) Rings and ring-centered fragments. An exhaustive ring-perception algorithm, based on the approach of Wipke and Dyott,[20] is used to classify rings and ring assemblies as (i) open (O), no additional direct connections between members of a cycle and (ii) envelopes (E), circuits around the periphery of fused systems. The sizes, elemental constitution, and bonding patterns in O and E rings are established. The algorithm also identifies atoms, bonds, and rings involved in fusion, bridging, and spiro-ring formation. The algorithm can handle up to 100 rings (O and E) of maximum size 50 atoms. Some 3.5% of the CSD

Table V. Efficiency of Element, Element Group, and 2D Chemical Connectivity Screens in Substructure Searching Process[a]

fragment	N_{pass}	S (%)	N_{hits}	t_1	t_2	t_2/t_1
Tr-Hal-Tr-Hal ring	981	98.5	800	8.3	97.6	11.8
N-C-C-S-C ring	56	99.9	34	5.2	52.9	10.2
$(C)_3$-P-N-$(AA)_2$	182	99.9	12	6.0	56.3	9.4
anilines	1072	98.4	452	6.8	52.0	7.7
any open C11 ring	1055	98.4	42	24.0	204.0	8.5
any open C11 ring with 1 double bond	764	98.8	33	7.5	77.9	10.4

[a] Statistics were generated by using an entry-sequential ASER file containing 65 737 entries. N_{pass} is the number of entries which pass the screening step. S (%) is the percentage of entries eliminated by the screening. N_{hits} is the number of hits after the atom-by-atom, bond-by-bond matching. t_1 is the time (s) for the screened search. t_2 is the time (s) for a nonscreened search. The ratio of these times (t_2/t_1) is also tabulated. Search times are for an IBM-3084 mainframe.

entries exceed these limits, and a single bit indicates this fact when ring screens (the final 62 bits in the map) are ignored for these entries.

Performance of the Screening System. The ultimate test of the screen sets covering elemental constitution and 2D chemical structure is their practical efficiency over a very wide range of substructure search queries. Some indication of efficiency is presented in Table V, which records screenout percentages and comparative timings for screened and unscreened searches for a variety of chemical fragments. The results show a mean screenout of 99% and screened searches are an order of magnitude faster than their unscreened counterparts.

2D CHEMICAL SIMILARITY SEARCHING

The classification of 2D chemical structures on the basis of attributes derived from their connection tables has a history dating back some 20 years.[21] Indeed, calculation of the similarity or dissimilarity of objects is a fundamental prerequisite for a variety of autoclassification and pattern recognition techniques, which employ some form of cluster analysis based on these (dis)similarity measures.[22]

Recently,[23-25] Willett, Bawden, and others have extended and systematized these basic concepts to provide methodologies for similarity (or browsing) searches and for the clustering of similar compounds that are applicable to large databases. For chemical "objects", the bit-encoded screens derived from the connection table represent an ideal set of attributes for comparison. A variety of association coefficients[22-26] exist that provide a quantitative assessment of the similarity between pairs of fixed-length bit-encoded attribute sets. Thus, if the numbers of attributes (from a fixed set) exhibited by objects (molecules) p and q are N_p and N_q, and the number of attributes common to p and q is N_c, then two simple similarity measures are T_{pq}, the Tanimoto (Jaccard) coefficient, and D_{pq}, the Dice coefficient:

$$T_{pq} = \frac{N_c}{N_p + N_q - N_c}$$

$$D_{pq} = \frac{2N_c}{N_p + N_q}$$

We have implemented similarity searching within Version 3 of the CSD system with user choice of T_{pq} or D_{pq}. The connection table for a query molecule is encoded, with a maximum specificity of atom and bond properties, via the AT and BO records detailed in Table III. These instructions form a "packet" akin to the substructure search specification, but identified to the QUEST program via the global keyword SIMIL. Bit-screen attributes for the query (q) are established in the normal way and form a constant bit map which is systematically compared with the bit maps (p) stored for each CSD entry in ASER. Those comparisons which yield the 100 highest values of T_{pq} or D_{pq} are retained and presented in the user in descending order of the similarity score.

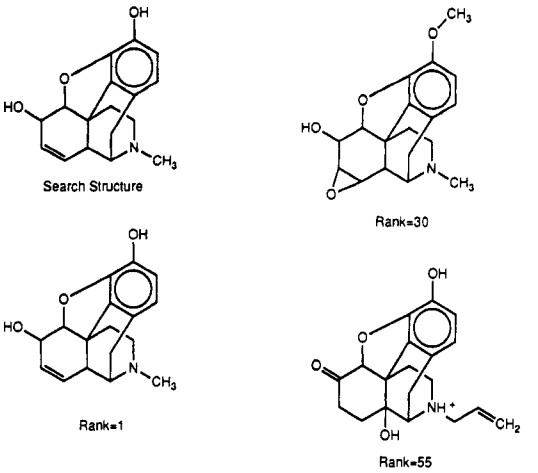


Figure 8. Example of a 2D similarity search within CSD System Version 4.

Obviously, this process is tedious in a system where the exhaustive encoding of the query molecule is alphanumeric and where results are presented in nongraphical (text) form. Similarity searching is designed as a browsing mechanism, and its full value within the CSD system has recently been realized through the graphical input/output facilities of Version 4 described in the next section. In direct-access packages, it is possible to use an existing CSD entry as the query by citing the appropriate CSD reference code. The results depicted in Figure 8 are derived from the Version 4 upgrade.

The application of similarity searching to the CSD presents two problems that are currently being resolved. The first concerns the attribute set, and may well apply to other databases also. The vast majority of CSD screens are defined in chemical terms, very few are purely topological in nature, i.e., refer to subgraphs in which the nodes may be "any atom" and the edges may be "any bond type". Hence a similarity analysis in which the query is encoded with minimal chemical specification is likely to present apparently questionable results to the user. We are currently devising some purely topological descriptors to enhance this type of similarity search, and which will also contribute to the screening mechanism for normal substructure searches.

The second problem is peculiar to the CSD. Crystal structures can refer to a crystal chemical unit that contains more than one discrete molecular species. Each of these different molecules contributes attributes to the overall bit map for the entry, but we have no record of which attribute was derived from which molecular unit. Thus a query, likely to be a single molecular species in most cases, is being compared against a composite bit map for 36% of the CSD entries. In practice, this problem is not as significant as it might seem, since the majority of additional units in crystal structures are small solvent molecules or discrete monoatomic ions. The latter contribute few, if any, screens to the composite set. Only in the case of molecular complexes, where the two moieties may be similar in size, does the problem show itself to its fullest extent. Again, remedial action is being sought.

VERSION 4 OF THE CSD SYSTEM

Development of Version 4 of the CSD system involved both the development of a new software package to allow menu-based searches with the existing QUEST program and the upgrade of the database, described earlier, to include 2D structure

Table VI. Version 4 Graphics Menus and Submenu Relationships

menu	function	submenu	function
BUILD	enables structure and substructure searches to be constructed		
		GROUPS	displays 26 common functional groups
		TEMPLATES	lists 20 template groups—selection of one of these groups produces a sub-submenu displaying their structures; also allows for use of the Feldmann notation for fused ring systems
		CONSTRAIN	displays menu options that can be used to add chemical constraints to a structure
		Periodic Table	allows selection of any element in the periodic table, any CSD group of elements, or preparation of user-defined groups of elements
SEARCH	enables nonstructural searches to be set up and global QUEST options changed		
		TEXT	all textual searches that can be carried out in QUEST
		NUMERIC	all numerical searches that can be carried out in QUEST
		SCREENS	allows screens to be set and used as selection criteria for the whole database, or in combination with other tests
		REDUCED-CELLS	allows crystallographic reduced cell tests to be set up
		PRINT	enables the standard print options for QUEST to be altered
		Periodic Table	enables searches for individual elements or groups
EDIT	enables various types of structure editing		
FILES	allows structures to be stored and retrieved in later sessions		

diagram coordinates for each entry, thus enabling the display of diagrams for hit structures.

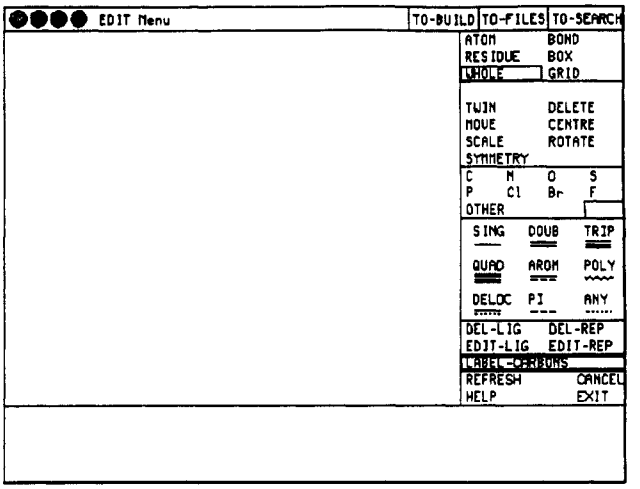
Graphics Menus. The menu-based graphics system is summarized in Table VI. The system is based on four main menus, each of which can be accessed from the other three main menus, and a series of submenus, which can only be accessed from the parent menu and, in turn, can only access the parent menu. An example of a main menu, the EDIT menu can be seen in Figure 9a, and the TEXT submenu is shown in Figure 9b as an example of one of the submenu layouts.

These menus provide a graphical interface to both the structural and the nonstructural search capabilities of the QUEST program described above for Version 3. The main BUILD menu allows structures and substructures to be input using a mouse or cursor keys and a series of graphics commands. The BUILD menu has four submenus: two of these are concerned with speeding the process of structure building by providing a series of predefined chemical functional groups and templates; the third comprises a graphic display of the periodic table (see Figure 3) enabling the selection of any element, a predefined element group or a user-defined element group; the final submenu enables any chemical constraints to be added to the structural skeleton. The SEARCH menu provides the facilities for defining all other nonstructural searches and global commands; its six submenus are described in Table VI. The EDIT menu allows structures to be edited, either in whole or in part. The FILES menu allows structures to be stored and retrieved again in future searches.

A survey of our users showed that the majority had access to either a Tektronix terminal or a good emulator of the Tektronix standard, so this was chosen as the first platform on which to develop the graphics. Recently, the graphical interface has been customized for a second platform, the Silicon Graphics 4D series, as these workstations are being used by an increasingly large proportion of our users.

Version 4 has been implemented on a wide range of Tektronix terminals, from 4010s and 4014s, right through to the 4200 series. On the most primitive of these, the 4010, Version 4 has to function without even the ability to erase lines and text once they have been drawn on the screen. For this reason, a REFRESH button was provided in every menu to allow the screen to be redrawn with the removal of any unwanted graphical or text items. On Tektronix 4100 and 4200 series terminals, Version 4 makes use of the panel drawing capability.

a



b

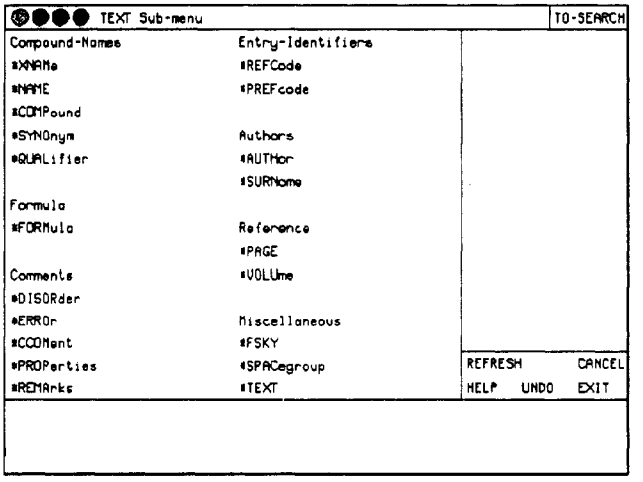


Figure 9. (Panel a) Version 4 EDIT menu. (Panel b) Version 4 TEXT submenu.

Thus when changing menus, for example, the previous menu can be erased without the need to redraw any structure that is present in the drawing area.

On 4200 series, and older, Tektronix terminals, the rate-determining step for the displaying of images is the transfer

of the escape sequences that comprise the Tektronix instructions from the host computer to the terminal. Therefore, any technique which reduces the number of characters sent to the terminal will increase the speed of the display process. Version 4 uses segments of terminal memory to store each menu once it has been displayed, so that the next time it needs to be drawn, it is redrawn from terminal memory rather than having to be transmitted again from the host. From Version 4.4, all Tektronix terminals buffer the escape sequences into lines of up to 80 characters in length, thus reducing the number of carriage returns and line feeds that are set. This has produced a further noticeable increase in the speed of the graphical output in Version 4.

Operation of the Interactive System. The diagram in Figure 1 shows that Versions 3 and 4 can be used interactively or in batch mode. Version 4 allows the user to enter commands interactively using either the alphanumeric query language or the graphics menus. The details printed for each hit can include up to 53 pieces of information, 10 of which are default (refcode, compound and synonym name, compound name qualifier, molecular formula, author, journal coden and name, volume and page number, and year). The other 43 can be requested individually, but an item will always be displayed if that information field is being tested as part of the search question, e.g., if the search is for structures with an *R* factor which is less than 0.05, the *R* factor for each of the hit entries will be added at the end of the default list. Figure 12b shows the default hit information for the entry ABHYTZ. Version 4 will normally display 2D structural diagrams for all hits unless this feature is deselected.

The QUEST program prints a summary after every 1000 database entries listing the number of hits found so far and the percentage screenout. When a hit is found, a number of interactive options are presented to the user. These options provide the ability (a) to terminate the current search and either return to the QUEST command input module or exit the QUEST program completely, (b) to get help on a particular keyword or to print more information about QUEST's progress through the database, (c) to obtain more detailed information on the current hit, or (d) to continue the search without stopping at every hit. These options enable a very flexible, interactive search of the database to be carried out.

A record of the hits resulting from a search will be kept in a default "journal" file. Several other files containing hit information can also be written during a QUEST search (see Figure 1). Two of the most important files are the FDAT file, which Figure 1 shows is used as input to GSTAT and PLUTO, and a database (ASER) subset file, which enables further, more detailed searches to be carried out rapidly on a small number of entries.

3D SEARCH AND NUMERICAL ANALYSES: PROGRAM GSTAT

The flow chart of Figure 1 shows that the functionality of the program GSTAT is common to both Versions 3 and 4. Indeed, GSTAT is an onward development of the original GEOM78 included in Version 1[5] and of the program GEOSTAT[6,7] included in Version 2. In all versions, the program reads an FDAT interface file created by QUEST (or its predecessors) in response to a specific search query, i.e., the screened search capabilities of QUEST act as a primary filter for the 3D process.

GSTAT has two primary modes of operation. First, there is an entry-by-entry mode in which intramolecular geometry (bond lengths, valence angles, and torsion angles), metal coordination sphere geometry (distances and angles), or intermolecular contact distances between specified elements are calculated systematically for every entry in the retrieved subset. Such listings are presented in terms of the crystallographic atom labeling and can only be interpreted in conjunction with labeled 3D plots or by reference to the original paper. The variety of labeling schemes employed, even for closely similar molecules or substructures (fragments), mitigates against the use of entry-by-entry listings for any systematic work. Their use is, therefore, restricted to the examination of one (or a very few) highly specific entries.

Table VII. Possible Geometrical Parameters (Internal Coordinates) That Can Be Calculated by Program GSTAT Operating in Fragment Mode[a]

parameter type	specification
distance	atom–atom, atom–centroid, atom–plane centroid–centroid, centroid–plane
angle	involving atoms, centroids, vectors, and normals to planes
torsion	involving atoms and/or centroids
ring pucker	Cremer-Pople[42] for any ring, Altona-Sundaralingam[43] for five-membered rings
spherical polar coordinates	with respect to sp^3 and sp^2 geometries around target atom: quantifies a direction of approach

[a] Modifications and linear combinations of these basic parameters can also be generated (see text).

Table VIII. Typical 3D Search Queries Posed to the CSD System

locate all amide groups O=C—NH— and tabulate their bond lengths and angles
locate a triangular pharmacophore involving groups A, B, and C within A–B, B–C, and C–A distance constraints
what is the average value for a C(sp^2)–Cl bond
find the average C–N–C valence angle in azacycloheptane
classify and rank the observed preferred conformations of azacycloheptane
what is the mean O···O distance in O—H···O=C hydrogen bonds
what is the direction of approach of the O–H vector to the O=C—$(X)_2$ unit

It is the "fragment" mode of GSTAT that provides the facilities for 3D searching and for the systematic numerical and statistical analyses of derived geometry that underpin so many of the research projects based upon the CSD. Here, the program permits (a) the definition and location of a substructural fragment, a procedure that automatically imposes the user's own numbering onto each hit, (b) the calculation of a wide variety of user-specified geometrical parameters (internal coordinates) for each occurrence of the fragment, (c) the selection of fragments on the basis of geometrical criteria—the essence of what is termed 3D searching of the experimental results stored in the CSD, (d) the generation of a final systematic tabulation of N_p parameters for each of the N_f selected fragments, and (e) the statistical and numerical analysis of this table, expressed in the form of a data matrix $\mathbf{G}(N_f,N_p)$.

In providing 3D search capabilities in the CSD, we must recognize that queries will normally be formulated in terms of geometrical data which are calculated from the stored crystallographic coordinates, i.e., in terms of internal coordinates rather than in terms of external coordinates referred to some arbitrary origin. The internal coordinates can be related to the chemistry or 3D spatial characteristics of a given crystal structure, molecular species, or substructural fragment. For the CSD, which has both intramolecular and intermolecular information implicit in the stored coordinate and symmetry records, the range of possible 3D search terms is very large: typical examples are given in Table VII. Further, the types of query posed in 3D may be far less specific than the 1D and 2D QUEST searches discussed above. A variety of typical queries is illustrated in Table VIII.

These considerations tend to broaden the meaning of the word "search" in a 3D context. They highlight the need for data analysis facilities to be combined effectively with wide-

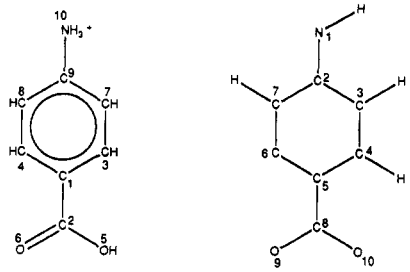


Figure 10. Chemical vs crystallographic connectivity.

ranging search options. This, in turn, highlights our own lack of knowledge of the information content of the 3D data: we cannot search for very long C–Cl bond lengths, for example, until we can define the concept "very long" in numerical terms. No such mystery surrounds the 1D and 2D data items which are, by and large, explicit information and/or a codification of well-established chemical conventions. In this section, we summarize the current status of 3D search and data analysis within the CSD system. The ongoing challenges provided by the 3D data are discussed further in subsequent sections.

Location of Substructures. The connection table available to GSTAT in the FDAT interface file differs considerably from that available to QUEST for substructure searches. The table is distance based: links between atoms are established by using the experimental 3D coordinate data and covalent radii criteria. Covalent radii have been established for all 85 elements represented in the CSD after many years experience with our check software;[3] numerical values are available in CSD system documentation. The result is a 2D representation of molecular topology in which the 3D crystallographic coordinates are node (atom) properties. The chemical and crystallographic connectivities are compared in Figure 10. The hydrogen-depleted graphs are topologically identical, but they have different atomic enumerations, and there is no chemical concordance between them. The crystallographic representation lacks bond-type information and details of atomic charges. Further, some or all of the hydrogen atoms may not have been located in the diffraction experiment. Thus, the chemical representation is not directly related to the 3D coordinate sets. This separation of connectivity representations explains the current two-pass (QUEST, GSTAT) 3D search mechanism in the CSD system. The establishment of a direct link between the two connectivities forms the basis for Version 5 and will be discussed later.

The two-pass mechanism requires that a 2D substructure search for the fragment, specifying all required chemical constraints, is first processed by QUEST, which generates the required FDAT interface. The same fragment, coded alphanumerically with a subset of the instructions of Table III, is then used by GSTAT to perform a topological search of the crystallographic connectivity. Since QUEST essentially functions as a chemical screening mechanism, the possibilities for a mismatch in GSTAT are very significantly reduced. However, chemical bond-type specifications can be approximated in GSTAT by use of distance constraints, e.g., a C=O double bond may be approximated by a distance range of, say, 1.15–1.20 Å. Extensive tables of reference data are now available[26,27] to provide suitable distance criteria.

Within the CSD system, the concept of distance-based connectivity can, of course, be extended to encompass intermolecular interactions by application of the crystallographic symmetry operators to the stored coordinate sets. Thus, for a study of hydrogen bonding involving (say) O-H···O′=C systems, all O...O′ nonbonded interactions less than a specified limit (e.g., 3.2 Å) may be calculated. Sometimes atom O′ will belong to the same molecule as the fixed target O (intramolecular contact). Typically, however, O′ will occur in a symmetry-related molecule, and the O...O′ link, together with the complete connectivity of the symmetry-related molecule, can be added to the overall distance-based representation. An extended definition of the H-bond topology can now be located via normal atom-by-atom, bond-by-bond techniques, and its geometrical descriptors may be calculated, tested, and analyzed as described below.

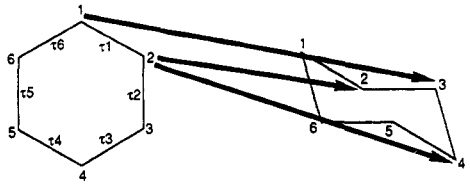


Figure 11. Multiple mappings due to 2D topological symmetry of search fragment.

Geometric (3D) Searches. A given 2D chemical fragment may, of course, exhibit a variety of conformations and other geometrical characteristics in 3D. The topological search results in the imposition of a user-defined atomic enumeration upon each example of the fragment located in the CSD. Further, it provides a direct link between these atoms and their 3D coordinates. The atomic enumeration provides a basis for user-specification of those geometrical parameters (internal coordinates) that are required for (a) further selection of fragments to complete the 3D search, (b) subsequent in-depth numerical analysis of fragment geometry, or (c) combination of both of these activities. Further, the program permits the user to assign his or her own mnemonic to each parameter for later use in the instructions associated with search or data analysis. A selection of possible parameters is given in Table VII. Values may be calculated not only by user reference to specific atomic nodes but also by reference to centroids, nonbonded vectors, or normals to mean planes, whose generation are themselves under user control.

The calculation of geometrical criteria is enhanced in GSTAT by a TRANSFORMATION mechanism. Individual parameters or pairs of parameters may be acted upon by a variety of FORTRAN-like operators to generate variants that may be used in searching or data analysis. Thus, if we have defined D1, D2, and D3 to be three bond lengths within each located fragment, then the simple pairwise constructs

```
TRA      DA =  D1  +  D2
TRA      DB =  DA  +  D3
TRA   DMEAN =  DB  /  3.0
```

will assign the name DMEAN to the mean value of the three distances. Each and every geometrical parameter that can be calculated by GSTAT can be used in 3D searching via a SELECT command, e.g.:

```
SEL  DMEAN   n1   n2
```

where n1 and n2 specify the permitted range for the parameter.

The atom-by-atom, bond-by-bond substructure search routine in GSTAT is a modification of that used in QUEST, so as to allow for some of the fundamental differences between 2D and 3D fragment searching. First, a 3D search must be exhaustive: all occurrences of the fragment must be located in the crystallographic topology of each entry since each occurrence will have different geometry. This exhaustivity is not required in 2D searching, where a single occurrence is (normally) sufficient to classify a particular CSD entry as a hit.

Second, the 3D routine must take account of possible 2D topological symmetry in the search fragment. This is illustrated in Figure 11: atom 1 of a cyclohexane search fragment (D_{6h} symmetry in 2D) may map to any one of the six atoms

in a CSD entry, say to atom number 3. Atom 2 of the fragment may now map to one of two alternatives, number 2 or number 4. The symmetry, therefore, gives rise to $6 \times 2 = 12$ alternative and 2D-equivalent mappings. However, in 3D this symmetry is not preserved, and different mappings generate different sequences of derived geometrical parameters, only one (or a few) of which may satisfy a sequence of selection criteria input by the user. This is particularly true for conformational searches, where torsion angles are used as search terms. GSTAT performs an exhaustive atom mapping until (a) a mapping is found which satisfies the 3D search criteria or (b) all possible mappings have failed the criteria. If no search criteria are specified, then the results of the first mapping are accepted.

Finally, structural inversions must also be taken into account in 3D searches of the CSD, where the coordinate set will define one of the two possible 3D enantiomorphs. In many cases, the two enantiomorphs are equivalent and the compound is achiral. If, however, a sequence of torsion angles with signs attached has been used as the selection criteria, then the signs, of themselves, will imply one of these enantiomers specifically. Hence it is necessary to test the criteria against both the CSD-encoded 3D structure and its mirror image. In some cases, however, the CSD structure will correspond to the actual absolute chirality of the compound. Here it may be undesirable to perform the structural inversion operation which is, therefore, under user control. The effects of topological symmetry and structural inversion on 3D search and data analysis procedures are more fully discussed elsewhere.[28]

Statistical and Numerical Analyses. The user-specified geometrical characteristics of each selected fragment form a data matrix $\mathbf{G}(N_f,N_p)$, where N_f is the number of fragments located, and N_p is the number of parameters specified. A mnemonic name is assigned by the user to each parameter, and this is used for reference in many of the analytical functions. In a large database such as the CSD, many structural variants of the same chemical fragment will exist in different crystal structures. One of the main purposes of data analysis is to classify these variants into homogeneous subgroups and, perhaps, derive mean geometry for each group. A variety of techniques, both mathematical and graphical, are used for this purpose. The techniques themselves are best classified on the basis of the number of variables treated by each.

Univariate Statistics and Display. A simple statistical summary of the distribution of each of the N_p parameters (maximum, minimum, and mean values, standard deviations, etc.) is automatically provided when the **G**-matrix is listed. Histograms of individual variables, identified by parameter name, can also be generated. These items are invaluable for assessing the chemical features of a retrieved subset, deriving suitable selection criteria, detecting outlying observations, and generating mean parameter values[26,27] for research and modeling purposes.

Bivariate Statistics and Display. It is often necessary to study and quantify relationships between pairs of parameters. Three functions are provided by GSTAT: (a) generation of a pairwise correlation/covariance matrix, (b) simple linear regression of one variable upon another, (c) graphical display of bivariate distributions by use of simple 2D scatterplots. Often the formation of discrete clusters of observations in the latter indicate some underlying classification criteria.

Multivariate Analyses. Differences in fragment geometry or conformation normally depend on variations in more than two parameters. Two numerical methods are available in GSTAT to treat these situations. First, principal component analysis (PCA)[29-31] uses eigenanalysis to reduce the dimensionality of the problem by expressing the total variance in the dataset in terms of some small number of principal components. Pairwise plots of the dataset, referred to these PC axes, can often reveal some underlying classification. Second, we are introducing a set of cluster analysis algorithms into GSTAT.[28,32,33] This is a numerical technique, as opposed to the visual approach of PCA, that is commonly used in many disciplines as a basis for unsupervised learning and classification of large datasets. The standard algorithms require modification to handle topological symmetry, enantiomeric inversions, and some of the circular data types (e.g., torsion angles) used to define conformational variants. This area of the CSD system is currently receiving considerable attention, since it forms a possible basis for automatic knowledge acquisition for use in molecular scene analysis[34] and other modeling applications.

3D GRAPHICAL DISPLAY: PROGRAM PLUTO

The PLUTO program is common to both Versions 3 and 4 (see Figure 1). It operates from a retrieved FDAT file, generated by QUEST, together with a set of alphanumeric instructions that control the style and content of the 3D illustration. The package is capable of generating mono or stereo pictures of individual molecules or complete crystal structures. A variety of styles are possible (line drawings, ball-and-spoke with perspective, space-filling models), and the user has control of the view direction, labeling of atoms, etc. More complete details are in the earlier reference.[5] PLUTO is currently undergoing a major restructuring for inclusion in the menu-driven upgrade to Version 5 described below.

AN EXAMPLE OF CSD SYSTEM USE

In a paper such as this, it is not possible, nor is it appropriate, to provide fine detail of all available system options; that is the role of system documentation. Rather, we provide a specific example to give readers a feel for the capabilities of the CSD system. The example chosen is a study of hydrogen bonding[35] between N–H donors and thiocarbonyl (C=S) acceptors. The two-stage QUEST–GSTAT approach, in terms of user input and system output, is summarized in the various parts of Figure 12. Input to QUEST is shown in alphanumeric form, although this would normally be encoded graphically within Version 4. The figure legends provide detailed commentary on the complete process; this text provides summary notes only.

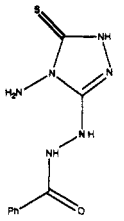
The purpose of the study is to establish the geometrical characteristics of N–H···S=C hydrogen bonds and to provide a visual survey of any variations in that geometry that might, eventually, be correlated with chemical environment and other features. We would stress that Figure 12 represents the results of an initial search and survey of the available data in the CSD. The first-stage process begins with a QUEST search for entries that contain both the donor and the acceptor groups. The search, as coded, will result in hits where the two may coexist in the same bonded unit, or in different bonded units. While such a search can be allowed to proceed in "automatic mode", it is preferable (within Version 4) to scan the chemical diagrams for the hits, both to ensure that the query has been correctly encoded and to obtain an overview of the types of compounds that are being retrieved. Such general observations can be of considerable use in rationalizing the gross features of the ultimate numerical distributions. In this example, there were 447 hits that satisfied the search criteria.

The second-stage process using GSTAT first establishes an "extended topological connectivity", as described in an earlier section, where the extension is based on S···N contacts which are (here) less than 3.7 Å. The FRAGMENT coded for the 3D search represents the complete, linked hydrogen-bonded motif, since the S···N connection is now logged and can be

a

```
SAVE FDAT
SCREEN 57
T1 *COOR .GE. 1
T2 *CONN
AT1  S  1  E
AT2  C  1  T3
BO  1  2  2                    C=S
END
T3 *CONN
AT1  N  99  1,2,3,4            NH (H=1-4)
END
QUES T1 + T2 + T3
```

b

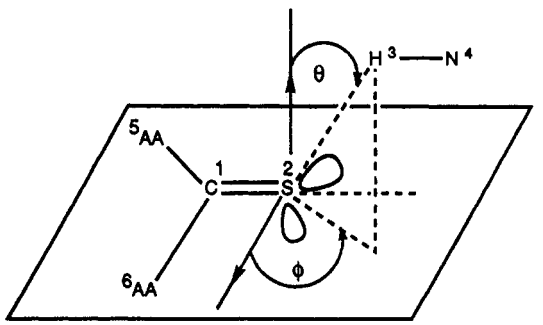


```
*REFC=ABHYTZ // *COMP=4-Amino-3-(beta-benzoylhydrazino)-5-mercapto-1,2,4-triazol
e // *FORM=C9 H10 N6 O1 S1 // *AUTH=R.C.Seccombe,C.H.L.Kennard // *CODE=188(J.Ch
em.Soc.,Perkin 2) // *VOLU= // *PAGE= 4 // *YEAR=1973 // *COOR=27 //
```

c

```
HNORM
CALC  INTER  FROM  S  2.1  TO  N  1.6  EXT
FRAG
AT1 C 3
AT2 S 2
AT3 H 1
AT4 N 2
AT5  AA  1
AT6  AA  1
BO 1 2
BO 3 4
BO  1  5
BO  1  6
END
DEF C=S 1 2
DEF S...H 2 3
DEF S.N 2 4
DEF SHN 2 3 4
DEF *LP2 Theta Phi 6 1 2 3
DEF *NBO BT24  2  4  -1
SEL  S.N  2.0  3.7
SEL  SHN  100.0  180.0
HIST SHN CONE
HIST S...H
HIST S.N
SCAT Theta Phi
SCAT  S...H    SHN
```

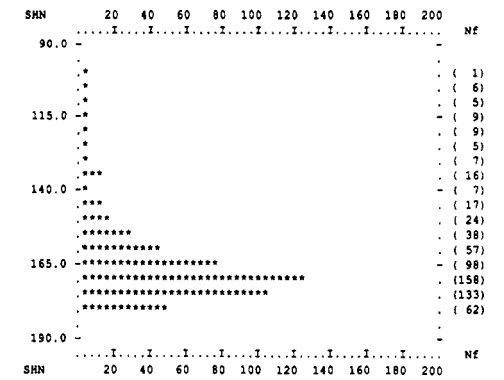
d



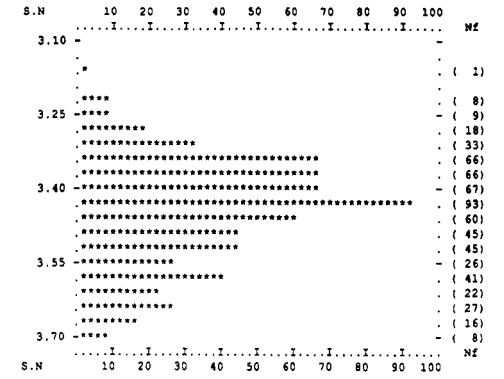
e

Nfrag	Refcod	C=S	S...H	S.N	SHN	THETA	PHI	BT24
1	ABHYTZ	1.669	2.329	3.310	163.747	68.107	15.839	-1
2	ACRTUR	1.692	2.464	3.412	156.282	59.177	15.329	-1
3	ACRTUR	1.692	2.392	3.390	169.771	80.064	15.414	-1
4	ACTCBZ	1.690	2.764	3.543	134.319	42.214	30.650	-1
	Mean	1.696	2.582	3.447	150.133	59.539	21.353	0.000
	S.D.Sample	0.032	0.241	0.106	20.806	23.779	23.678	0.000
	S.D.Mean	0.001	0.009	0.004	0.815	0.932	0.928	0.000
	Minimum	1.613	2.172	3.163	101.129	3.251	-80.588	0.000
	Maximum	1.818	3.333	3.699	178.957	90.000	89.511	0.000
	Nobs	651	651	651	651	651	651	0

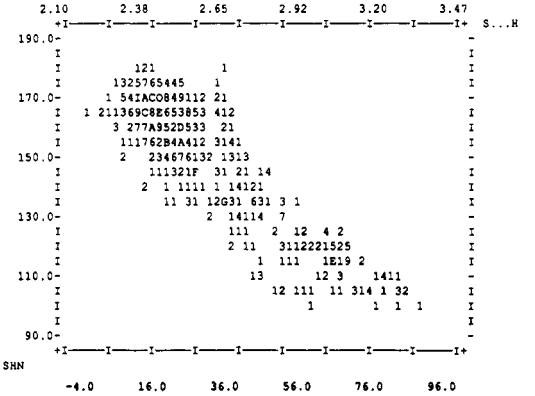
f



g



h



i

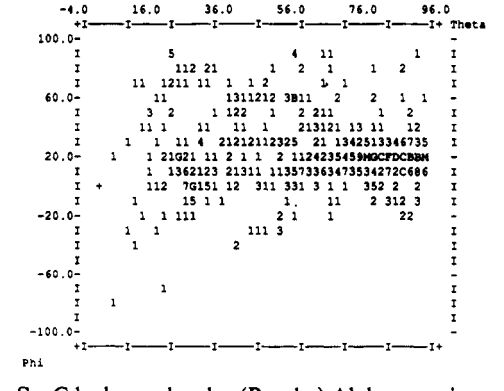


Figure 12. Example of CSD system use: a search and geometrical survey of N—H···S=C hydrogen bonds: (Panel a) Alphanumeric equivalent QUEST input. The search is restricted to organic compounds only (SCREEN 57) which have atomic coordinates available. There are two substructures to be located. The C=S is coded with S required to have [E]xactly one non-H connection, the C must have at least one such connection, but must be sp^2 by the T3 definition. For the N–H donor, only the N needs to be coded, to have any number of non-H connections (99 means skip the test on mca values), and from 1 to 4 attached H atoms. The question is the Boolean intersection of the three individual tests. (Panel b) Screen display for one of the hits generated by Version 4 in response to the instructions coded at (a) above. Located substructures are highlighted by thicker bonds for monochrome terminals, otherwise color discrimination is used. The text displayed is the standard option. The user may keep or reject the hit interactively or use a variety of other options to discover more about the entry. The program may also be set to automatic mode. (Panel c) Alphanumeric instruction set for GSTAT. HNORM demands that the C–H, N–H, and O–H positions are normalized along the X–H vector to yield their neutron diffraction determined bond lengths. The CALC INTER command sets up the extended connectivity (see text) within which the H-bonded fragment is to be located. Fragment topology only is given; AA is any atom. The required geometrical parameters are DEFined by a suitable name and in terms of atom numbers in the search FRAGment. The *NBO record requires that atoms 2 (S) and 4 (N) are in different bonded residues. The angles θ and ϕ calculated by *LP2 are defined in (d) below. HISTogram and SCATtergram requests are self-explanatory. (Panel d) Definition of the spherical polar coordinates θ and ϕ. S is the origin of the sphere and the $(AA)_2$-C=S system (assumed planar) lies in the equatorial plane, which will also approximate the S lone-pair plane. The angles describe the direction of approach of the H atom to the lone pairs. (Panel e) Section of the G-matrix printout from GSTAT. The complete summary statistics are shown; the parameter BT24 cannot be averaged since it is a flag indicating an intermolecular contact (intramolecular = +1); (Panel f) Histogram of N-H···S angles. The conical correction[36] is applied. Panels f–i of this figure are output by GSTAT directly to a line printer or terminal screen. (Panel g) Histogram of S···N distances. (Panel h) Scattergram of the S···H distance vs the N–H···S angle. Here, and in panel i, A = 10 instances, B = 11, C = 12, etc. (Panel i) Scattergram of the angle θ vs the angle ϕ defined in panel d.

located in the search. The 3D search criteria are initial, and very broad limiting values for the acceptance of a given S···N contact as an H-bond. Note that the inclusion of the H atom in the search motif automatically restricts the hits to those in which hydrogen atoms have been located in the diffraction experiment. This allows us to specify that the N–H···S angle shall be in the range normally accepted for H-bond formation. Further, the use of the NBOND keyword limits the search to intermolecular contacts: the S and N atoms must be located in different bonded residues, either part of the original asymmetric unit or symmetry related to it. The remaining coding instructs the program to calculate, for each located fragment, the geometrical characteristics of interest in the study. These are fairly standard for an initial survey in this area of chemistry. Finally, the visual display of results is requested. The 447 entries contained 689 examples of the H-bonded motif within the specified geometric criteria, of these 651 are classified as intermolecular.

The results from GSTAT are shown in Figure 12, panels e–i. The (truncated) printout of the G-matrix provides the self-explanatory summary statistics referred to above. For some search queries, e.g., those involving a search for a specific mean bond length, etc., these results are sufficient. In other cases, the mean values taken together with the data ranges and standard deviations might indicate some revision of the 3D search constraints. In all cases, but particularly here where we are studying a low-energy interaction in which considerable variation in individual parameters is expected, the visual display of data distributions is essential. Thus, the histogram of N–H···S angles (Figure 12f), with the conical correction factor[36] applied, indicates the expected preference for values approaching linearity at 180°. Despite this, there remains a small peak at an angle close to 140°, perhaps indicative of some special chemical environment effect or bifurcation that should be further investigated. The histogram of S···N distances shows a normal distribution in the range 3.3–3.7 Å, but exhibits two isolated peaks on the downward slope, again worthy of more attention.

The scattergrams of Figure 12h and 12i are of particular interest. The plot of the S···H distance versus the N–H···S angle shows the already observed preference for a linear three-atom system. Further, it exemplifies the well-known rule that H-bond length increases, i.e., the interaction weakens, as the angle at H departs from linearity. The plot of θ versus ϕ, illustrated in the geometrical construct of Figure 12d, shows that (a) θ values in the range 70–90° predominate, the H atom is approaching S in the plane of the $(AA)_2$–C=S system; and (b) ϕ values in a similar narrow range of +10 to +30° are preferred, the H is approaching S in directions which are likely to be occupied by the lone pairs; the mean C=S···H angle is, in fact, 106° over all fragments.

The study illustrated is obviously preliminary, more analysis is required to fully characterize the H-bond geometry in various chemical environments. It is, however, simple to perform and yields results that are of significant interest to those who wish to model such systems. It highlights some of our earlier comments concerning the broader definition of the word "search" in the 3D context and, indeed, shows that search and research can be synonymous terms in this area.

CURRENT DEVELOPMENTS IN THE CSD SYSTEM

The current work of the Centre is focused on two main goals, the integration of 3D searches into the interactive menu-driven QUEST package and the provision of software to enable external users to convert their in-house crystal structure results into the CSD ASER file format.

The integration and extension of the 3D search capabilities of GSTAT within the QUEST program, to yield a new system upgrade designated Version 5, has already been noted. As with Versions 3 and 4, the new development will be underpinned by a further upgrade of the ASER database file, together with software enhancements to take advantage of new data items. The presence of two quite separate connectivity representations, chemical and crystallographic (see Figure 10), has already been identified as the underlying reason for the two-stage approach to 3D searching. There is an obvious need for a unified connectivity representation that links the 3D coordinate data, as additional atom properties, to the conventional 2D chemical description. Further, we need to extend the search heuristics to include some bit-encoded representations relating to the likely 3D search terms.

The matching of the two connectivity representations is not straightforward, due to the nature of the crystallographic results. The chief problems are as follows: (a) The presence of more than one bonded residue in the crystal structure, the residues can be chemically different, in which case a 1:1 mapping of the two representations is preserved, or they may be chemically identical (more than one molecule in the asymmetric unit), whence a 1:many mapping will result. (b) Some or all of the hydrogen atoms may be missing in the reported crystal structure. (c) There may be local or complete topological symmetry in the hydrogen-depleted crystallographic connectivity, but not in the full 2D chemistry. (d) All or part of the crystal structure may be disordered, giving a 1:many mapping for just a few affected atoms. (e) The crystal structure may be polymeric on distance criteria, in which case reduction to the monomer for matching can present difficulties. Despite these problems, we have now developed a graph-theoretic algorithm that has succeeded in matching the connectivities of some 89% of CSD entries for which atomic coordinates are available. This information is now being checked and validated before inclusion in an extended ASER file.

The generation of additional bit-screen heuristics for 3D searches is currently a topic of considerable research interest.[37-39] The chief problem is to identify those internal coordinates (Table VII) which are most likely to be used in 3D selection criteria and then to devise suitable bit-encoded descriptors. However, within the planned fully integrated 1D, 2D, and 3D search capabilities of Version 5 QUEST, the existing screens will, in many cases, act as an efficient primary filter. We need only concern ourselves with very general search problems, e.g., those posed by pharmacophoric patterns defined with minimal chemical content, and only with internal coordinates involving the most common elements. A pharmacophore is a set of structural features in a drug molecule, often defined in 3D geometrical terms, which is recognized at a receptor site. A further, but related problem, concerns the selection of descriptors to facilitate 3D similarity searches. It is not yet clear that these will correspond exactly to the 3D screens, as they do in the 2D case discussed above.

Software modifications are currently in hand to take advantage of both of these file upgrades. Thus, the 3D search methodology of GSTAT has been integrated into QUEST, and further requested extensions are being introduced. The interactive menu-driven graphical interface is being upgraded to accommodate these improved search capabilities. Subroutines derived from PLUTO are being restructured and integrated into QUEST to provide interactive 3D graphical display of hits. A full description of Version 5 will be presented in the near future.

Finally, we are undertaking a thorough review of our in-house software system for the input, checking, and structuring of crystallographic data. Eventually some parts of this system may be offered to individual crystallographic laboratories as an aid to manuscript preparation and to permit the comparison

of new unpublished structural results with the existing body of knowledge. As a first step, we have developed a program that will process the results of a crystal structure analysis, prepared in CCDC internal formats, through to the corresponding ASER entry. It is assumed that the dataset has been generated via an automated diffractometer system and that the user has transformed it to the required input format. Hence, data checks in this prototype are kept minimal. Further developments are being planned to enhance our in-house processing activities in the first instance; again, they will be reported in due course.

IMPLEMENTATIONS OF THE CSD SYSTEM

Machine-Independent Implementation. As stated earlier, this implementation consists of an entry-sequential ASER file (currently Version 4 requires 204 Mbytes of disk space) together with the FORTRAN77 source code for QUEST, GSTAT, and PLUTO. In practice, the database is supplied in a formated form (FSER) in EBCDIC or ASCII characters, together with a small FORTRAN77 program for conversion to a binary file (ASER) compatible with the host computer. No attempt is made to create direct-access files from FSER, since this would require the addition of suitable FORTRAN77 direct-access READ's within the source code. Thus, possible problems of the machine-dependency of direct-access file structures are avoided. Within Version 4, the graphics system operates via Tektronix escape codes, and a suitable proprietary terminal, or another equipped with a satisfactory emulator, is essential for running the system. Suitable emulators (e.g., Tgraf, VersaTerm Pro) are available for IBM PC's and clones, and for Apple Macintosh microcomputers connected to the host machine. System-dependent "job control" instructions must be generated at each host site.

DEC-VAX/VMS Implementation. The database file has been restructured for VAX/VMS in order to minimize the disk storage it occupies (currently Version 4 requires 102 Mbytes of disk space) and the CPU resources needed for data access.

Furthermore, the performance of the software was improved by modifying the code to take advantage of VMS-specific system services. The database modifications are

Division of the Database According to Record Type. The SCREEN records of all entries are collected together and stored in a pair of files. The TEXTCONN and DATA records form another file. This division is chosen because all screen records are read during a search while only selected TEXTCONN and DATA records are accessed. Most of the CPU time consumed by a search is used to bring the screen records into the program memory. This is minimized by using mapping, i.e., by use of calls to the SYS$CRMPSC system service.

Use of an Indexed-Sequential File. The TEXTCONN and DATA records are written as an indexed file with one 9-character key made up from the 8-character refcode and the suffix C or D indicating either TEXTCONN or DATA record.

Compression of TEXTCONN and DATA Records. These records make up the bulk of the database but are rarely accessed. They are therefore compressed at the Centre using the DCX$COMPRESS_DATA system service. The compressed records and the compression/expansion function are then exported. When the search software requires these records for an entry, then they are expanded using the DCX$EXPAND_DATA system service. This compression reduces the size of the records by an average of 50%.

The search software was modified to take advantage of interprocess communications available in VAX/VMS. The search executes as two concurrent processes. A background process applies only the screens and queues selected entry identifiers to the foreground process. The foreground process acts upon the remaining TEXTCONN and DATA records (if necessary) and finally displays any hits to the user. In this way there is a minimal pause between the display of hits.

Silicon Graphics Unix Workstation Implementation. The database file has been restructured for Unix (currently Version 4 requires 196 Mbytes of disk space) to enable direct access to the TEXTCONN and DATA records selected after screening and to minimize the CPU resources required to read all of the SCREEN records.

The performance of the software has been improved by using modules written in C and Unix system services. The Silicon Graphics proprietary graphical library is used to display graphics on the workstation itself, while Tektronix terminals attached to the workstation are also supported.

Database Modifications. The SCREEN records of all entries are collected together and stored as a single file (e.g., c_dbase.msk) and the remaining TEXTCONN and DATA records form another (e.g., c_dbase.tcd). The database I/O is performed using FORTRAN calls to C subroutines which in turn make use of the *lseek* and *read* Unix functions. The argument for the *lseek* function (the offset for the TEXTCONN and DATA records) is stored in the SCREEN record for that entry. The buffer length argument to the *read* function is calculated from information already in the SCREEN record.

Software Modifications. The search software has been modified to operate as two concurrent processes in a similar way to the VAX/VMS implementation described above. Interprocess communication is via modules written in C, making use of the *fork* and *execlp* Unix functions and a pipe.

DISTRIBUTION AND UTILIZATION

The CSD system is widely distributed, both to academic and for-profit Institutions. In 1991, Affiliated Centres in 31 countries are involved in worldwide distribution to academic users. Affiliated Centres receive master copies of the latest system version and make it available to individual laboratories within an agreed geographical area. For-profit Institutions are supplied directly from Cambridge, except for those in Japan where the Japanese Association for International Chemical Information acts for the Centre. Once the master database is installed at a given site, new and edited database entries and new copies of the maintained software are issued at intervals of 6 months, in January and July of each year.

Obviously, the CSD system is heavily used for the retrieval of literature references, for rapid surveys of the geometry of particular subsets, and to provide 3D structural input to molecular modeling and drug design programs. It also forms the "experimental" basis for a wide range of detailed research projects.[40,41] These relate to the development of new numerical analysis methodologies, the determination of mean molecular dimensions, the study of substituent effects and chemical bonding, studies of conformational variations and preferences, stereochemistry, and the systematic analysis of the hydrogen-bonded and nonbonded interactions that lie at the heart of molecular recognition processes. Some 200 papers of this type have been published to date, and we are currently including these references, as a separate category, within the CSD itself. A list of references, believed complete to the end of 1986, has already been published.[41]

CONCLUSION

This paper has described the development, implementation, and use of the CSD systems designated as Versions 3 and 4; further development to a Version 5 is summarized. We would reiterate that, although the system has two components (the database and the software), development depends crucially on

symbiotic upgrades in both areas. The CSD is, of itself, a typical database. It contains a wealth of data items that can be searched explicitly or implicitly. More importantly, these data items can be used to acquire new knowledge which can be stored in some suitable representation. This knowledge can be used to enhance existing search capabilities; the bit-screen heuristics employed in many databases are an example of encoded knowledge acting in this way. Further, the knowledge representations can themselves form the basis for new specific searches and be employed in automated reasoning and inference procedures. Hence, the symbiotic relationship between the database and the related software is fundamental to all future developments.

Nowhere are these arguments more valid than in the area of 3D structural chemistry. This paper has shown that we know very little about the systematics, the formal grammar, of this important area. In many cases, an initial broad survey of CSD is needed to establish even preliminary 3D search terms. Indeed, as we have pointed out, the word "search" takes on a much broader meaning in 3D, and "search" and "research" can, in many cases, become closely related activities. There is no doubting the complexity of the rules that govern the assembly of substructural units to form bonded 3D molecules. There is no doubting that the rules by which these molecules associate one with another are more fuzzy and an order of magnitude (at least) more complex. However, the huge reservoir of data now available provides a basis for further knowledge acquisition and for attempts to deduce the complex grammars involved in 3D structure in its broadest sense. These are the challenges of the next decade.

ACKNOWLEDGMENT

We express our deep gratitude to those staff members of the CCDC, both past and present, who have been responsible for the building and maintenance of the Cambridge Structural Database. We also thank Dr. Sam Motherwell and Dr. Peter Murray-Rust for their specific and fundamental contributions to the developing CSD system, and those many users whose suggestions, contributions, and advice have been so constructive.

REFERENCES AND NOTES

(1) *The Royal Society Scientific Information Conference Report*; The Royal Society: London, 1948.
(2) Watson, D. G. Printed Information Sources in Crystallography. In *Crystallographic Databases*; Allen, F. H., Bergerhoff, G., Sievers, R., Eds.; International Union of Crystallography: Chester, U.K., 1987; pp 25–29.
(3) Allen, F. H.; Kennard, O.; Motherwell, W. D. S.; Town, W. G.; Watson, D. G.; Scott, T. J.; Larson, A. C. The Cambridge Crystallographic Data Centre. 3. The Unique Molecule Program. *J. Appl. Crystallogr.* **1974**, *7*, 73–78.
(4) *Molecular Structures and Dimensions*; Kennard, O., Watson, D. G., Allen, F. H., Bellard, S., Cartwright, B. A., Eds.; Reidel: Dordrecht, The Netherlands, 1971–1984; Vols. 1–15.
(5) Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. The Cambridge Crystallographic Data Centre: Computer-Based Search, Retrieval, Analysis and Display of Information. *Acta Crystallogr.* **1979**, *B35*, 2331–2339.
(6) Murray-Rust, P.; Raftery, J. Computer Analysis of Molecular Geometry VI: Classification of Differences in Conformation. *J. Mol. Graphics* **1985**, *3*, 50–59.
(7) Murray-Rust, P.; Raftery, J. Computer Analysis of Molecular Geometry VII: The Identification of Structural Fragments in the Cambridge Structural Data File. *J. Mol. Graphics* **1985**, *3*, 60–68.
(8) Machin, P. A. Programming Aspects of Crystallographic Data Files: Interactive Retrieval from the Cambridge Database. In *Crystallographic Computing 3*; Sheldrick, G. M., Kruger, C., Goddard, R., Eds.; Oxford University Press: Oxford, U.K., 1985; pp 106–118.
(9) See, e.g.: Ash, J. E.; Chubb, P. A.; Ward, S. E.; Welford, S. M.; Willett, P. *Communication, Storage and Retrieval of Chemical Information*; Horwood-Wiley: Chichester, U.K., 1985; Chapter 6.
(10) Crowe, J. E.; Lynch, M. F.; Town, W. G. Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. 1. Non-Cyclic Fragments. *J. Chem. Soc. (C)* **1970**, *23*, 990–996.
(11) Adamson, G. W.; Lynch, M. F.; Town, W. G. Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. 2. Atom-Centered Fragments. *J. Chem. Soc. (C)* **1971**, *24*, 3702–3706.
(12) Adamson, G. W.; Lambourne, D. R.; Lynch, M. F. Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. 3. Statistical Association of Fragment Incidence. *J. Chem. Soc., Perkin Trans. 1* **1972**, 2428–2433.
(13) Adamson, G. W.; Cowell, J.; Lynch, M. F.; Town, W. G.; Yapp, M. A. Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. 4. Cyclic Fragments. *J. Chem. Soc., Perkin Trans. 1* **1973**, 863–865.
(14) Willett, P. A. Screen Set Generation Algorithm. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 159–162.
(15) Hodes, L. Selection of Descriptors According to Discrimination and Redundancy: Applications to Chemical Structure Searching. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 88–93.
(16) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. The CAS ONLINE Search System. 1. General System Design and Selection, Generation and Use of Search Screens. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.
(17) Harrison, M. C. Implementation of the Substring Test by Hashing. *Commun. ACM* **1971**, *14*, 777–779.
(18) van Rijsbergen, C. J. *Information Retrieval*; Butterworths: London, 1975.
(19) Rich, E. *Artificial Intelligence*; McGraw-Hill: London, 1983.
(20) Wipke, W. T.; Dyott, T. M. Use of Ring Assemblies in a Ring-Perception Algorithm. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 140–147.
(21) Adamson, G. W.; Bush, J. A. A Comparison of the Performance of Some Similarity and Dissimilarity Measures in the Automatic Classification of Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 55–58.
(22) Everitt, B. *Cluster Analysis*; Halsted-Heinemann: London, 1980.
(23) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nearest-Neighbor Searching in an Online Chemical Structure Search System. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36–41.
(24) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press-Wiley: Letchworth, U.K., 1987.
(25) Bawden, D. Browsing and Clustering of Chemical Structures. In *Chemical Structures: the International Language of Chemistry*. Warr, W. A., Ed.; Springer-Verlag: Berlin, 1988.
(26) Allen, F. H.; Kennard, O.; Watson, D. G.; Brammer, L.; Orpen, A. G.; Taylor, R. Tables of Bond Lengths Determined by X-ray and Neutron Diffraction. 1. Bond Lengths in Organic Compounds. *J. Chem. Soc., Perkin Trans. 2* **1987**, S1–S19.
(27) Orpen, G. A.; Brammer, L.; Allen, F. H.; Kennard, O.; Watson, D. G.; Taylor, R. Tables of Bond Lengths Determined by X-ray and Neutron Diffraction. 2. Organometallic Compounds and Coordination Compounds of the d- and f-Block Metals. *J. Chem. Soc., Dalton Trans.* **1989**, S1–S83.
(28) Allen, F. H.; Doyle, M. J.; Taylor, R. Automated Conformational Analysis from Crystallographic Data. 1. A Symmetry-Modified Single-Linkage Clustering Algorithm for 3D Pattern Recognition. *Acta Crystallogr.* **1991**, *B47*, 29–40.
(29) Murray-Rust, P.; Motherwell, W. D. S. Computer Retrieval and Analysis of Molecular Geometry. 3. Geometry of the β-1′-Aminoribofuranoside Fragment. *Acta Crystallogr.* **1978**, *B34*, 2534–2546.
(30) Taylor, R. The Cambridge Structural Database in Molecular Graphics: Techniques for the Rapid Identification of Conformational Minima. *J. Mol. Graphics* **1986**, *4*, 123–131.
(31) Auf der Heyde, T. P. E. Analyzing Chemical Data in More Than Two Dimensions. *J. Chem. Educ.* **1990**, *67*, 461–469.
(32) Allen, F. H.; Doyle, M. J.; Taylor, R. Automated Conformational Analysis from Crystallographic Data. 2. Symmetry-Modified Jarvis-Patrick and Complete-Linkage Clustering Algorithms for 3D Pattern Recognition. *Acta Crystallogr.* **1991**, *B47*, 41–49.
(33) Allen, F. H.; Doyle, M. J.; Taylor, R. Automated Conformational Analysis from Crystallographic Data. 3. 3D Pattern Recognition within the Cambridge Structural Database System, Implementation and Practical Examples. *Acta Crystallogr.* **1991**, *B47*, 50–61.
(34) Glasgow, J. I.; Fortier, S.; Allen, F. H. Crystal and Molecular Structure Determination through Imagery. In *Artificial Intelligence and Molecular Biology*; Hunter, L., Ed.; AAAI Press: San Francisco, 1991.
(35) Taylor, R.; Kennard, O. Hydrogen-Bond Geometry in Organic Crystals. *Acc. Chem. Res.* **1984**, *17*, 320–326.
(36) Kroon, J.; Kanters, J. A.; van Duijneveldt-van de Rijt, J. G. C. M.; van Duijneveldt, F. B.; Vliegenthart, J. A. O-H···O Hydrogen Bonding in Molecular Crystals: A Statistical and Quantum-Chemical Analysis. *J. Mol. Struct.* **1975**, *24*, 109–129.
(37) Jakes, S. E.; Willett, P. Pharmacophoric Pattern Matching in Files of 3D Chemical Structures: Selection of Interatomic Distance Screens. *J. Mol. Graphics* **1986**, *4*, 12–20.
(38) Cringean, J. K.; Pepperrell, C. A.; Poirette, A. R.; Willett, P. Selection of Screens for 3D Searching. *Tetrahedron Comput. Methodol.* **1991**, in press.
(39) Sheridan, R. P.; Nilakantan, R.; Rusinko, A., III; Bauman, N.; Haraki, K. S.; Venkataraghavan, R. 3DSEARCH: A System for Three-Dimensional Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 255–260.

(40) Allen, F. H.; Kennard, O.; Taylor, R. Systematic Analysis of Structural Data as a Research Technique in Organic Chemistry. *Acc. Chem. Res.* **1983**, *16*, 146-153.
(41) Allen, F. H.; Kennard, O. The Cambridge Structural Database: Current Applications and Future Developments. In *Crystallographic Databases*; Allen, F. H., Bergerhoff, G., Sievers, R., Eds.; International Union of Crystallography: Chester, U.K., 1987; pp 55-76.
(42) Cremer, D.; Pople, J. A. A General Definition of Ring-Puckering Coordinates. *J. Am. Chem. Soc.* **1975**, *97*, 1354-1358.
(43) Altona, C.; Sundaralingam, M. Conformational Analysis of the Sugar Ring in Nucleosides and Nucleotides: A New Description Using the Concept of Pseudorotation. *J. Am. Chem. Soc.* **1972**, *94*, 8205-8212.

Chemical Abstracts Service Chemical Registry System. 13. Enhanced Handling of Stereochemistry†

J. E. BLACKWOOD, P. E. BLOWER, JR.,* S. W. LAYTEN, D. H. LILLIE, A. H. LIPKUS, J. P. PEER, C. QIAN, L. M. STAGGENBORG, and C. E. WATSON

Chemical Abstracts Service, P.O. Box 3012, Columbus, Ohio 43210

Received January 14, 1991

CAS registers stereoisomers using *text descriptors* related to the stereochemical descriptors of the corresponding chemical names. This system works well for the unique registration of stereoisomers, but it is difficult to relate the text descriptor to the atoms and bonds of the Registry connection table. This limits its usefulness for substructure search or display of stereochemistry in the structure diagram. CAS is currently augmenting the Registry connection table with atom/bond-specific stereodescriptors. This article focuses on two aspects of this work: the representation of stereochemistry in the connection table and techniques for converting the Registry structure file to the stereoaugmented format.

INTRODUCTION

The Chemical Abstracts Service (CAS) Chemical Registry System marked its 25th anniversary in 1990. The Registry System is a computer-based system that identifies substances on the basis of their molecular structure. Begun originally to support substance indexing for *Chemical Abstracts* (CA), the Registry System has influenced the work of chemists and other scientists around the world. Perhaps best known as the source of CAS Registry Numbers and of the Registry File on the STN International network, the Chemical Registry System provides a foundation for substance identification used by the scientific community worldwide, with over 10 million substances currently on file. The CAS Registry Number, which links the structure with the *CA* index name and other data, is used for chemical substance identification by many governmental agencies and industrial organizations.

The Registry System has evolved over a period of years. The initial 1965 version, Registry I, was designed to be as specific as possible. It registered only fully defined organic compounds. Other substances such as inorganic compounds, polymers, and compounds with partially known structures were manually assigned Registry Numbers on the basis of their structural diagrams, names, or molecular formulas. In 1968 the second version of the CAS Registry System, Registry II, extended machine registration to include inorganic substances, coordination compounds, polymers, mixtures, alloys, and certain incompletely defined substances. Registry III, the current version of the Registry System, became operational in late 1973. Although no changes were made in the basic algorithmic techniques for registration, Registry III made a major adjustment to the Registry records so that the system would more effectively support CAS index nomenclature generation and computer-based structure output. The design, content, functions, statistics, special features, and input structure conventions have been described in detail in previous papers.[1-11]

The boundaries of the Chemical Registry System are constantly being extended, both in content and the manner in which information is added. In the past 25 years the CAS Chemical Registry System has evolved from a production tool for CAS publications and services to a vital, useful service for the scientific community worldwide. And with this expanded role, the Registry System and related services will become even more responsive to the needs of scientists and engineers.

CAS is currently developing Registry IV. Improvements currently underway include enhanced alloy processing, improvements in structure display, faster registration due to a new online editorial input system, addition of biosequence data, and introduction of display and search of stereochemical information in the Registry File. Additional items being investigated include modifications to allow scientists to search using a variety of conventions for representing substances and better support for industrial and engineering users of information about metals and alloys, polymers, ceramics, and composites.

Stereochemical representation[3,12,13] has been an integral part of the CAS Chemical Registry System since its inception as Registry I. Stereoisomers are considered to be different, individual substances, and each is recognized as unique. The unique identification of stereoisomers permits the storage and retrieval of information collected from scientific literature about a specific isomer. Currently stereochemical data is only recorded in the chemical name and a controlled vocabulary field known as the CAS Text Descriptor. This is being enhanced by a technique for recording specific atom and bond stereochemistry in the connection table,[1] the Registry structure record maintained for each substance.

REPRESENTATION

Several different techniques will be used to represent atom/bond-specific stereochemistry in the connection table record, depending on the type of stereocenter being described. Tetrahedral stereocenters, stereogenic double bonds, and allenes are described by using a parity descriptor, similar to that described by Petrarca et al.[14] and adapted by Wipke and Dyott[15] in their Stereochemically Extended Morgan Algorithm. The relative Cahn-Ingold-Prelog[16,17] (CIP) ranks of the neighboring atoms are recorded along with the parity data. This information allows straightforward calculation of an *R*

† Dedicated to Professor Michael Lynch, whose participation in work on computer representation of stereochemistry more than 20 years ago helped provide the basis for the work reported here.

0095-2338/91/1631-0204$02.50/0