International Tables for Crystallography (2006). Vol. F, Section 24.3.3, pp. 665–666.

ity mapping that finally confirms (or not) the presence of the required query substructure.

24.3.2.7. *Data validation*

All data entering the CSD are subject to stringent check and evaluation procedures. Some of these are visual, but the majority are automated within the CSD program *PreQuest*. The checks ensure that the 1D and 2D information fields abstracted by CCDC staff are accurately encoded, and that the 3D crystallographic coordinates are consistent with both the chemical description of the structure and with the geometrical description supplied by the authors. Most typographical errors in original papers can be corrected by the CCDC but, in the case of serious discrepancies, the original authors are consulted.

24.3.2.8. *The CSD-Use database*

CSD-Use is a database of scientific research papers in which the CSD was used as the principal or sole source of experimental information. The database comprises more than 700 literature citations classified according to the type of systematic study undertaken. Each CSD-Use entry also contains a short summary of the major findings of the research. The database is growing rapidly over time, and is expected to be a valuable resource in the future, since it contains a fully retrospective overview of the data-mining methods and research applications of the CSD.

24.3.3. The CSD software system

24.3.3.1. *Overview*

The CSD is supplied with a suite of fully interactive graphical software modules which provides users with facilities to: (*a*) interrogate all of the 1D, 2D and 3D information fields; (*b*) display entries graphically in a variety of styles; (*c*) retrieve relevant data for search hits, including geometrical parameters derived from the stored coordinates; and (*d*) display the derived numerical information, *e.g.* as histograms, scattergrams *etc.*, generate descriptive statistics and perform more complex numerical analyses. More recently, software has been added that permits users to transform their own in-house structural data to CSD formats for inclusion in these processes. A summary of the overall CSD software system is given in Fig. 24.3.3.1 which shows the functional relationships between the four major applications programs.

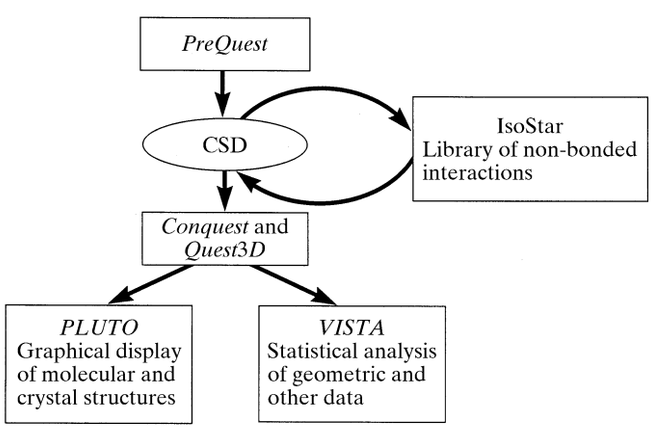

Fig. 24.3.3.1. Summary of the software components of the Cambridge Structural Database system (CSDS).

24.3.3.2. *PreQuest*

PreQuest is a data-validation and data-conversion program which is used to create high-quality structural data files in CSD format from, *e.g.*, raw input data from a CIF. *PreQuest* is used routinely by CCDC's scientific editors to create and validate entries for inclusion in the master CSD archive, hence the program is constantly being maintained and upgraded. The released version enables users to build a private CSD-format database of their own structures which can then be searched independently of, or in conjunction with, the master CSD files using the database access programs described below.

24.3.3.3. *Searching the CSD: Quest3D and ConQuest*

Quest3D has been the main search engine and information-retrieval program for the CSD since the late 1980s. Its main features are summarized below. However, since 1997, the CCDC has been developing its successor, the *ConQuest* program, which was first released as part of the CSD system in April 2000. During an interim period, perhaps two years, *ConQuest* and *Quest3D* will both form part of the released CSD system on certain computing platforms while the functionality of the new program is being fully developed. Further details of *ConQuest* are provided in Section 24.3.3.5, indicating in particular how it differs from, and improves upon, the facilities available in *Quest3D*.

24.3.3.4. *Quest3D*

Quest3D is the main search engine and information-retrieval program for the CSD. It permits interrogation of all information fields: (*a*) 19 text fields, (*b*) 38 individual numerical fields, (*c*) element symbols and element counts, (*d*) full or partial molecular formulae, (*e*) direct access to over 150 bit screens, (*f*) extensive 2D chemical substructure search capabilities, and (*g*) 3D substructure searching at the molecular level or at the extended crystal-structure level. A search of a specific information field is termed a *test* of that field, and is constructed graphically *via* the menu system; menu components correspond to the categories of searches identified above. A complete *query* is then constructed by combining a number of separate *test* components using Boolean logic.

Substructure searching is the most important and frequently used facility. At the molecular level, the substructure (chemical fragment) query is entered graphically and is defined using the formal covalent bond types present in the 2D chemical connectivity tables of the CSD. The process can be extended to locate non-bonded contacts in the complete crystal structure. Here, the individual atoms or chemical groups involved in the contact must be specified, and a limiting non-bonded contact distance must be provided, along with any other geometrical criteria required to define the contact more precisely.

All substructure searches begin with the user drawing the required chemical unit(s) *via* the BUILD menu. Chemical variability and precision are controlled through (*a*) the PERIODIC TABLE sub-menu, which allows for specification of variable element types at specific atomic sites, (*b*) the 2D-CONSTRAIN menu, which allows further chemical restrictions to be specified, such as cyclicity/acyclicity of bonds, exact hydrogen-atom counts, total coordination numbers for atoms *etc.*, and (*c*) the 3D-CONSTRAIN menu, which permits the user to specify a list of geometrical parameters to be calculated by the program for each instance of the fragment located in the CSD; any of these geometrical parameters may be used as criteria to limit the scope of the search, especially at the intermolecular level. A file of calculated geometrical information is output by *Quest3D* and may be read by *Vista*, or by external data analysis software. Other

Quest3D output files allow CSD search results to be communicated rapidly to proprietary modelling software.

24.3.3.5. *ConQuest*

The overall aim of the *ConQuest* project is to replace *Quest3D* with graphical search software that makes best use of modern computing environments. The primary objective has been to create an interface that is both simple and intuitive to use, so as to encourage use of the CSD by a broader spectrum of scientists. Thus, *ConQuest* provides: (*a*) text and numeric searches *via* pop-up windows, (*b*) a new sketcher window within which to encode 2D and 3D substructure searches, pharmacophore searches, and searches for non-bonded contacts in crystal structures, and (*c*) the immediate viewing of hits with facilities for backward and forward scrolling within hit lists. *ConQuest* is provided with full documentation and tutorials, both online and in printed form, and with context-dependent help facilities. Version 1.0, released in April 2000, contains most of the functionality available within the *Quest3D* program, and it is expected that *Quest3D* capabilities will soon be exceeded by the new program.

A most important feature of *ConQuest* is its availability on PC-Windows platforms, as well as its implementation under Unix/Linux. Initially, *ConQuest* and the CSD will be the only parts of the full CSD system available under PC-Windows, but *Vista* (or *Vista*-like facilities, see Section 24.3.3.6), a new visualizer and provision of the CCDC's knowledge bases (IsoStar and Mogul) will follow as planned developments in the PC area.

24.3.3.6. *Vista*

Vista reads geometrical table(s) generated by *Quest3D* and provides extensive facilities for the graphical representation and statistical analysis of the numerical data. Graphical facilities include histograms and scattergrams referred to Cartesian or polar axes, with a hyperlink back to the original CSD entries to permit immediate investigation of, *e.g.*, outlying observations. The contents of plots can be edited interactively, and all illustrations can be output in PostScript format for inclusion in reports and publications. Additionally, *Vista* will generate descriptive statistics for a distribution, carry out simple linear regressions and perform principal-component analyses.

24.3.3.7. *Pluto*

Pluto is used to visualize crystal and molecular structures in a variety of styles, including stick diagrams and ball-and-spoke and space-filling representations of individual molecules or extended crystal structures.

24.3.3.8. *Use of the CSD software system: an example*

The preceding sections can only give a flavour of the extensive search, analysis and visualization capabilities of *Quest3D*, *ConQuest*, *Vista* and *Pluto*, which are fully documented in manuals available online *via* the web address given below, or in printed form from the CCDC.

In this section, we illustrate the application of the CSD system to one specific example: a CSD-based analysis to examine the O—H···O hydrogen-bonding ability of the keto oxygen of Fig. 24.3.3.2. This example illustrates a number of key features of the software system. The example is constructed in terms of *Quest3D* terminology, but identical facilities are available in the *ConQuest* program.

(1) Draw the two component substructures: the keto group and the O—H donor group. Constrain the *total coordination number* of C_1, C_3 (Fig. 24.3.3.2) to be 4, thus defining them as $C(sp^3)$ atoms.

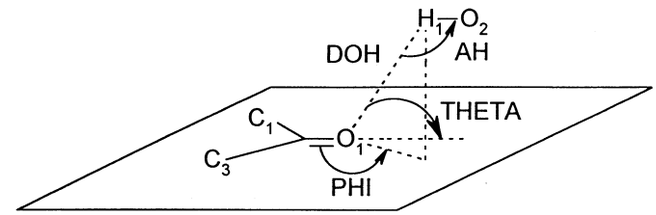

Fig. 24.3.3.2. The keto···hydroxyl fragment described in the example of CSDS usage (see Section 24.3.3.8), illustrating the parameters DOH, AH, THETA and PHI used to describe the hydrogen-bonded system.

(2) Define a non-bonded contact between keto O_1 and hydroxy donor H_1. Require that this contact (DOH) is less than 2.62 Å, the sum of van der Waals radii, after normalization of the H-atom position to correspond to a standard O—H bond length as determined by neutron diffraction [X-ray location of H atoms is imprecise – *X*—H distances are usually foreshortened – so the system will reposition H atoms along the *X*—H vector and at an *X*—H distance that corresponds to the mean value from neutron diffraction experiments (Allen *et al.*, 1987)].

(3) Define the geometrical parameters shown in Fig. 24.3.3.2, comprising the H···O distance (DOH), the O—H···O angle (AH), and the angles THETA and PHI that describe the angle of approach of H to the putative lone-pair plane of the keto oxygen atom. THETA is the angle of approach of the donor H atom to the plane of the keto group, PHI is the angle of rotation of the projection of the O···H vector in that plane; THETA = 0°, PHI = ±120° would correspond to H-atom approach along an O-atom lone-pair direction. The search is further constrained so that hits are only accepted if AH > 90°.

(4) At this stage, the 3D-CONSTRAIN menu will show a graphic which closely resembles Fig. 24.3.3.2. Test 1 is now defined.

(5) Since there will be large numbers of examples of keto-O···H—O hydrogen bonds in the CSD, a secondary constraint based on the crystallographic *R* factor is applied so that examples are only located in the more precise structure determinations. To do this, we access the NUMERIC search menu to define RFACT < 0.075 as test 2.

(6) Enter the QUEST menu, which summarizes all current tests, select the *organic structures only* bit screen, and complete the full query by combining test 1 and test 2 *via* a Boolean .AND. operator.

Searches can be performed interactively or allowed to run to completion without further intervention from the user. In interactive mode, *Quest3D* presents each hit as it is located, as illustrated in Fig. 24.3.3.3, and can then display the 1D bibliographic information, a 2D structural diagram, the 3D molecular structure, or a 3D packing diagram by toggling between display options. For an intermolecular search, as exemplified here, the non-bonded contact that triggered the hit is clearly identified. For the example described above, a file of the four user-defined geometrical parameters (DOH, AH, THETA, PHI) for each hit is created for use by *Vista*.

Vista displays the geometrical parameters in the form of an interactive spreadsheet; the user may include or exclude specific substructures on the basis of numerical criteria during the data analysis, *e.g.* to focus on a specific range of DOH values, exclude outlying observations *etc*. Hyperlinking between *Vista* and the master CSD file means that all of the database information of Fig. 24.3.1.1 is immediately available during a *Vista* session, either by clicking on a particular fragment in the spreadsheet or on a particular data point in a histogram or scattergram. Use of *Vista* is illustrated for the >C=O···H—O example in Figs. 24.3.3.4, 24.3.3.5 and 24.3.3.6.

references