Acta Cryst. (1998). A**54**, 758–771

The Development, Status and Scientific Impact of Crystallographic Databases

Frank H. Allen

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England.
E-mail: allen@ccdc.cam.ac.uk

(*Received 8 July 1998; accepted 5 August 1998*)

Abstract

Nearly 300 000 crystal structures have been reported in the scientific literature and all of them are accessible through the crystallographic structural databases. The historical development, information content and current status of these databases are described, with special reference to methods for data acquisition and structure validation. The relationships that exist between authors, journals and databases are discussed in the light of statistics that predict more than 800 000 structural database entries by the year 2010, more than doubling the output of the last 30 years in less than half the time. Use of the structural databases for data mining and knowledge acquisition is summarized. So far, the vast majority of this research activity has centred around the databases that record small-molecule and macromolecular structures. The creation of knowledge-based libraries of structural information from the existing molecular databases suggests a new era of two-level information provision: the raw-data level and the derived-knowledge level. The crystallographic knowledge bases are encouraging the development of software systems that access the stored knowledge to solve complex problems in structural science.

Frank Allen was born in Reading, UK, and is a graduate of Imperial College London. Following postdoctoral work at the University of British Columbia, Vancouver, he joined the embryo CCDC in 1970, becoming Deputy Director in 1991 and Scientific Director in 1997. His research interests are in database design, search and retrieval, and in applications of crystallographic information in structural chemistry. He is the current Editor of Acta Crystallographica Section B, Vice-President of the British Crystallographic Association and a Council Member of the European Crystallographic Association. He received the UK Royal Society of Chemistry Award for Structural Chemistry in 1994. He has interests in education, travel and sport (now at the observational level only!).

1. Introduction

Literature is news that STAYS news. (Ezra Pound)

The literature of any scientific discipline is crucial to its onward progression. However, it is usual for earlier work to lose its immediacy and be superseded gradually over time. Not so in structural crystallography. Here each publication has a timeless quality, being rediscovered by successive readers, often from different disciplines and for very different reasons. For example, a structure initially performed to establish molecular stereochemistry may later prove vital because it exhibits unusual conformational features, displays certain molecular recognition properties or is related to a novel drug or material. It is this permanent and multidisciplinary value that has driven the development of specialist journals and, particularly, printed compendia and computerized databases to preserve the primary numerical results of crystal structure analyses.

Because the early literature of crystallography was widely spread, the production of printed compendia preceded the establishment of specialist journals. *Strukturbericht* (Ewald & Hermann, 1929) appeared only 16 years after the publication of the first crystal structure and continued into the mid-1980s as *Structure Reports* published by the International Union of Crystallography (IUCr). These volumes recorded bibliographic, chemical and primary crystallographic data, including atomic coordinates, and were the printed forerunners of the computerized databases.

The first act of the newly formed IUCr was indeed to create a major specialist primary journal. In the words of its founding Editor, *Acta Crystallographica* was designed to be '. . . a central place for publication and discussion of all research in this vast and ever expanding field' (Ewald, 1948). Later in that same Editorial, Ewald (1948) remarked that 'The best scheme for the publication of scientific investigations is a problem of outstanding importance for the sound development of science, and is particularly acute in those fields that touch on many different branches of study'. This was explicit recognition of the guiding philosophy that had

underpinned the creation of *Strukturbericht* and that would now contribute to the development of *Acta*.

Crystallographers have generated many other printed secondary publications, compendia and indexes. Notable among these are Crystal Data (National Institute of Standards and Technology, 1998) and the Powder Diffraction File (International Centre for Diffraction Data, 1998), material that was ideally suited for re-expression in the form of computerized databases.

Despite the continued value of these and other crystallographic database projects, this article will concentrate exclusively on the crystallographic *structural* databases, *i.e.* those databases that have now replaced *Structure Reports* as curators of the world's output of primary crystal structure data. The article will try to look forwards as well as backwards and attempt to summarize (*a*) the crystallographic databases themselves, with special reference to data acquisition and structure-validation methods, including some assessment of the future in these areas, (*b*) the current status of software systems for database access together with an overview of the major research applications of crystallographic data, particularly in the area of molecular structure, and (*c*) the derivation of crystallographic knowledge bases from existing database content and likely future applications of derived structural knowledge in conjunction with problem-solving software.

2. The crystallographic structural databases

All science is either physics or stamp collecting! (Ernest Rutherford)

It is one thing to have a literature, it is quite another for it to be accessible to specialists and nonspecialists alike. The rapid development of computer technology in the 1950s and 1960s began to provide the means for the organization and dissemination of scientific information, particularly the storage and retrieval of bibliographic and other textual material. In the chemical context, much pioneering work was also carried out on the computerized representation, manipulation and interrogation of chemical structural diagrams, the visual language of molecular chemistry (see *e.g.* Gray, 1986, and references therein). These planar two-dimensional (2D) representations of atomic connectivity were encoded as undirected graphs with chemical atoms as the nodes and chemical bonds as the edges. The inclusion of stereochemical information, in the form of 'wedge' and 'dot' bond attributes enhanced the description to a '2.5-dimensional' level. The development of chemical connection tables permitted a full array of graph-theoretical algorithms to assist in the analysis and classification of constituent units, *i.e.* rings, chains and chemical functional groups. Most importantly, algorithms and heuristics began to be developed for the rapid location of chemical substructures, specified as graphical search queries, within very large collections of chemical graphs.

Table 1. *Current entry statistics (June 1998, to the nearest thousand) for the crystallographic structural databases*

CSD	186000
PDB	8000
NDB	1000
ICSD	48000
CRYSTMET†	45000
Total	**288000**

† 19 000 primary structure determinations, 26 000 assigned structures.

Olga Kennard working in Cambridge in the mid-1960s had the vision to drive these developments to their ultimate goal in structural crystallography, through the creation of a numerical database that would store the primary results of all three-dimensional (3D) crystal structure analyses of organocarbon compounds. The timing was perfect. Worldwide output of just a few hundred small-molecule structures per year made the abstracting of current publications a tractable proposition, while assimilation of retrospective material was made possible through the existence of *Structure Reports*. Other pioneers soon followed so that within a decade the complete chemical spectrum, from metals and alloys to proteins and viruses, was covered by an active crystallographic structural database.

The following subsections provide a very brief overview of the five structural databases now available, full details can be obtained from the references cited. Current overall database statistics are given in Table 1.

2.1. *Cambridge Structural Database (1998) (CSD)*

Compilation of the CSD (Allen *et al.*, 1979, 1991; Kennard & Allen, 1993) began in 1965 in the University Chemical Laboratories, Cambridge, England, an operation that led to the formation of the present Cambridge Crystallographic Data Centre. The CSD is concerned with organics, organometallics and metal complexes and the information content of each entry (summarized in Fig. 1) comprises bibliographic and chemical text, 2D structural diagrams (as connection tables with display coordinates), and the cell dimensions, symmetry and atomic coordinates that describe the 3D molecular and crystal structure. Growth of the CSD is summarized in Fig. 2(*a*).

2.2. *Protein Data Bank (1998) (PDB)*

Originally founded in 1971 by the late Walter C. Hamilton at Brookhaven National Laboratory, USA, the PDB was directed by Thomas Koetzle until 1993. The present director is Joel L. Sussmann. From slow beginnings, the PDB of recent years (Abola *et al.*, 1997) has shown dramatic growth, depicted in Fig. 2(*b*), so that

the few hundred structures of the 1980s has grown to a current total of 8000 entries (Table 1). PDB entries have always been deposited in electronic form, and the issue of a PDB ID code is a pre-publication requirement of the major professional organizations and of most major scientific journals. Broad categories of PDB information mirror those in the CSD, except that amino-acid sequence information (a text field in which formal chemistry is implicit) parallels the CSD connectivity records.

2.3. *Nucleic Acid Data Bank (1998) (NDB)*

For many years, structures of nucleic acids were split between the CSD and the PDB on the basis of size. Recently, all of these structures (currently 731: Table 1) have been brought together in one database under the direction of Helen Berman at Rutgers University, USA (Berman *et al.*, 1992). Information types again mirror the CSD basis, with oligomer codes acting as implicit connectivity records, and with the database also recording standard geometrical information.

2.4. *Inorganic Crystal Structure Database (1998) (ICSD)*

Building on the Bond Index to the Determination of Inorganic Structures (Brown, 1969–1981), the ICSD was established as a computerized resource in the laboratory of Guenter Bergerhoff at the University of Bonn, Germany. The work of database creation is now carried out at the Fachinformationszentrum, Karlsruhe, Germany. The ICSD (Bergerhoff *et al.*, 1983, 1998) covers inorganics and minerals, and has broad information categories that mirror the CSD, except that formal chemical connectivity is not generally applicable and is not recorded. There is a small area of agreed overlap between the CSD and the ICSD in the area of 'molecular inorganics'.

2.5. *Metals Data File (CRYSTMET, 1998)*

CRYSTMET is the National Research Council of Canada Metals Crystallographic Data File, covering metals, alloys and intermetallics. An original compilation of Don Cromer and Alan Larson was taken to NRC, Ottawa, Canada, by the latter in 1974. The database was developed and extended in collaboration with the late Larry Calvert and, since 1981, with John Rogers. In 1996, NRC assigned its interest in CRYSTMET (1998) to Toth Information Systems. Of the 45 000 structures currently available in CRYSTMET and noted in Table 1, 26 000 are structures assigned on the basis

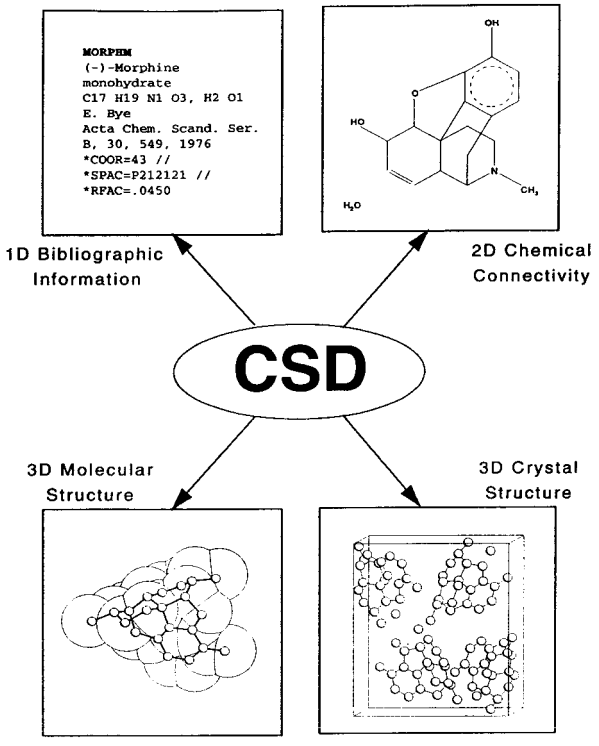


Fig. 1. Information content of entries in the Cambridge Structural Database. The information content of other crystallographic databases can be related to this chart, as indicated in the text.

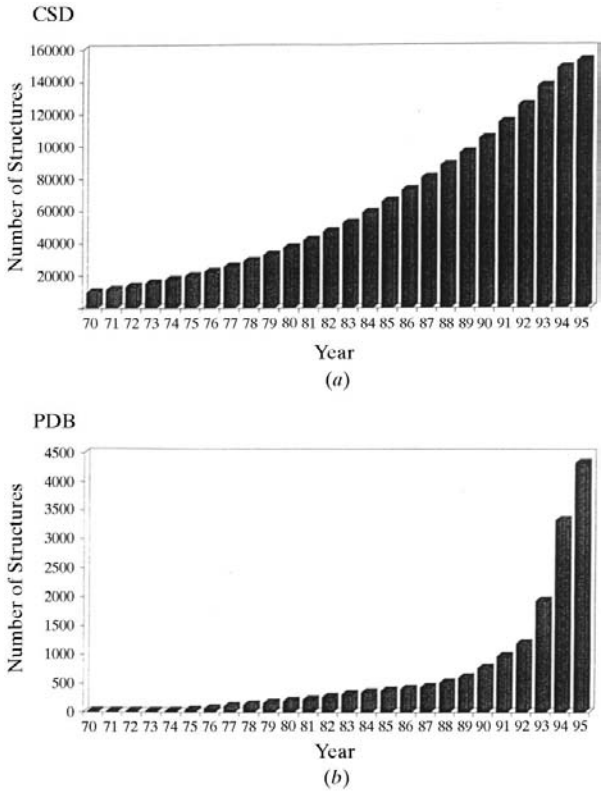


Fig. 2. Cumulative growth of (*a*) the Cambridge Structural Database (CSD) and (*b*) the Protein Data Bank (PDB) over the period 1970–1995.

of isotypism. Again, CRYSTMET contains the bibliographic and crystallographic data categories of Fig. 1, but formal connectivity is not applicable.

2.6. *Data acquisition and validation*

For a database to be of true scientific value, it must meet the twin criteria of *completeness* and *accuracy*. With the exception of the PDB, where completeness is essentially implicit through its role in the publication process, the other crystallographic databases reflect the published literature (although the CSD does contain some 'private communications' where authors have no intention to proceed to formal printed publication). All of the databases make great efforts to be complete and act as official depositories of numerical crystal structure data for many major journals. All of the databases perform extensive validation checks on new input data and will usually repeat those checks if validation standards change in any way. Indeed, one of the scientific uses of the accumulated data is to derive or upgrade basic standard values, such as covalent and van der Waals radii for atoms, standard geometrical features of functional groups or structure types *etc.* Apart from their own intrinsic interest, these standards are fundamental to validation procedures. It is important also that the results of (unresolved) validation checks are clearly recorded in the databases. This allows the user choice over which entries to include in any particular database analysis.

Historically, much database input has been re-keyboarded from original literature or supplementary 'deposition documents', hence the validation checks were originally conceived to eliminate local keyboarding errors. However, it soon emerged that the printed literature contained significant numbers of errors in the primary crystallographic data. Although time consuming to untangle, the vast majority of these errors can be classed as typographical, ranging from simple errors of manual data entry or transcription to more complex cases (*e.g.* the misassociation of coordinate sets with cells and space groups from other structures *etc.*). However, it is rare for validation processes to uncover 'wrong' structures – a very positive reflection on the checks and balances involved in formal publication mechanisms. The accuracy of the databases has also been significantly enhanced by the cooperation of the original authors and by the highly constructive feedback received from end-users, who have reported typographical and structural errors over many years. Thus, the crystallographic databases represent value-added products in the hands of their users, both in terms of the accuracy of the recorded data and through the storage of additional searchable information that is derived from the primary input, *e.g.* structure classifiers, protein secondary structure descriptors, bit encoded information to enhance search speeds *etc.*

2.7. *Influence of the crystallographic databases on publication mechanisms*

The crystallographic databases are commonly used to avoid duplication of effort, both on the part of experimentalists and journals. This simple but important check, usually based on crystal data together with available chemical information, has already saved significant amounts of data-collection time and computational resources, as well as valuable refereeing and editorial time for journals that wish to decline duplicate structure determinations. These simple checks will become progressively more valuable as the number of structures continues to increase.

However, the most important effect of the data-validation procedures was to draw attention to the accuracy of numerical crystallographic results reported in the primary literature. The retyping of lengthy tables of coordinates and structural geometry, often at several stages in the publication process, resulted in the occurrence of at least one numerical error in 15% of all published papers or deposition documents. By the early 1980s, many journals began to require the deposition of original computer outputs and, at the end of the decade, the IUCr Editorial Office began to perform numerical checks on incoming material. From the mid-1980s, suggestions that journals and databases should accept information in computerized form became more frequent. However, the plethora of formats and electronic media then in use mitigated against a large-scale implementation of electronic input. The Crystallographic Information File (CIF: Hall *et al.*, 1991) has addressed these concerns for small molecules and inorganic structures, and the CIF concept has spread to cover macromolecular structures and beyond, as decribed elsewhere in this Special Issue by Hall (1998).

In practical terms, the inclusion of CIF generators in the most popular structure-determination packages, together with rapid growth of Internet communications during the 1990s, has greatly aided the uptake of CIF. Major journals increasingly require deposited information in CIF format, and increasingly use the databases as their official depositories. As a result, 40% of current CSD input arrives in CIF form. However, CIF deposition systems are not yet perfect: whilst CIF generators will produce files of high integrity with respect to crystallographic data items that are known to, or are generated by, the software package, some data items must still be incorporated by manual editing to complete the CIF for onwards transmission. Principally these are bibliographic data, descriptive text and chemical details, and any lack of conformance with CIF dictionary rules causes problems for interpretation software. It is to be hoped that local or Internet-accessible software checks of CIF integrity will soon become the norm before files are deposited with journals and databases.

One of the more difficult areas for a molecular database such as the CSD is to encode the 2D chemical information that is so crucial to database searchability and research use. An important CSD validation check is that the 2D connectivity record corresponds to the connectivity deduced from the 3D coordinate set using standard bonding radii. If not, there may be coordinate errors or the bonding radii must be adjusted to account for rarer bonding situations, usually involving metal centres. CSD data-processing software can assign chemical connectivity directly from geometrical information but it is reassuring to know that the resulting representation is correct, especially as chemical structures become ever more complex. While *Acta Crystallographica* requires the mandatory presence of a 2D chemical diagram, many other journals do not, especially if crystallographic detail is consigned to a brief deposition document. In the future, it should be possible to incorporate 2D chemical information within electronic depositions, through use of data names already in the CIF dictionary, or using those from the CIF companion molecular information file (MIF: Allen *et al.*, 1995).

2.8. *Statistical inferences and future publication scenarios*

Fig. 2 shows the growth rate of the CSD compared with that of the PDB for the period 1970–1995. With a current total of 186 000 structures and a doubling period of approximately 7 years, we may predict that the CSD will contain at least 500 000 structures by the year 2010. The PDB archive has a current total of 8000 structures and a doubling period of about 2 years. This rapid growth rate is reminiscent of the CSD of the early 1970s and will almost certainly abate over time. Nevertheless, by the year 2010, we might expect the PDB to contain upwards of 100 000 structures. Growth rates in the inorganic and metals areas appear to be a little slower than for the CSD, corresponding to a doubling period of about 8–9 years, and yielding year 2010 totals of *ca* 140 000 structures in each case (including assigned structures).

A conservative estimate of the joint holdings of the crystallographic databases in 2010 is well in excess of 800 000 structures. Put more starkly, authors, journals and the crystallographic databases must process almost twice as many structures in the next 12 years as they have in the previous 30 years. It may be more than that. The data collection and computer technologies of the 1990s provide structures in hours rather than days or weeks, a quantum increase that parallels the wholesale move away from photographic methods and limited computing resources that began some 30 years ago. Note that authors are also challenged by these rapid developments. It is clear that many laboratories already have backlogs of crystal structures awaiting publication and the scientific community is thereby deprived of results of lasting value. How, then, will we all cope with the output of increasingly efficient machines over the next decade?

Electronic data-deposition mechanisms, associated with scientific text that may appear electronically or in printed form, are currently being put in place by many major journals. Unfortunately, at the submission level, slightly different procedures are being adopted by different journals, while some still appear to be ignoring the electronic revolution. Thus, crystallographic publication scenarios are currently in a state of flux and the databases must cope with this transition state. More importantly, however, the databases must continue to work with journals and authors alike, so as to contribute their experience to the ongoing developments. Issues to be addressed include: the information content and CIF integrity of electronic submissions, standardization of submission and validation procedures involving journals and databases, the role of the expert referee in assessing novel structural work, consideration of the databases as primary publication media, and the academic status that will be accorded to such contributions. The pace of change in communications technology is unrelenting and poses challenges for all scientists; hence these and other issues must be debated sooner rather than later. Structural crystallography looks forward to an exciting and highly productive future and it is vital that the results of this increased productivity continue to be captured efficiently and effectively for the long-term benefit of the scientific community.

3. Software systems for database access

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. (J. D. Bernal)

3.1. *Basic search facilities*

All of the crystallographic databases provide software systems for search, retrieval and visualization of stored information and full details can be obtained from the references cited above or directly from the appropriate data centre. All of the systems provide for direct searching and browsing on the basis of text and simple numerical fields – the types of information denoted as 'one-dimensional' in Fig. 1 – and can retrieve a variety of structural information files on the basis of such searches, principally coordinate sets and/or geometrical information. In most systems, searches of individual data fields can be combined into an overall query using Boolean logic or, in the case of relational models, using the structured query language (SQL).

In the molecular databases – the CSD and the PDB – the formal descriptions of chemical structure, the 2D chemical connection tables of the CSD and the amino-acid sequence information of the PDB, are also available

for search purposes, and it is the relationship between chemical and geometrical structures that underpins many research applications. Amino-acid sequence data are encoded as one-dimensional text strings and are searched using extensions of basic text-string-matching techniques. It is only for the diverse molecular species recorded in the CSD that we need resort to graph-theoretic 2D search techniques. Since the CSD comprises more than two thirds of current database holdings, and the vast majority of its research applications begin with a substructure search, it is appropriate to review these search methods in a little more detail.

3.2. *Substructure searching at the molecular and supramolecular levels*

The schema for the current (Version 5) CSD software system is illustrated in Fig. 3. Apart from providing the basic search mechanisms noted above, the *Quest3D* program also provides extensive search facilities at both the molecular and supramolecular (extended crystal structure) levels, as illustrated in Fig. 4. The key to these searches is the graph-theoretical mapping of the formal 2D chemical structure representation onto the connectivity of the 3D crystal structure. This procedure can be considered as adding the formal chemical description of atoms and bonds to the experimentally determined 3D structure, so that both 2D chemical criteria and geometrical criteria derived from the 3D coordinates can be applied in the search process. Thus, *Quest3D* is able (*a*) to perform 'standard' substructure searches, in which the query fragments are wholly defined by chemical formalisms (Fig. 4*a*), (*b*) to extend the process to intramolecular atomic arrangements that are defined by both chemical and geometrical criteria (such as the pharmacophoric pattern of Fig. 4*b*) and, most importantly, (*c*) to further extend the process to the supramolecular (crystal structure) level (Fig. 4*c*) through a similar combination of chemical and user-defined geometrical search criteria to permit the location and study of specific noncovalent interactions.

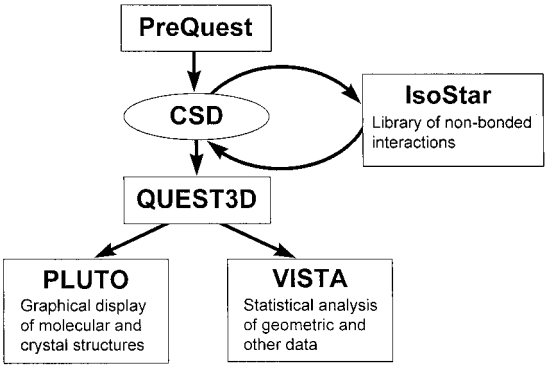


Fig. 3. The Cambridge Structural Database System, comprising the CSD itself, software for its creation (*PreQuest*) and for search, analysis and display (*Quest3D*, *VISTA* and *PLUTO*). The knowledge-based library, *IsoStar*, is regularly updated from the growing data content of the CSD.

3.3. *Post-processing of search results*

The crystallographic databases all provide their own visualization modules, *e.g. PLUTO* in the CSD System of Fig. 3, or provide links to other commonly available packages. In the case of the NDB, structural knowledge, in the form of stored geometrical characteristics can also be displayed (Berman *et al.*, 1992). The post-processing of numerical information retrieved by database searches can either be performed within the database software packages themselves or through links to external software. This software can range from commercially available graphical, statistical and modelling packages to one-off local programs written for specific research applications.

This is particularly true for ICSD and CRYSTMET, where current research applications concentrate on studies of *e.g.* interatomic distances, structural similarities, isotypism, standard representations of isotypes, structural topologies, and classifications of coordination polyhedra (Bergerhoff *et al.*, 1998, and references therein). Many of these classifiers will probably become incorporated into the databases as additional 'derived' data items to improve searchability. Research applications of the 'nonmolecular' databases are developing, and there is no doubt that users of these databases are beginning to provide the ideas and tools for systematization in inorganic structural chemistry. As indicated by Bergerhoff *et al.* (1998), this work will be significantly advanced by improved descriptions of structural organization – akin to the connectivity representations of molecular chemistry, by the availability of 3D geometrical (structural) search facilities and by the routine application of appropriate statistical and numerical analysis techniques to geometrical descriptions of inorganic structures. The bond-valence approaches and systematic studies of Brown & Altermatt (1985), Brown

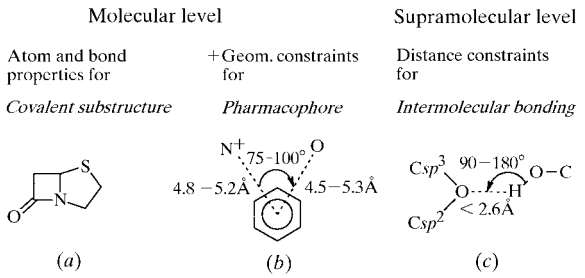


Fig. 4. Substructure search capabilities of the CSD System at the molecular and supramolecular levels using chemical-bonding conventions explicit in the 2D chemical connectivity records of Fig. 1 and geometrical criteria derived from the 3D coordinate information.

(1992, 1997), Brese & O'Keefe (1992) and others are beginning to provide the necessary basis for the development of a systematic description of inorganic structures.

By contrast, the molecular databases have seen very extensive research use in the past twenty years. Until recently, the small numbers of protein and nucleic acids structures made it possible to select the required structure(s) from printed lists of database holdings. However, in the last decade, search and browse mechanisms have been made available for these archives (Berman *et al.*, 1992; Abola *et al.*, 1997). Systematic studies of protein structure have also depended on commercial or public domain software products, or on software written by its users. Often these systematic studies have involved a transformation of the PDB into other database or knowledge base structures (see *e.g.* Islam & Sternberg, 1989; Thornton & Gardner, 1989).

This survey of software capability presents a somewhat fragmentary picture. However, the primary goal of each database group is to create and manage a numerical archive of crystallographic information. The development of applications software was either outside of original funding remits or a secondary issue which was only seriously addressed as the processes of database creation and management became stabilized. The development of CSD System software followed this course, with the first distributed system (Allen *et al.*, 1979) becoming available nearly 15 years after the inception of the CSD archive. Since that time, the development of applications software has formed an integral part of CCDC's operations, so that software facilities have been continuously upgraded in response to the needs of researchers in academia and industry. Nevertheless, it takes time to build distributed software systems with acceptable user interfaces. The Version 5 System (Fig. 3) released in 1993 was the first CSD System to provide integrated searching of all information fields up to the extended structure level exemplified in Fig. 4(*c*). Even then the integrated system is unable to answer each and every query posed to the database, or to process adequately the information that is retrieved. External software packages and local code or scripts still have a role to play in the more complex analyses of CSD information.

4. Research applications

> *Crystals are windows on the world of atoms.* (Chet Raymo)

The aim of basic research using the crystallographic databases is to convert sets of retrieved facts or data items into scientific knowledge. This process can be divided into three conceptual stages: (*a*) definition and retrieval of those facts that best describe the problem under study; (*b*) selection of an appropriate method for the systematic analysis of those facts; and (*c*) the scientific interpretation of the analytical results. The inorganic research applications noted above certainly involve all of these stages. However, their principal aim is to provide fundamental classifications of the data, *i.e.* a systematic description of the structural chemistry of inorganic compounds that will underpin future research applications. The standard electron-pair model of chemical bonding already provides that fundamental description of structure within the molecular databases, and the importance of formal 2D chemical descriptions within the CSD has already been stressed. For this reason, and because of the fundamental significance of molecular, supramolecular and biological structures to modern science, it is the CSD and the PDB that have so far been most widely used as bases for fundamental research.

It is clearly impossible to provide a thorough overview of the many hundreds of research applications of the CSD and the PDB. Nor is this necessary, the 888 pages of the recently published book *Structure Correlation* (Bürgi & Dunitz, 1994) perform this task in a definitive manner, while other books published during the last decade (Desiraju, 1989; Jeffrey & Saenger, 1991; Glusker *et al.*, 1994) provide further examples of database analyses in specific fields. Further, two other articles in this Special Issue address major areas of database applications: structure correlation (Bürgi, 1998) and supramolecular structural organization and crystal engineering (Nangia & Desiraju, 1998). Quite rightly, all of these publications concentrate on the scientific knowledge that can be deduced from database analyses – the stage (*c*) referred to above. The present section will therefore concentrate more on stages (*a*) and (*b*): the processes of search and retrieval of appropriate facts and methods of analysis. At the end, we will summarize briefly those broad areas of structural science that have benefited most from these approaches, before indicating likely future directions for crystallographic information provision in the molecular area. Most of the example material will be drawn from the CSD, since it offers an integrated software framework for experiments in data mining. However, these principles and methods have clear general applicability across all of the databases.

4.1. *Information retrieval from the CSD*

Within the context of molecular chemistry, structural knowledge is most simply expressed in terms of geometrical descriptions of structure that are derived from the underlying coordinate information. The vast majority of CSD applications are concerned with systematic analyses of the geometrical characteristics of subsets of related chemical substructures. The CSD System program *Quest3D* (Fig. 3) will calculate a set of user-defined geometrical descriptors for each instance of

a substructure that is located during the search process. A very wide range of geometrical descriptors is available, ranging from interatomic distances, valence angles and torsions to more complex parameters that involve mid-points and centroids of atomic groupings, ring-puckering parameters and the directionalities of noncovalent interactions. Simple constructs permit any of these parameters to be modified (*e.g.* to obtain the absolute value of a signed quantity) or combined (*e.g.* to obtain the mean value of a number of basic parameters). Often, the choice of the most suitable parameters for a particular problem is an iterative process, the best set emerging during the processes of data analysis and interpretation. The file of information retrieved by *Quest3D* is a matrix of geometrical data, $\mathbf{G}(N, p)$, containing the p user-defined geomerical descriptors for each of the N substructural fragments located during the search process. Visualization and analysis of this matrix is central to the vast majority of data-mining experiments.

4.2. *Analysis and visualization of geometrical information*

The CSD program *VISTA* provides a wide variety of facilities for the visualization and analysis of geometrical matrices generated by *Quest3D*. Visualizations of univariate or bivariate distributions are provided by histograms and scattergrams referred to Cartesian or polar axes. Fig. 5 illustrates a polar histogram of the torsion angle TOR defined for the substructure shown.

A wide variety of statistical and numerical methods can be applied to the analysis of geometrical data. Initially, interest focused on the analysis of univariate distributions, *e.g.* in the derivation of mean bond lengths, using simple descriptive statistics: means, medians, standard deviations of means and samples, and upper and lower percentiles. Regression analysis was also applied to bivariate distributions, for example in studying the variation of one parameter with another (structure–structure correlation), or the relationships between geometry and physical properties (structure–property correlations). In 1978, Murray-Rust and colleagues (Murray-Rust & Bland, 1978; Murray-Rust & Motherwell, 1978) recognized the importance of analysing a complete **G** matrix comprising $p > 3$ geometrical descriptors, *e.g.* in the systematic analysis of molecular conformations, and introduced multivariate statistical methods such as principal component analysis and cluster analysis for this purpose. There have been many developments over the years, particularly through the application of other statistical techniques and the adaptation of these methods to crystallographic usage. This area has recently been extensively reviewed (Taylor & Allen, 1994).

4.3. *Principal areas of research interest*

4.3.1. *Mean molecular dimensions*. The work of Pauling (1939) represented the first systematic attempt to derive mean values for bond lengths and valence angles from the limited structural data available at that time. This work resulted in the definition of covalent bonding radii for the common elements and had a seminal influence on the development of chemistry over the past half century. Further tabulations appeared sporadically until the publication in 1956 and 1959 of the major compilation *Tables of Interatomic Distances and Configuration in Molecules and Ions*, edited by L. E. Sutton (1959, 1965) for the Chemical Society of London. In the mid-1980s, the CCDC and its collaborators compiled updated tables of mean bond lengths for both organic (Allen *et al.*, 1987) and organometallic and metal-coordination compounds (Orpen *et al.*, 1989). For each bond length, both compilations present the mean, its estimated standard deviation and the sample standard deviation, together with the median value of the distribution, and its upper and lower quartile values. The organic section describes 682 discrete chemical bond types involving 65 element pairs, while the organometallic and metal complex compilation presents similar statistics for 325 different bond types involving *d*- and *f*-block metals. More recently, Engh & Huber (1991) have generated sets of mean bond lengths and valence

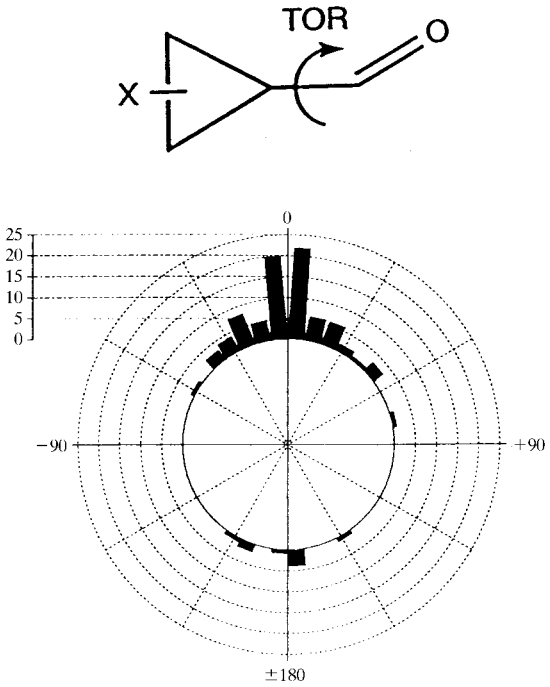


Fig. 5. Polar scatterplot (*VISTA*) of the torsion angle TOR (O═C—C—C—*X*, *X* is mid-point of bond), showing a preference for the *cis*-bisected conformation of the carbonyl substituent with respect to the cyclopropane ring.

angles from peptidic structures retrieved from the 80 000 entries then available in the CSD. Their compilations are based on 31 atom types, which are most appropriate to the protein environment and are well represented in CSD structures. Apart from their intrinsic value in structural science, a knowledge of standard molecular dimensions is important in the determination, refinement and validation of novel structures, particularly in protein crystallography.

4.3.2. *Conformational analysis*. It is a simple matter to use the CSD System to display the distribution of crystallographically determined torsion angles about a specific bond, as shown in Fig. 5. It is also possible to study the inter-relationship of two torsion angles by plotting them on a conventional 2D scattergram. In this case, we may hope to map conformational interconversion pathways or to identify areas of high population density that correspond to conformational preferences. The multivariate statistical methods noted above permit this process to be extended to conformational problems of higher dimensionality, *i.e.* to problems that must be defined using more than three torsion angles. Work of this type falls under the topic of structure correlation and is fully referenced elsewhere in this Special Issue (Bürgi, 1998). Obviously, crystallographic conformations represent energetically accessible forms and the message from these correlation experiments is that there is a clear qualitative relationship between the observed conformational distributions derived from the condensed-phase crystallographic data and the major features of the computed (*in vacuo*) potential energy hypersurface. This is clearly illustrated in Fig. 6 for two example substructures (Allen *et al.*, 1996). Results of this type for a series of 12 substructures enabled these authors to assess the value of crystal structure information within a molecular modelling environment as follows: (*a*) torsion angles with higher strain energies (> 4 kJ mol^{-1}) are rare in crystal structures, occurring in less than 5% of cases; (*b*) the effects of crystal packing on conformation appear to be the exception rather than the rule; and (*c*) crystal structure observations are good guides to the conformational preferences of isolated molecules and are a valuable adjunct to computational methods, especially in 'difficult' systems, *i.e.* those for which force-field or other computational parameters are not readily available.

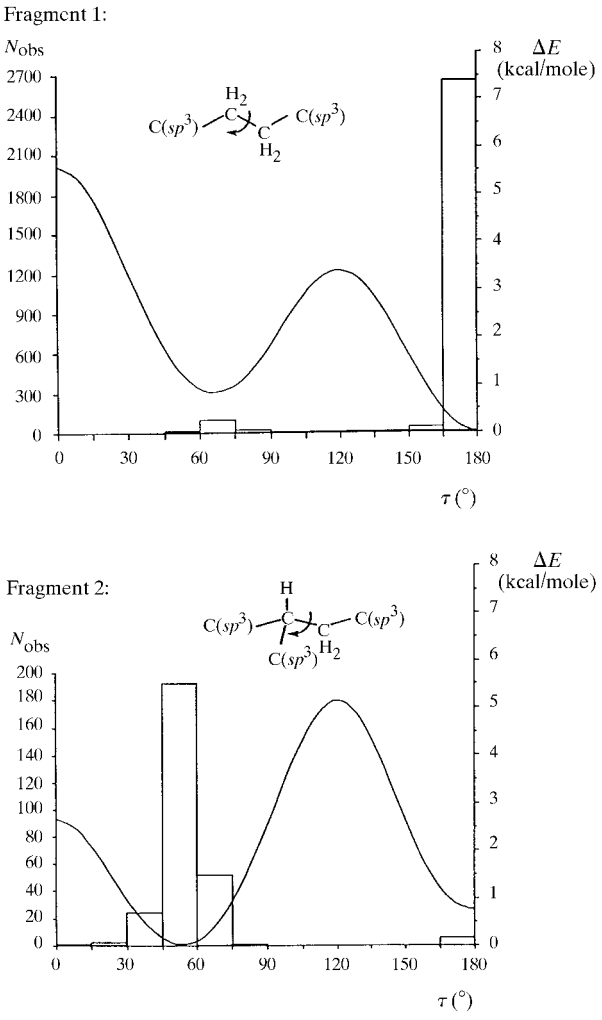


Fig. 6. Torsional distributions from crystallographic results (CSD data) compared with energy profiles computed using 6-31G* basis sets for two of the 12 substructures studied by Allen *et al.* (1996).

4.3.3. *Studies of noncovalent interactions*. A crystal structure is the archetypal supermolecule and crystal structure analysis is the only technique that provides routine experimental observations of the types and geometries of noncovalent interactions. Because of this uniqueness, crystallographic database analyses are providing scientific knowledge that is crucial in many areas, *e.g.* in the determination and validation of new protein structures, in structure-based drug design and the study of protein–ligand interactions and in crystal engineering and the design of novel materials with predictable structures and properties.

Using the nonbonded search modules of *Quest3D* and the graphical features of *VISTA*, it is now a simple matter to retrieve and display the geometry of specific noncovalent interactions. This process is illustrated for O—H···O hydrogen bonds by the intermolecular search fragment of Fig. 7(*a*), which leads to the graphical displays of appropriate geometrical descriptors provided in Fig. 7(*b*)–(*d*). Apart from strong hydrogen bonds, the CSD has been used to study a variety of weaker hydrogen bonds, *e.g.* *X*—H···S=C, C—H···O, *X*—H···Metal and *X*—H···π, together with a variety of interactions that are not mediated by hydrogen, *e.g.* halogen···halogen, O···halogen and carbonyl···carbonyl interactions. The importance of database analyses in identifying and describing the most important and reproducible noncovalent interactions is

highlighted more extensively elsewhere in this Special Issue (Nangia & Desiraju, 1998).

However, despite the undoubted value of crystallographic analyses of noncovalent interactions, these studies can only provide a very qualitative estimate of their relative strengths. Recent research (*e.g.* Nobeli *et al.*, 1997; Lommerse *et al.*, 1996) has now shown the value of combining CSD analyses with high-level *ab initio* molecular-orbital calculations, for example the intermolecular perturbation theory (IMPT) of Hayes & Stone (1983). This technique is computationally intensive but the examination of the potential-energy hypersurface can be effectively guided by the crystallographic data: the calculations being restricted to geometrical locations that are close to highly populated areas of the crystallographic data distributions.

5. The future: structural knowledge bases

> *Knowledge is of two kinds. We know a subject ourselves, or we know where we can find information upon it.* (Samuel Johnson)

So far this paper has discussed the primary numerical data content of the crystallographic databases – atomic coordinates, cell dimensions, *etc*. – and shown how these data can be retrieved and analysed to generate systematic structural knowledge at the molecular and supramolecular levels. The article has also shown that this process, now being termed knowledge discovery or data mining, has made a significant impact in many areas of structural science. In the late 1980s, the CCDC began to collect references to scientific applications of the CSD and today this collection (which is publicly available as the *DBUSE* module of the distributed CSD System) comprises more than 600 references. Many of these published analyses have demanded considerable expertise for their execution and, while it is valuable to have easy access to their literature citations, it would be much better to have direct access to the structural knowledge generated by these studies. This is not a novel concept. Structural knowledge is a component part of the NDB (Berman *et al.*, 1992), specific knowledge-based systems have already been derived from the PDB (*e.g.* Islam & Sternberg, 1989; Thornton & Gardner, 1989) and the value of knowledge-based approaches to protein structure determination and molecular modelling have been discussed by Allen *et al.* (1990). In 1995, the CCDC began a programme to derive libraries of structural knowledge from the raw data content of the CSD. The first of these libraries – *IsoStar*: a library of information on intermolecular interactions – is briefly summarized below. A second library, containing torsional distributions and other features of intramolecular geometry, is currently under development. It is planned that these libraries will be updated regularly, so as to keep pace with the increasing size of the CSD.

5.1. *IsoStar: a library of nonbonded interactions from the CSD and the PDB*

The amount of data about noncovalent interactions in the CSD is vast, and this information has wide applicability. Further, the type of information that is required is relatively standard: having identified an interaction of interest, we wish to know details of its geometry and directional properties, as depicted in Fig. 7. To provide structured and direct access to a comprehensive set of derived information, a knowledge-based library of nonbonded interactions (*IsoStar*: Bruno *et al.*, 1997) has been developed at the CCDC. *IsoStar* is based on experimental data, not only from the CSD but also from the PDB, and contains some theoretical results calculated using the IMPT method. Version 1.0 of *IsoStar* was released October 1997 and contains information on nonbonded interactions formed between 277 common functional groups, referred to as *central groups*, and 28 *contact groups*, *e.g.* hydrogen-bond donors, water, halide ions *etc.* Information is displayed in the form of scatterplots for each interaction. Version 1.0 contains 6683 scatterplots: 5296 from the CSD and 1387 from the PDB. *IsoStar* also reports results for 867 theoretical potential-energy minima. For a given contact between a central group (A) and a contact group (B), CSD search results are transformed into an easily visualized form by overlaying the A moieties. This results in a 3D distribution (scatterplot) showing the experimental distribution of B around A, which can be manipulated and viewed interactively using the *RASMOL* visualizer (Sayle, 1996). Fig. 8(a) shows an example of a scatterplot: the distribution of OH groups around carboxylate anions, illustrating hydrogen-bond formation along the lone-pair directions of the carboxylate O atoms. The *IsoStar* software also enables the user to quickly inspect the original crystal structure in which a specific contact occurs *via* a hyperlink to the original CSD entries. Another tool generates contoured surfaces from scatterplots, which show the density distribution of the contact groups. Contouring aids the interpretation of the scatterplot and the analysis of preferred geometries. Fig. 8(b) shows the contoured surface of the scatterplot in Fig. 8(a): the lone-pair directionality of the hydrogen bonds becomes even more obvious. The PDB scatterplots in *IsoStar* only involve interactions between noncovalently bound ligands and proteins, *i.e.* side-chain–side-chain interactions are excluded. The *IsoStar* library contains data derived from almost 800 complexes having a resolution better than 2.5 Å. Fig. 8(c) shows the distribution of OH groups around carboxylate groups as derived from the PDB; the close similarity between Fig. 8(c) (PDB) and Fig. 8(a) (CSD) is obvious.

5.2. *A library of torsional distributions derived from the CSD*

The CSD can be viewed as a huge library of individual molecular conformations. However, to be of general value, it is necessary to distil, store and present this knowledge in an ordered manner, in the form of torsional distributions for specific atomic tetrads *A–B–C–D*. Protein-specific libraries of this type, derived from high-resolution PDB structures, are commonly used as aids to protein structure determination, refinement and validation. Torsional information can either be stored in external databases or hardwired into a program in the form of rules. CSD usage has tended to concentrate on analyses of individual substructures, both for their intrinsic interest and as testbeds for the development of new data-analysis techniques. However, Klebe & Mietzner (1994) have recently described the generation of a small library containing 216 torsional distributions derived from the CSD, together with 80 determined from protein–ligand complexes in the PDB. The library was used in a knowledge-based approach for predicting multiple conformer models for putative ligands in the computational modelling of protein–ligand docking.

As part of its knowledge-base development programme, the CCDC has just embarked on the generation of a more comprehensive torsional library (Cole *et al.*, 1998). Here, information is being hierarchically ordered according to the level of specificity of the chemical substructures for which torsional distribu-

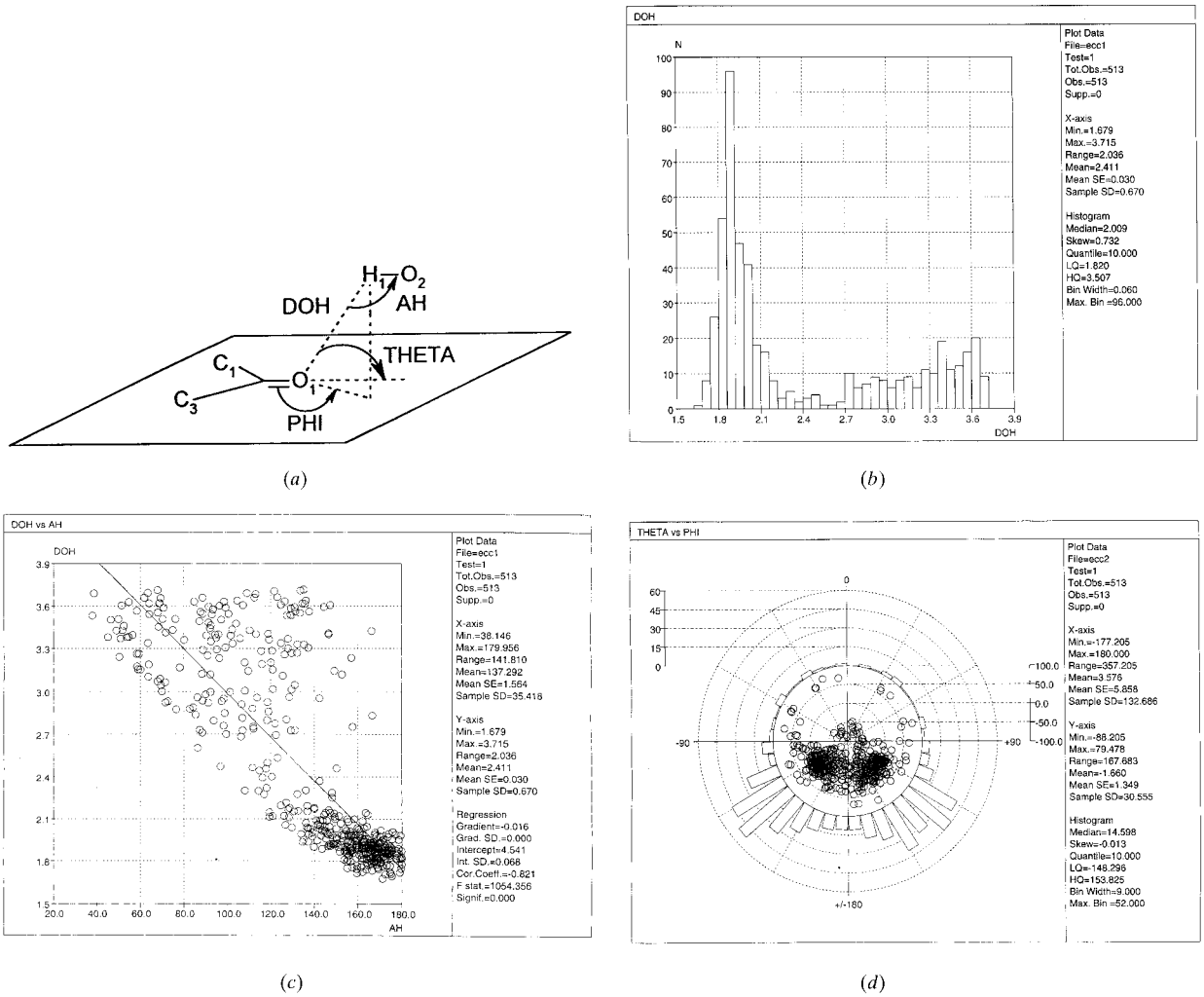
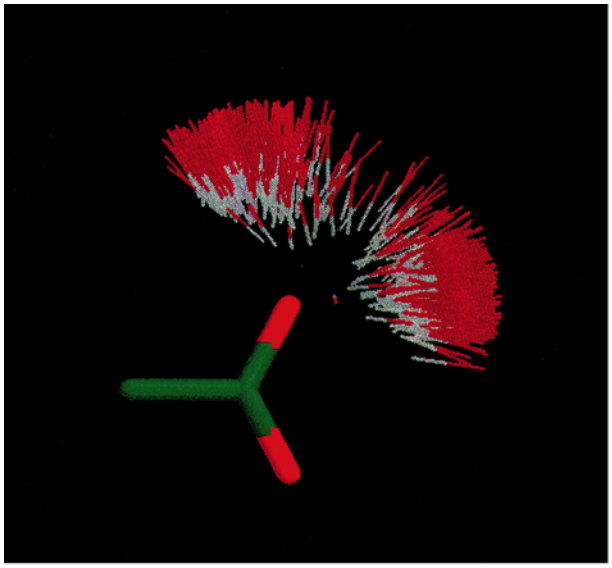


Fig. 7. (*a*) Definition of a supramolecular substructural query and associated geometrical parameters for the study of O—H···O (keto) hydrogen bonds. A limiting value of DOH normally acts as the primary geometrical constraint. Limiting values for other geometrical parameters can also be set. THETA and PHI define the direction of approach of donor-H to acceptor-O, thus THETA 0°, |PHI| 120° is indicative of lone-pair directionality at O. (*b*) Histogram of the >C═O···H hydrogen-bond length DOH, (*c*) plot of DOH *versus* *A*H, the angle O···H—O, and (*d*) polar scatterplot of THETA *versus* PHI showing lone-pair directionality at acceptor-O. Plots were generated using *VISTA* (Fig. 3)

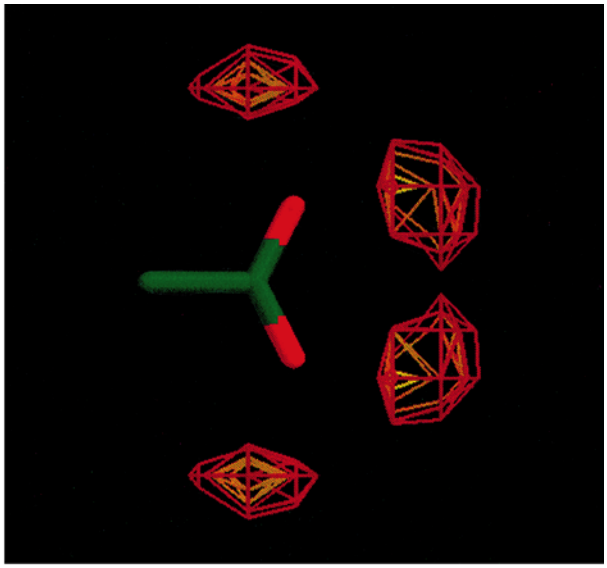
tions are available in the library. So far, 6.8 million torsion angles have been computed for possible inclusion in the library and software for retrieval and display of torsional distributions is under development.

5.3. *Software applications of structural knowledge bases*

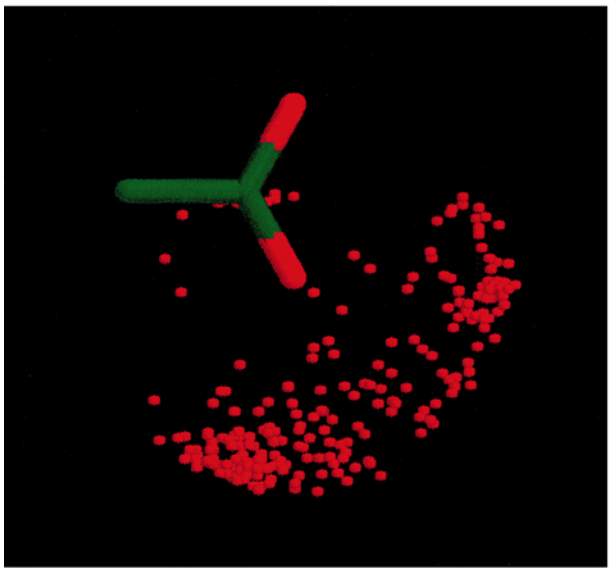
Just as the information in the crystallographic databases can be accessed by applications software, so the structural information stored in knowledge bases such as *IsoStar* can be accessed by software that is designed to solve problems in structural chemistry and biology. Currently, the CCDC is involved in the development of two approaches to the prediction of protein–ligand interactions. The program *GOLD* (Jones *et al.*, 1997) is a genetic algorithm approach which also uses conformational information from the CSD and nonbonded interaction information from *IsoStar*. The program *SuperStar* (Verdonk & Taylor, 1998) attempts to use only the experimental knowledge stored in *IsoStar* to generate 'preferred interaction' surfaces for the binding of the functional groups of a specific ligand to a protein active site. It seems likely that structural knowledge will also play a major role in solving the structures of flexible molecules from powder diffraction data (see *e.g.* Shankland *et al.*, 1998).



(*a*)



(*b*)



(*c*)

Fig. 8. *IsoStar* scattergrams: (*a*) O—H donors (contact group) around a carboxylate-O acceptor (central group) using CSD data, (*b*) contoured density plot of the distribution in (*a*), and (*c*) distribution of O—(H) oxygen atoms around carboxylate-O using PDB data.

6. Conclusions

Science is built up of facts, as a house is built of stones; but an accumulation of facts is no more a science than a heap of stones is a house. (Henri Poincaré)

This short article has attempted to cover the developmental history of the crystallographic databases and to summarize their scientific applications. In particular, it has described their influence and impact in a wide range of fields including crystallographic publication mechanisms, chemical informatics and structural science. The latter is by far the most important area of impact, by virtue of the very wide range of structural knowledge that can be obtained from data-mining experiments. While most of the examples cited here have arisen from molecular and supramolecular science, there is no doubt that crystallographic databases are also beginning to play a major role in developments in inorganic structural chemistry.

The article has also tried to look to the future, not only by raising issues relating to the publication of crystallographic results and their ultimate capture by the databases, but also by indicating how knowledge-base development will change and improve the provision of structural information to the end user. It is now very clear that software access to crystallographic information must be at two levels: a raw-data level and a derived-knowledge level. The onward development of structural knowledge bases from the underlying data provides for the preservation, storage and regular updating of the results of data-mining experiments, thus avoiding repetition of standard experiments and providing instant access to complex derivative information. Most importantly, a suitably structured knowledge base can be acted on by software tools designed to solve complex problems in structural chemistry. The availability of knowledge bases derived from experimental observations is likely to be a crucial factor in the solution of those two analogous and currently intractable problems in the small-molecule and protein-structure domains: crystal structure and polymorph prediction on the one hand and protein folding on the other.

It is to be hoped, at least, that this brief summary has shown that some sort of scientific edifice is being created from the accumulation of facts that reside in the crystallographic databases!

References

Abola, E. E., Sussman, J. L., Prilusky, J. & Manning, N. O. (1997). *Methods Enzymol.* **277**, 556–571.

Allen, F. H., Barnard, J. M., Cook, A. P. F. & Hall, S. R. (1995). *J.Chem. Inf. Comput. Sci.* **35**, 412–427.

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T. W. A., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *Acta Cryst.* B**35**, 2331–2339.

Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. G. (1991). *J. Chem. Inf. Comput. Sci.* **31**, 187–204.

Allen, F. H., Harris, S. E. & Taylor, R. (1996). *J. Comput. Aided Mol. Des.* **10**, 247–254.

Allen, F. H., Kennard, O., Watson, D. G., Brammer, L., Orpen, A. G. & Taylor, R. (1987). *J. Chem. Soc. Perkin Trans. 2*, pp. S1–S19.

Allen, F. H., Rowland, R. S., Fortier, S. & Glasgow, J. I. (1990). *Tetrahedron Comput. Methodol.* **3**, 757–774.

Bergerhoff, G., Berndt, M., Brandenburg, K. & Degen, T. (1998). Unpublished.

Bergerhoff, G., Hundt., R., Sievers, R. & Brown, I. D. (1983). *J. Chem. Inf. Comput. Sci.* **23**, 66–69.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinavasan, A. R. & Schneider, B. (1992). *Biophys. J.* **63**, 751–759.

Brese, N. E. & O'Keefe, M. (1992). *Acta Cryst.* A**48**, 663–669.

Brown, I. D. (1969–1981). *BIDICS – Bond Index to the Determination of Inorganic Crystal Structures.* Institute for Materials Science, McMaster University, Hamilton, Ontario, Canada.

Brown, I. D. (1992). *Acta Cryst.* B**48**, 553–572.

Brown, I. D. (1997). *Acta Cryst.* B**53**, 381–393.

Brown, I. D. & Altermatt, D. (1985). *Acta Cryst.* B**41**, 244–247.

Bruno, I. J., Cole, J. C., Lommerse, J. P. M., Rowland, R. S., Taylor, R. & Verdonk, M. L. (1997). *J. Comput. Aided Mol. Des.* **11**, 525–537.

Bürgi, H.-B. (1998). *Acta Cryst.* A**54**, 873–885.

Bürgi, H.-B. & Dunitz, J. D. (1994). *Structure Correlation.* Weinheim: VCH Publishers.

Cambridge Structural Database (1998). Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England.

Cole, J. C., Kessler, M. & Taylor, R. (1998). Private communication.

CRYSTMET (1998). Toth Information Systems Inc., 2045 Quincy Avenue, Gloucester, Ontario K1J 6B2, Canada.

Desiraju, G. R. (1989). *Crystal Engineering: the Design of Organic Solids.* New York: Academic Press.

Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–398.

Ewald, P. P. (1948). *Acta Cryst.* **1**, 1–2.

Ewald, P. P & Hermann, C. (1929). *Strukturbericht 1913–28.* Leipzig: Akademische Verlagsgesellschaft.

Glusker, J. P., Lewis, M. & Rossi, M. (1994). *Crystal Structure Analysis for Chemists and Biologists.* Weinheim: VCH Publishers.

Gray, N. A. B. (1986). *Computer-Assisted Structure Elucidation.* New York: Wiley.

Hall, S. R. (1998). *Acta Cryst.* A**54**, 820–832.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *Acta Cryst.* A**47**, 655–685.

Hayes, I. C. & Stone, A. J. (1983). *J. Mol. Phys.* **53**, 83–105.

Inorganic Crystal Structure Database (1998). ICSD, Fachinformationszentrum Karlsruhe, D-7514 Eggenstein-Leopoldshafen, Germany.

International Centre for Diffraction Data (1998). ICDD, 12 Campus Boulevard, Newtown Square, PA 19073–3273, USA.
Islam, S. E. & Sternberg, M. J. E. (1989). *Protein Eng.* **2**, 431–442.
Jeffrey, G. A. & Saenger, W. (1991). *Hydrogen Bonding in Biological Structures.* Berlin: Springer Verlag.
Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. (1997). *J. Mol. Biol.* **267**, 727–748.
Kennard O. & Allen, F. H. (1993). *Chem. Des. Autom. News*, **8**, 1, 31–37.
Klebe, G. & Mietzner, T. (1994). *J. Comput. Aided Mol. Des.* **8**, 583–594.
Lommerse, J. P. M., Stone, A. J., Taylor, R. & Allen, F. H. (1996). *J. Am. Chem. Soc.*, **118**, 3108–3116.
Murray-Rust, P. & Bland, R. (1978). *Acta Cryst.* B**34**, 2527–2533.
Murray-Rust, P. & Motherwell, W. D. S. (1978). *Acta Cryst.* B**34**, 2534–2546.
Nangia, A. & Desiraju, G. R. (1998). *Acta Cryst.* A**54**, 934–944.
National Institute of Standards & Technology (1998). Gaithersburg, MD 20899, USA.
Nobeli, I., Price, S. L., Lommerse, J. P. M. & Taylor, R. (1997). *J. Comput. Chem.* **18**, 2060–2074.
Nucleic Acid Data Bank (1998). NDB, Department of Chemistry, Rutgers University, PO Box 939, Piscataway, NJ 08855–0939, USA.
Orpen, A. G., Brammer, L., Allen, F. H., Kennard, O., Watson, D. G. & Taylor, R. (1989). *J. Chem. Soc. Dalton Trans.* pp. S1–S83.
Pauling, L. (1939). *The Nature of the Chemical Bond.* Ithaca, NY: Cornell University Press.
Protein Data Bank (1998). PDB, Biology Department, Brookhaven National Laboratory, PO Box 5000, Upton, NY 11973–5000, USA.
Sayle, R. (1996). *RASMOL – a Program for Structure Visualization.* Glaxo Wellcome Research Centre, Stevenage, Hertfordshire, England.
Shankland, K., David, W. I. F., Csoka, T. & McBride, L. (1998). *Int. J. Pharm.* **165**, 117–126.
Sutton, L. E. (1959). *Table of Interatomic Distances in Molecules and Ions.* Spec. Publ. No. 11. London: The Chemical Society.
Sutton, L. E. (1965). *Table of Interatomic Distances in Molecules and Ions.* Spec. Publ. No. 18. London: The Chemical Society.
Taylor, R. & Allen, F. H. (1994). *Statistical and Numerical Methods of Data Analysis.* In *Structure Correlation*, edited by H.-B. Bürgi & J. D. Dunitz. Weinheim: VCH Publishers.
Thornton, J. M. & Gardner, S. P. (1989). *Trends. Biochem. Sci.* **14**, 300–304.
Verdonk, M. L. & Taylor, R. (1998). In preparation.