International Tables for Crystallography (2006). Vol. F, Section 22.4.3, pp. 559–560.

for primary publication or hard-copy deposition. Thus, the PDB has always acquired data through direct deposition in electronic form, and authors have usually been involved in the validation of their entries. Further, it is a mandatory requirement of the vast majority of journals, and a clear recommendation of appropriate professional organizations, that prior deposition with the PDB is an essential precursor to primary publication. This key involvement of the PDB in the publication process acts as a vital guarantee of the completeness of the archive. The prior-deposition rule must be rigidly adhered to for the long-term benefit of science.

22.4.2.3. *Standard formats: CIF and mmCIF*

The CSD, on the other hand, reflects the published literature, and much of its data content has been re-keyboarded from hard-copy material. The Cambridge Crystallographic Data Centre (CCDC) is now beginning to receive significant amounts of electronic input, a development that owes much to the rapid international acceptance of an agreed standard electronic interchange format, the crystallographic information file or CIF (Hall *et al.*, 1991), and the rapid incorporation of CIF generators within most major structure solution and refinement packages. The CIF offers many advantages, some of which are only just being addressed within the CSD: (*a*) a clear definition of input data items and their representation; (*b*) a significant reduction in time spent correcting simple typographical errors; and (*c*) the possibility of enhancing the overall database content through the electronic availability of *all* information from the analysis, *i.e.* more than could reasonably be re-typed from hard-copy material. For the PDB, the recent adoption of the macromolecular CIF (mmCIF) as the agreed international standard offers similar advantages. This development, together with advances in communications technology, now make it possible to automate the deposition process more effectively, but the advantages of mmCIF can only be fully realized once it also becomes a standard output format of all of the relevant software packages.

22.4.2.4. *Structure validation*

The value of research results derived from the CSD and the PDB depends crucially on the accuracy of the underlying data [see *e.g.* Hooft *et al.* (1996) with respect to protein data]. As with the early CSD, much current research involves use of data from the developing PDB to establish rules and protocols for the validation of new protein structures (see *e.g.* Laskowski *et al.*, 1993). This activity, in turn, means that earlier entries in the archive may have to be reassessed periodically to bring their representations into line with best current practice. This sequence of events was commonplace in the CSD of the 1970s and, even now, new structure types entering the CSD can still provoke a reassessment of subclasses of earlier entries.

Secondly, it is important that errors and warnings raised by validation software have clear meanings and that validation results are clearly encoded within each entry. The end user can then make informed choices about which entries to include (or not) in any given application. Recent moves to apply a range of agreed and unambiguous primary checks to new data, and to require resolution of any problems prior to the issue of a publication ID code, represent an important development.

22.4.3. Structural knowledge from the CSD

22.4.3.1. *The CSD software system*

Structural knowledge from the CSD is reflected principally in the geometries of individual molecules, extended crystal structures and, most importantly, through systematic studies of the geometrical characteristics of large subsets of related substructural units. Software facilities for search, retrieval, analysis and visualization of CSD information are fully described in Chapter 24.3. The system allows for the calculation of a very wide range of geometrical parameters, both intramolecular and intermolecular. Most importantly, chemical substructural search fragments may be specified using normal covalent bonding definitions (single, double, triple *etc.*), limiting non-covalent contact distances and other geometrical constraints. For each instance of a search fragment located in the CSD, the system will compute a user-defined set of geometrical descriptors. The full matrix, $G(N, p)$, of the p geometrical parameters for each of the N fragments located in the CSD can then be analysed using numerical, statistical and visualization techniques to display individual parameter distributions, to compute medians, means and standard deviations, and to examine the geometrical data for correlations or discrete clusters of observations that may exist in the p-dimensional parameter space.

22.4.3.2. *CSD structures and substructures of relevance to protein studies*

Table 22.4.3.1 presents statistics for the 3137 structures of amino acids and peptides that are available in the CSD of April 1998 (containing 181 309 entries). Although this represents less than 2% of CSD information, some may consider that these are the only entries of real interest in molecular biology. In certain cases, *e.g.* for the derivation of very precise molecular dimensions and for some conformational work, this may be true. However, the real issue concerns the *transferability* of CSD-derived information to the protein environment. It is the biological relevance of a chemical

Table 22.4.3.1. *Summary of amino-acid and peptide structures available in the CSD (April 1998, 181 309 entries)*

(*a*) Overall statistics

Structures	No. of entries
α-Amino acids (any organic) *	3137
Peptides (standard or modified standard α-amino acids) †	1430

(*b*) Peptide statistics

	No. of CSD entries	
No. of residues	Acyclic	Cyclic
2	543	123
3	249	45
4	76	50
5	62	44
6	20	73
7	14	15
8	19	32
10	16	19
11	4	10
12	2	11
13	—	—
14	1	—
15	3	2
16	3	—

* Any organic structure containing the α-amino acid functionality.

† The standard amino acids (those normally found in proteins) may be modified by substitution in these peptides.

Table 22.4.3.2. *CSD entry statistics for selected metal-containing structures*

CSD entries ($R < 0.10$) containing M and (N or O). No additional transition metals were allowed to occur in the Na, K, Mg and Ca structures cited.

Metal	No. of CSD entries
Na	1189
K	987
Mg	510
Ca	469
Zn	1996

substructure (inter- or intramolecular) that is important, and this consideration immediately brings much larger subsets of CSD entries into play. Information such as van der Waals radii can be derived from the CSD as a whole, while more specific information concerning, for example, biologically important metal coordination geometries can be derived from appreciable subsets of the total database, as shown in the statistics of Table 22.4.3.2.

22.4.3.3. *Geometrical parameters of relevance to protein studies*

Precise geometrical knowledge from atomic resolution studies of small molecules is important in the macromolecular domain since it provides: (*a*) geometrical restraints and standards to be applied during protein structure determination, refinement and validation; (*b*) model geometries for liganded small molecules and information about their preferred modes of interaction with the host protein; (*c*) details of metal coordination spheres and geometries that are likely to be observed in metalloproteins; and (*d*) information from which force field and other parameters may be derived. Thus, the types of study discussed in this chapter are concerned with retrieving systematic knowledge concerning:

(1) molecular dimensions: bond lengths and valence angles;

(2) conformational features: torsion angles that describe acyclic and cyclic systems;

(3) metal coordination-sphere geometries: coordination numbers, metal–ligand distances and inter-ligand valence angles;

(4) general non-bonded contact distances: van der Waals radii;

(5) hydrogen-bond geometries: distances, angles, directional properties;

(6) other non-bonded interactions: identification and geometrical description;

(7) formation of preferred atomic arrangements or motifs involving non-covalent interactions.

In this short overview, which deals with such a broad range of structural information, our literature coverage is, of necessity, highly selective. In each area, we have tried to cite the more recent papers, from which leading references to earlier studies can be located. We also draw attention to a number of recent monographs in which a variety of CSD analyses are comprehensively cited and discussed: *Structure Correlation* (Bürgi & Dunitz, 1994), *Crystal Structure Analysis for Chemists and Biologists* (Glusker *et al.*, 1994), *Hydrogen Bonding in Biological Structures* (Jeffrey & Saenger, 1991) and *Crystal Engineering: the Design of Organic Solids* (Desiraju, 1989). Finally, we note the CCDC's own database of published research applications of the CSD. The DBUSE database currently contains literature references and short descriptive abstracts for nearly 700 papers. It forms part of each biannual CSD release and is fully searchable using the *Quest3D* program.

22.4.4. Intramolecular geometry

22.4.4.1. *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 Sutton (1956, 1959), by The Chemical Society of London. Kennard (1962) extended the available data for bonds between carbon and other elements.

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). Both compilations were based on the CSD of September 1985 containing 49854 entries. Of these, 10324 organic structures and 9802 organometallics or metal complexes satisfied a variety of secondary selection criteria, and were used in the analysis. 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. Of these, 511 (75%) involve carbon, and 428 (63%) involving only carbon, nitrogen and oxygen are relevant to protein studies. The organometallic and metal complex compilation presents similar statistics for 325 different bond types involving *d*- and *f*-block metals. It is planned to automate and systematize the production of such tabulations, so that they can be dynamically updated in computerized form, as part of CCDC's ongoing development of knowledge-based structural libraries.

More recently, Engh & Huber (1991) have generated sets of mean bond lengths and valence angles from peptidic structures retrieved from the 80000 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. These authors note that such knowledge, together with torsional and other information, is vital to the determination, refinement and validation of protein structures. Prior to their detailed CSD analysis, some of the parameters used for these purposes had been determined with a lower accuracy than was required by the diffraction data. For this reason, and particularly for use with higher-resolution protein data, they recommend that the most accurate parameters possible should always be used.

Systematic use of CSD data generates mean values together with standard deviations for both the sample and the mean. The sample standard deviations provide information about the spread of each parameter distribution, *i.e.* information about the variability of each parameter which can be parameterized as force constants. Comparative refinements of selected proteins showed that the new CSD-based parameters yielded significant improvements in R factors and in geometry statistics. Finally, Engh & Huber (1991) remark that their results should be updated regularly as the quantity and quality of data in the CSD increase with time. Apart from producing more precise estimates of mean values, incorporation of more protein-relevant atom types into the schema should then be possible.

22.4.4.2. *Conformational information*

Torsion angles are the natural measures of conformational relationships within molecules. If we specify a chemical substructure involving a central bond of interest, then the CSD system

references