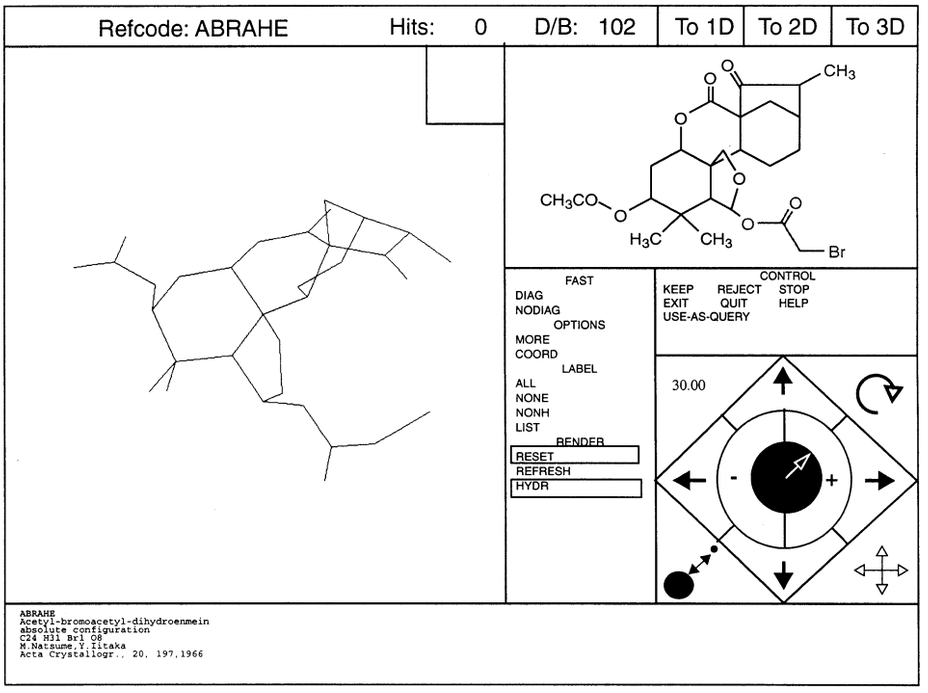

Fig. 24.3.3.3. A typical *Quest3D* graphics screen showing how search hits are visualized and manipulated.

24.3.4. Knowledge engineering from the CSD

24.3.4.1. *Databases versus knowledge bases*

As illustrated in earlier sections, the CSD represents a collection of primary *data* resulting from diffraction experiments on crystals of small molecules – in particular the fractional coordinates, space group and cell dimensions that define the 3D crystal and molecular structure. However, the user of the system is usually interested in structural *knowledge* – in the form of bond lengths, angles, intermolecular contact distances and other parameters – that can be synthesized from the raw data by the use of the CSD system software. Thus, each detailed analysis carried out using the CSD system represents an experiment in data mining, and considerable operational and intellectual effort is employed in performing such analyses.

At the present state of development of the field, three facts are apparent: (*a*) many data-mining activities centre around a set of standard geometrical data types that are essential for major applications, particularly in structural chemistry, molecular modelling and rational drug design; (*b*) the expertise required to carry out data-mining experiments is not inconsiderable and the time required can be lengthy; and (*c*) as the size of the CSD is increasing rapidly and any compilations of structural knowledge should be updated on a regular basis, the increasing database size makes this operation very time consuming for individual users.

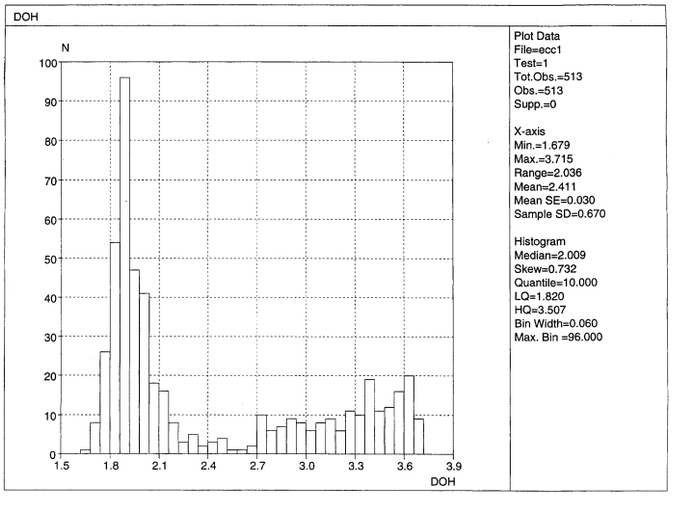

Fig. 24.3.3.4. A *Vista* histogram of the hydrogen-bond distance, DOH, showing a sharp peak in the range 1.8–2.2 Å, well below the sum of van der Waals radii (2.62 Å). This peak can be isolated in *Vista* to obtain an estimate of the mean O···H separation in >C=O···H—O systems.

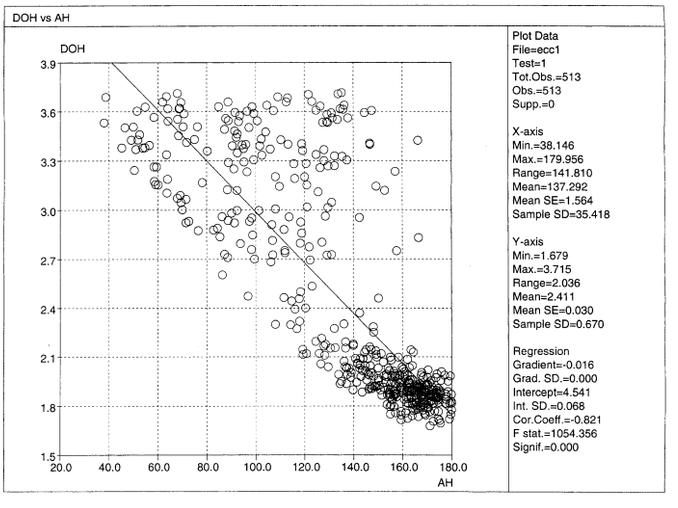

Fig. 24.3.3.5. A *Vista* scatterplot of the hydrogen-bond length (DOH) *versus* the O—H···O angle (AH). The plot shows a major clustering of observations having short DOH values and hydrogen-bond linearity (AH = 180°): stronger hydrogen bonds prefer to be linear.

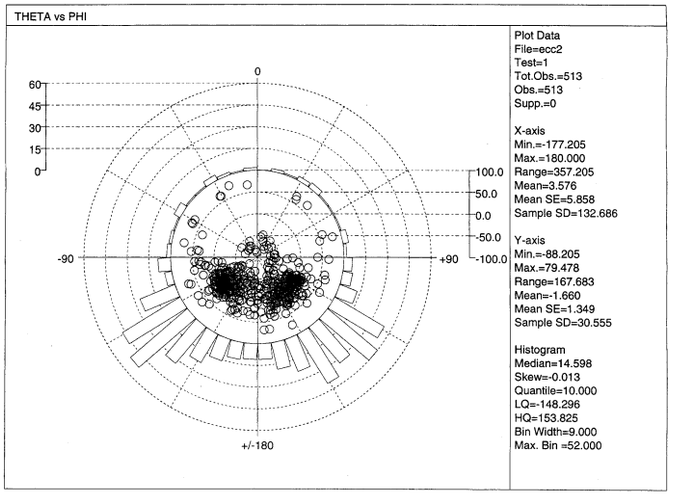

Fig. 24.3.3.6. A *Vista* polar scatterplot of THETA *versus* PHI, the angles that define the direction of approach of the donor H atom to the >C=O plane. There are clear indications of lone-pair directionality: H prefers an in-plane approach to O (THETA = 0°), with preferred PHI values in the range 120–135°.

These considerations indicate that access to CSD information should be at two levels, the raw-data level and the structural-knowledge level, and, since 1995, the CCDC has started 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 (Bruno *et al.*, 1997) – is briefly summarized below. A second library, Mogul, containing bond lengths, valence angles and torsional distributions, is currently under development in mid-2000. Such a knowledge base has obvious applications in crystallography, structural chemistry and molecular biology, not least in providing precise geometrical parameters which can be used in 3D model building, structure refinement and reality checking of developing and refined structures. The scientific applications of non-bonded contact geometries and conformational (torsional) information are more fully discussed in Chapter 22.4.

24.3.4.2. *IsoStar: a library of knowledge about intermolecular interactions*

IsoStar (Bruno *et al.*, 1997) is based on experimental data, not only from the CSD but also from the PDB, and contains some theoretical results calculated using the *ab initio* intermolecular perturbation theory (IMPT) method of Hayes & Stone (1984). The experimental data in the CSD and the PDB have been used to display interaction geometries involving *central groups* (*A*) and *contact groups* (*B*). CSD search results of the type exemplified above 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*. A web-browser front end permits rapid access to these scatterplots, which can be viewed in *RasMol* (Sayle, 1996), interrogated interactively, converted into contoured surfaces *etc.*

Version 1.1 of IsoStar, released in October 1998, contains information on non-bonded interactions formed between 310 central groups and 45 contact groups. Version 1.1 contains over 12 000 scatterplots: 9000 from the CSD and 3000 from the PDB. IsoStar also reports results for 867 theoretical potential-energy minima calculated using the IMPT procedure. The library will be updated on a regular basis using automated software procedures developed at the CCDC.

Chapter 22.4 contains illustrative examples from IsoStar, together with a more complete description of the knowledge base and its applications.

24.3.5. Accessing the CSD system and IsoStar

24.3.5.1. *Release mechanisms*

The CSD system, comprising the CSD and CSD-Use databases, and all of the applications software described above, is available on CD-ROM for Unix and DEC-VMS platforms and for PCs operating under Linux. At the time of writing (mid-2000), *ConQuest* alone is available for PC-Windows platforms, but the full availability of other components of the CSD system is currently being addressed. The CSD is released twice yearly, in April and October, as an indexed sequential binary file, with full installation instructions contained within the CD. Versions of the CSD have been reformatted for use with proprietary software systems: the MACCS3D/Isis system from Molecular Design Limited, and the Sybyl-UNITY system from Tripos Associates.

Subscribers in academic and other not-for-profit institutions may obtain the CSD system through their local National Affiliated Centre (NAC). The names, addresses and other coordinates of these centres are contained in the CCDC's web pages (see below). Users in countries not covered by NAC arrangements, or users from for-profit companies and organizations, should contact the CCDC directly.

The IsoStar library, for Unix systems only, is released after each library update, currently planned to occur on an annual basis. IsoStar forms part of the distributed CSD system, and CDs are available through the same mechanisms as the main system.

24.3.5.2. *Information about the CCDC*

The CCDC maintains an extensive set of information on the web site http://www.ccdc.cam.ac.uk. The site describes the CSD system, IsoStar, and the associated research and development activities of the CCDC. These pages also provide access to CSD system documentation, provide lists of contact details for the National Affiliated Centres that service not-for-profit users worldwide, and give up-to-date information on how to contact the CCDC directly.

24.3.6. Conclusion

This chapter has provided an overview of the Cambridge Structural Database, its associated software system, other databases and IsoStar – the first library of structural knowledge to be derived from the CSD. This information is accurate at the time of writing (May 1998, but with some revision in mid-2000). However, the CSD itself, the knowledge bases derived from it and a wide variety of applications software are under continuous development and improvement, and articles such as this can only provide a snapshot of progress at any particular time. Readers of this article are therefore encouraged to visit the CCDC's website (http://www.ccdc.cam.ac.uk) to obtain the latest information on available products and services.