International Tables for Crystallography (2006). Vol. F, Section 22.4.4, pp. 560–562.

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 *Quest*3*D* 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

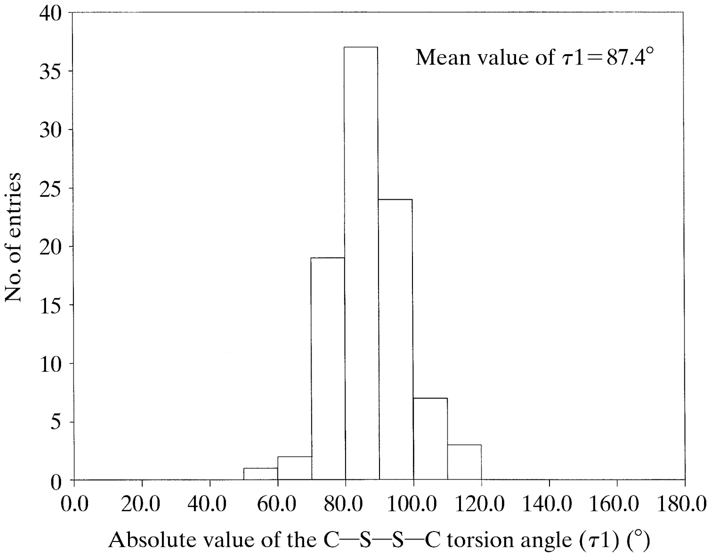

Fig. 22.4.4.1. Distribution of torsion angles in C(sp^3)—S—S—C(sp^3) substructures located in the CSD.

will display the distribution of torsion angles about that bond, computed from the tens, hundreds, or even thousands of instances located in the database. Examination of these *univariate distributions* will reveal any conformational preferences that may exist in small-molecule crystal structures. This approach is illustrated by the histogram of Fig. 22.4.4.1, which shows the torsional distribution about S—S bridge bonds in C(sp^3)—S—S—C(sp^3) substructures located in the CSD. Clearly, there is a preference for a perpendicular conformation in the CS—SC unit. This corresponds well with values observed for cysteine bridges in protein structures, and with theoretical calculations on small model compounds.

The interrelationship between two torsion angles can be visualized by plotting them against each other on a conventional 2D scattergram. In the small-molecule area, the distribution of data points in these scattergrams can reveal conformational interconversion pathways (Rappoport *et al.*, 1990) or show areas of high data density corresponding to conformational preferences (Schweizer & Dunitz, 1982). The best known *bivariate distribution* is the Ramachandran plot of peptidic φ–ψ angles, which is universally used to assess the quality of protein structures and to identify structural features. Ashida *et al.* (1987) performed an extensive analysis of peptide conformations available in the CSD and present torsional histograms, a Ramachandran plot, and a variety of other visual and descriptive statistics that summarize this data set.

It is often necessary to use three or more torsion angles to define the conformation of, *e.g.*, a side chain or flexible ring. Here, *multivariate* statistical techniques (Chatfield & Collins, 1980; Taylor, 1986) have proved valuable for extracting information from the matrix $T(N, k)$ that contains the k torsion angles computed for each of the N examples of the substructure in the CSD. Two methods, both available within the CSD system software described in Chapter 24.3, are commonly used to visualize the k-dimensional data set and to locate natural sub-groupings of data points within it.

Principal component analysis (PCA) (Murray-Rust & Motherwell, 1978; Allen, Doyle & Auf der Heyde, 1991, Allen, Howard & Pitchford, 1996) is a dimension-reduction technique which analyses the variance in $T(N, k)$ in terms of a new set of uncorrelated, orthogonal variables: the principal components, or PCs. The PCs are generated in decreasing order of the percentage of the variance that is explained by each of them. The hope is that the number of PCs, p, that explains most of the variance in the data set is such that $p \ll k$, so that a few pairwise scatter plots with respect to the new PC axes will provide useful visualizations of the complete data set. For cyclic fragments, PCA results are closely related to those obtained using the ring-puckering methodology of Cremer & Pople (1975). Cluster analysis (CA) (Everitt, 1980; Allen, Doyle & Taylor, 1991) is a purely numerical method that attempts to locate discrete groupings of data points within a multivariate data set. CA uses 'distances' or 'dissimilarities' between pairs of points in a k-dimensional space as its working basis, and a very large number of clustering algorithms exist. The mathematical basis of both of these techniques, the modifications that are needed to account for topological symmetry in the search fragment and examples of their application have been reviewed by Taylor & Allen (1994).

Preliminary work using the concepts of machine learning (Carbonell, 1989) for knowledge discovery and classification have also been carried out using the CSD (see *e.g.* Allen *et al.*, 1990; Fortier *et al.*, 1993). In particular, conceptual clustering methods have been applied to a number of substructures (Conklin *et al.*, 1996) and the results compared with those obtained by the statistical and numerical methods described above. Similar techniques are also being used for the classification of protein structures (see *e.g.* Blundell *et al.*, 1987).

22.4.4.3. *Crystallographic conformations and energies*

Crystallographic conformations obviously represent energetically accessible forms. However, for use in molecular-modelling applications, the key question must be asked: Are the condensed-phase crystallographic observations a good guide to conformational preferences in other phases? The indications are that the answer is 'yes' from the types of studies exemplified or cited in the previous section: there appears to be a clear *qualitative* relationship between crystallographic conformer distributions and the low-energy features of the appropriate potential energy hypersurface, although the estimation of absolute energies from the relative populations of these distributions is not appropriate (Bürgi & Dunitz, 1988).

Allen, Harris & Taylor (1996) addressed this question in a systematic manner for a series of 12 one-dimensional (univariate) conformational problems. All of the chosen substructures [simple derivatives of ethane, involving a single torsion angle (τ) about the central C—C bond] were expected to show one symmetric (*anti*, $\tau \simeq 180°$) energy minimum and two symmetry-related asymmetric (*gauche*, $\tau \simeq \pm 60°$) minima. For each substructure, the crystallographic torsional distribution was determined from the CSD and compared with the 1D potential-energy profile, computed using *ab initio* molecular-orbital methods and the 6-31G* basis set. Close agreement was observed between the experimental condensed phase results and the computed *in vacuo* data. Taken over all 12 substructures, the *ab initio* optimized values of the asymmetric (*gauche*) torsion angle vary from $< 55°$ to $> 80°$, and a scatter plot of these optimized values *versus* the mean crystallographic values for *gauche* conformers is linear, with a correlation coefficient of 0.831. Two other results of the study were that: (*a*) torsion angles with higher strain energies (> 4.5 kJ mol^{-1}) are rarely observed in crystal structures ($< 5\%$); and (*b*) taken over many structures, conformational distortions due to crystal packing appear to be the exception rather than the rule.

22.4.4.4. *Conformational libraries*

In essence, the CSD can be regarded 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 (Bower *et al.*, 1997; Dunbrack & Karplus, 1993). The information can either be stored in external databases, or hardwired into the program in the form of rules. However, CSD usage has tended to concentrate on analyses of individual substructures, as noted above, both for their intrinsic interest and to develop novel methods of data analysis. Recently, Klebe & Mietzner (1994) have 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. Conformer prediction is accomplished by the computer program *MIMUMBA*. As part of its programme for the development of knowledge-based libraries from the CSD, the CCDC has now embarked on the generation of a more comprehensive torsional library. Here, information is being hierarchically ordered according to the level of specificity of the chemical substructures for which torsional distributions are available in the library.

22.4.4.5. *Metal coordination geometry*

Some 54% of the information content of the CSD relates to organometallics and metal complexes. This reflects the crucial role of single-crystal diffraction analyses in the renaissance of inorganic chemistry since the 1950s, and the fundamental importance of the technique in characterizing the many novel molecules synthesized over the past 40 years. Since ligands containing nitrogen, oxygen and sulfur are ubiquitous, the CSD contains much information that is relevant to the binding of metal ions by proteins [*e.g.* zinc (Miller *et al.*, 1985), calcium (Strynadka & James, 1989) *etc*.]. Some statistics for the occurrence of some common metals having N and/or O ligands are presented in Table 22.4.3.2.

One of the earliest studies (Einspahr & Bugg, 1981) concerned the geometry of Ca–carboxylate binding, with special reference to biological systems. Since that time, a variety of other studies of biologically relevant metal coordination modes have appeared from the laboratories of Glusker, Dunitz and others (see *e.g.* Glusker, 1980; Chakrabarti & Dunitz, 1982; Carrell *et al.*, 1988, 1993; Chakrabarti, 1990*a*,*b*). These studies show, *inter alia*, that α-hydroxycarboxylates and imidazoles such as histidine tend to bind metal ions in their planes, but that alkali metal cations tend to bind carboxylate groups indiscriminately both in-plane and out-of-plane. Chapter 17 of Glusker *et al.* (1994) is a significant source of additional information and leading references to work in this area over the past two decades.

22.4.5. Intermolecular data

Non-bonded interaction geometries observed in small-molecule crystal structures are of great value in the determination and validation of protein structures, in furthering our understanding of protein folding, and in investigating the recognition processes involved in protein–ligand interactions. The CSD continues to provide vital information on all of these topics.

22.4.5.1. *van der Waals radii*

The hard-sphere atomic model is central to chemistry and molecular biology and, to an approximation, atomic van der Waals radii can be regarded as transferable from one structure to another. They are heavily used in assessing the general correctness of all crystal-structure models from metals and alloys to proteins. Pauling (1939) was the first to provide a usable tabulation for a wide range of elements, but the values of Bondi (1964) remain the most highly cited compilation in the modern literature. His values, assembled from a variety of sources including crystal-structure information, were selected for the calculation of molecular volumes and, in his original paper, Bondi (1964) issues a caution about their general validity for the calculation of limiting contact distances in crystals. In view of the huge amount of non-bonded contact information available in the CSD, Rowland & Taylor (1996) recently tested Bondi's statement as it might apply to the common nonmetallic elements, *i.e.* H, C, N, O, F, P, S, Cl, Br and I. They found remarkable agreement (within 0.02 Å) between the crystal-structure data and the Bondi values for S and the halogens, and agreement within 0.05 Å for C, N and O (new values all larger). The only significant discrepancy was for H, where averaged neutron-normalized small-molecule data yield a van der Waals radius of 1.1 Å, 0.1 Å shorter than the Bondi (1964) value. In the specific area of amino-acid structure, Gould *et al.* (1985) have studied the crystal environments and geometries of leucines, isoleucines, valines and phenylalanines. Their work provides estimates of minimum non-bonded contact distances and indicates the preferred van der Waals interactions of these primary building blocks.

22.4.5.2. *Hydrogen-bond geometry and directionality*

The hydrogen bond is the strongest and most frequently studied of the non-covalent interactions that are observed in crystal structures. As with intramolecular geometries, the first surveys of non-bonded interaction geometries all concerned hydrogen bonds, and were reported long before the CSD existed (Pauling, 1939; Donohue, 1952; Robertson, 1953; Pimentel & McClellan, 1960). The review by Donohue (1952) already contained a plot of N···O distances *versus* C—N···O angles in crystal structures (the C—N groups are terminal charged amino groups), while the review by Pimentel & McClellan (1960) contained histograms of hydrogen-bond distances. Up to the mid-1970s, numerous other studies appeared, *e.g.* Balasubramanian *et al.* (1970), Kroon & Kanters (1974) and Kroon *et al.* (1975), in which all of the statistical analyses were performed manually.

With the advent of the CSD and its developing software system, these kinds of studies became much more accessible and easier to perform, although the non-bonded search facility was only generalized and fully integrated within *Quest3D* in 1992. Thus, Taylor and colleagues reported studies on N—H···O=C hydrogen bonds (Taylor & Kennard, 1983; Taylor *et al.*, 1983, 1984*a*,*b*), Jeffrey and colleagues reported detailed studies on the O—H···O hydrogen bond (Ceccarelli *et al.*, 1981), hydrogen bonds in amino acids (Jeffrey & Maluszynska, 1982; Jeffrey & Mitra, 1984), and hydrogen bonding in nucleosides and nucleotides, barbiturates, purines and pyrimidines (Jeffrey & Maluszynska, 1986), while Murray-Rust & Glusker (1984) studied the directionalities of O—H···O hydrogen bonds to ethers and carbonyls. These studies indicated that hydrogen bonds are often very directional. For example, the distribution of the O—H···O hydrogen-bond angle, after correction for a geometrical factor, peaks at 180° (*i.e.* there is a clear preference for linear hydrogen bonds) and, in carbonyls and carboxylate groups, hydrogen bonds tend to form along the lone-pair directions of the O-atom acceptors (Fig. 22.4.5.1). For ethers, however, lone-pair directionality is not observed, as is illustrated in Fig. 22.4.5.2.

Software availability has facilitated CSD studies of a wide range of individual hydrogen-bonded systems in the recent literature, including studies of resonance-assisted hydrogen bonds (Bertolasi *et al.*, 1996) and resonance-induced hydrogen bonding to sulfur (Allen, Bird *et al.*, 1997*a*). These statistical studies are often combined with molecular-orbital calculations of interaction energies. Some of these studies are cited in this chapter, but the monograph of Jeffrey & Saenger (1991) and the CCDC's DBUSE database are valuable reference sources.