*International Tables for Crystallography* (2006). Vol. F, Section 22.4.3.1, p. 559.



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 |
|---|---|
| $\alpha$-Amino acids (any organic) * | 3137 |
| Peptides (standard or modified standard $\alpha$-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 $\alpha$-amino acid functionality.
† The standard amino acids (those normally found in proteins) may be modified by substitution in these peptides.





references