International Tables for Crystallography (2006). Vol. F, Section 22.4.5, pp. 562–567.

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 *Quest*3*D* 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.

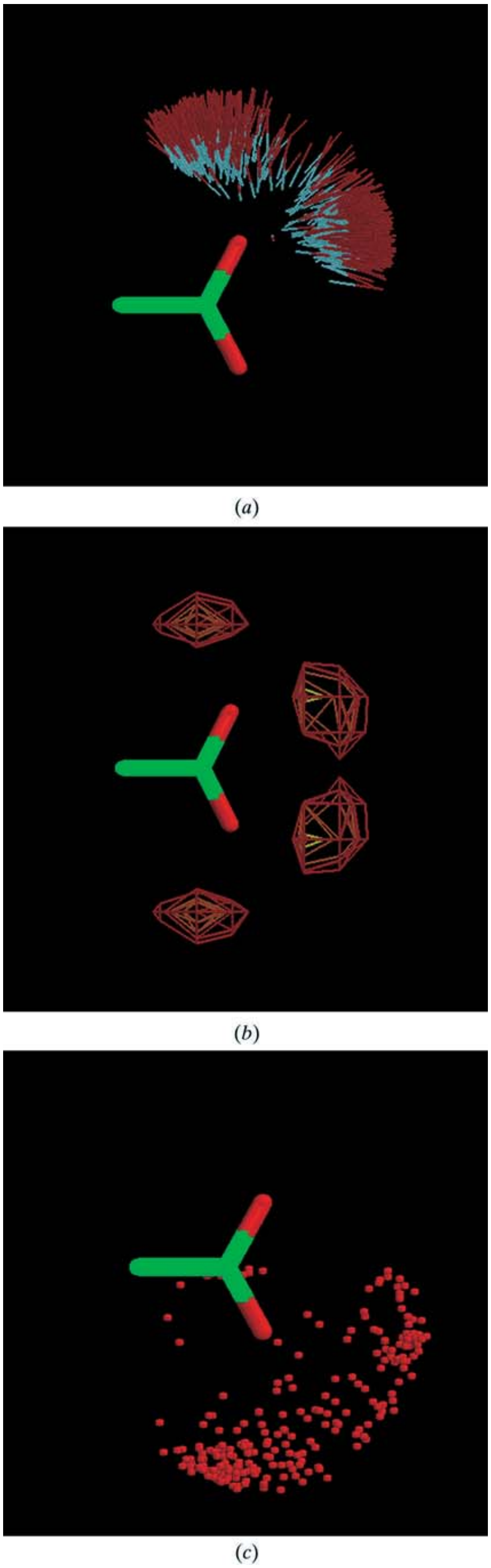

Fig. 22.4.5.1. The IsoStar knowledge-based library of intermolecular interactions: interaction of O—H donors (contact groups) with one of the >C=O acceptors of a carboxylate group (the central group). (*a*) Direct scatter plot derived from CSD data, (*b*) contoured scatter plot derived from CSD data and (*c*) direct scatter plot derived from PDB data.

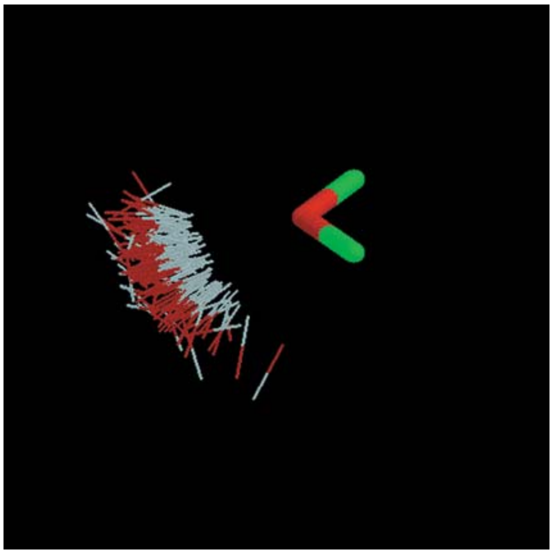

Fig. 22.4.5.2. Distribution of O—H donors around ether oxygen acceptors (CSD data from the IsoStar library, see text).

22.4.5.3. *C—H···X hydrogen bonds*

An important and often underestimated interaction in biological systems is the C—H···*X* hydrogen bond. These bonds have been extensively studied in small-molecule crystal structures, especially in relation to the ongoing discussion as to whether or not they should be called hydrogen bonds. Although Donohue (1968) concluded that the question 'The C—H···O hydrogen bond: what is it?' had only one answer: 'It isn't', a survey of 113 neutron-diffraction structures showed clear statistical evidence for an attractive interaction between C—H groups and oxygen and nitrogen acceptors (Taylor & Kennard, 1982). Later, more evidence for this hypothesis was found, and it was even shown that some C—H···O interactions are directional (Berkovitch-Yellin & Leiserowitz, 1984; Desiraju, 1991; Steiner & Saenger, 1992; Desiraju *et al.*, 1993; Steiner *et al.*, 1996). A continuing area of interest has been to establish the relative donor abilities of C—H in different chemical environments, since spectroscopic data had indicated that donor ability decreased in the order C(*sp*)—H > C(sp^2)—H > C(sp^3)—H. This general hydrogen-acidity requirement was noted by Taylor & Kennard (1982), and systematically addressed using CSD information by Desiraju & Murty (1987), and by Pedireddi & Desiraju (1992), who derived a novel scale of carbon acidity based on C···O separations in a wide variety of systems containing C—H···O hydrogen bonds. A recent paper (Derewenda *et al.*, 1995) highlights the importance of C—H···O=C bonds in stabilizing protein secondary structure.

22.4.5.4. *O—H···π and N—H···π hydrogen bonds*

Spectroscopic evidence for the existence of N,O—H···π hydrogen bonding to acetylenic, olefinic and aromatic acceptors is well documented (Joris *et al.*, 1968). To our knowledge, the first survey of these interactions in the CSD was carried out by Levitt & Perutz (1988), prompted by observations made in protein structures. A more recent CSD survey of this type of bonding (Viswamitra *et al.*, 1993) has shown that intermolecular examples are clearly observed and that these bonds, although very weak, can be both structurally and energetically significant. Recently, Steiner *et al.* (1995) have presented novel crystal structures, database evidence and quantum-chemical calculations on C$\equiv$C—H···π(C$\equiv$C) and π(phenyl) bonding. They cite H···C$\equiv$C (midpoint) distances as short as 2.51 Å and observe hydrogen-bond cooperativity in extended systems with hydrogen-bond energies in the range 4.2–

9.2 kJ mol^{-1}. Finally, we note that electron-rich transition metals can act as proton acceptors in hydrogen-bond interactions with O—H, N—H and C—H donors. Brammer *et al.* (1995) have reviewed progress in this developing area.

22.4.5.5. *Other non-covalent interactions*

The hydrogen bond, $X(\delta-)$—H$(\delta+)\cdots Y(\delta-)$—$Z(\delta+)$, can be viewed as an (almost) linear dipole–dipole interaction, whose ubiquity in nature is due to the presence of many donor–hydrogen dipoles. In a recent review of supramolecular synthons and their application in crystal engineering, Desiraju (1995) illustrates the structural importance of a wide range of attractive non-bonded interactions that do not involve hydrogen mediacy, and notes the long-term value of the CSD in identifying and characterizing these interactions. The area of weak intermolecular interactions is now a burgeoning one in which the combination of CSD analysis and high-level *ab initio* molecular-orbital calculations is proving important in establishing both preferred geometries and estimates of interaction energies. In this context, the intermolecular perturbation theory (IMPT) of Hayes & Stone (1984), a methodology which is free of basis-set superposition errors, is proving particularly useful.

Some of the earliest CSD studies concerned the geometry and directionality of approach of N and O nucleophiles to carbonyl centres, leading to the mapping of (dynamic) reaction pathways through systematic analysis of many examples of related (static) crystal structures (see Bürgi & Dunitz, 1983, 1994). This work was also extended to a study of the directional preferences of non-bonded atomic contacts at sulfur atoms, initially using S in amino acids but later including other examples of divalent sulfur (Rosenfield *et al.*, 1977). It was shown that C—S—C groups tend to bind positively charged electrophiles in directions that are approximately perpendicular to the C—S—C plane, while negatively charged nucleophiles prefer to bind to S along an extension of one of the C—S bonds.

The strong tendency for halogens X = Cl, Br and I to form short contacts to other halogens, and especially to electronegative O and N atoms (Nyburg & Faerman, 1985) is well known (Price *et al.*, 1994). Recent combined CSD/IMPT studies of C—$X\cdots$O=C (Lommerse *et al.*, 1996) and C—$X\cdots$O(nitro) (Allen, Lommerse *et al.*, 1997) systems showed a marked preference for the $X\cdots$O interaction to form along the extension of the C—X bond, with interaction energies in the range −7 to −10 kJ mol^{-1}. These interactions have been used (Desiraju, 1995) to engineer a variety of novel small-molecule crystal structures, and the few $X\cdots$O interactions observed in protein structures generally conform to the geometrical preferences observed in small-molecule studies.

Interactions involving other functional groups are also of importance, and Taylor *et al.* (1990) used CSD information to construct composite crystal-field environments for carbonyl and nitro groups in their search for isosteric replacements in modelling protein–ligand interactions. Their work showed that many of the short intermolecular contacts made by carbonyl groups are to other carbonyl groups in the extended crystal structure. More recently, Maccallum *et al.* (1995*a*,*b*) have demonstrated the importance of Coulombic interactions between the C and O atoms of proximal CONH groups in proteins as an important factor in stabilizing α-helices, β-sheets and the right-hand twist often observed in β-strands. Their calculations indicate an attractive carbonyl–carbonyl interaction energy of about −8 kJ mol^{-1} in specific cases, and they remark that these interactions are *ca* 80% as strong as the CO$\cdots$HN hydrogen bonds within their computational model. Allen, Baalham *et al.* (1998) have used combined CSD/IMPT analysis in a more detailed study of carbonyl–carbonyl interactions and have shown that (*a*) the interaction is commonly observed in small-molecule structures; (*b*) that the preferred interaction geometry is a dimer motif involving two antiparallel C$\cdots$O interactions, although numerous examples of a perpendicular motif (one C$\cdots$O interaction) were also observed; and that (*c*) the total interaction energies for the antiparallel and perpendicular motifs are about −20 and −8 kJ mol^{-1}, respectively, the latter value being comparable to that computed by Maccallum *et al.* (1995*a*,*b*). In studies with protein structures, it has also been noted that carbonyl–carbonyl interactions stabilize the partially allowed Ramachandran conformations of asparagine and aspartic acid (Deane *et al.*, 1999)

22.4.5.6. *Intermolecular motif formation in small-molecule crystal structures*

Desiraju (1995) has stressed that the design process in crystal engineering depends crucially on the high probabilities of formation of certain well known intermolecular motifs, *e.g.* the hydrogen-bonded dimer frequently formed by pairs of carboxylate groups. By analogy with molecular synthesis, he describes these general non-covalent motifs (which often contain strong hydrogen bonds) as supramolecular *synthons*, and points to their importance in supramolecular chemistry as a whole (see *e.g.* Lehn, 1988; Whitesides *et al.*, 1995). Since protein–protein and protein–ligand interactions are also supramolecular phenomena, it follows that information about common interaction motifs is also of importance in structural biology. A computer program is now being written at the CCDC to establish the topologies, chemical constitutions and probabilities of formation of intermolecular motifs directly from the CSD. Initial results (Allen, Raithby *et al.*, 1998; Allen *et al.*, 1999) provide statistics for the most common cyclic hydrogen-bonded motifs, and it is likely that motif information will be included in the developing IsoStar knowledge-based library described in Section 22.4.5.8.

22.4.5.7. *The answer ‘no’*

Previous sections have illustrated the location and characterization of some important non-covalent interactions. Equally important is a knowledge of when such interactions *do not* occur although chemical sensibility might indicate that they should. We provide four examples from the CSD: (*a*) only 4.8% of more than 1000 thioether S atoms form hydrogen-atom contacts that are within van der Waals limits, despite the obvious analogy with the potent

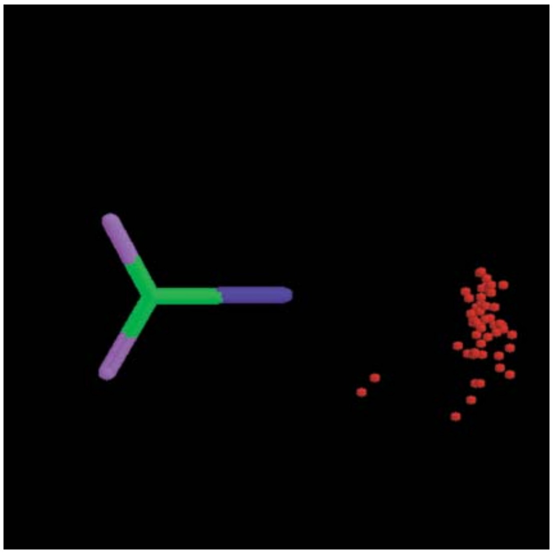

Fig. 22.4.5.3. Distribution of oxygen atoms around C(aromatic)—I (CSD data from the IsoStar library, see text).

acceptor C—O—C (Allen, Bird *et al.*, 1997*b*); (*b*) of 118 instances in which a furan ring coexists with N—H or O—H donors, the O atom forms hydrogen bonds on only three occasions (Nobeli *et al.*, 1997); (*c*) the ester oxygen $(R_1)(\mathrm{O}{=})\mathrm{C}{-}\underline{\mathrm{O}}{-}R_2$ almost never forms strong hydrogen bonds, although the adjunct carbonyl oxygen atom is well known as a highly potent acceptor (Lommerse *et al.*, 1997); and (*d*) covalently bound fluorine atoms rarely form hydrogen bonds (Dunitz & Taylor, 1997).

22.4.5.8. *IsoStar: a library of non-bonded interactions*

The previous sections show that the amount of data in the CSD on intermolecular geometries is vast, and CSD-derived information for a number of specific systems is available in the literature at various levels of detail. If not, the CSD must be searched for contacts between the relevant functional groups. To provide structured and direct access to a more comprehensive set of derived information, a knowledge-based library of non-bonded interactions (IsoStar: Bruno *et al.*, 1997) has been developed at the CCDC since 1995. IsoStar is based on experimental data, not only from the CSD but also from the PDB, and contains some theoretical results calculated using the IMPT method. Version 1.1 of IsoStar, released in October 1998, contains information on non-bonded interactions formed between 310 common functional groups, referred to as *central groups*, and 45 *contact groups*, *e.g.* hydrogen-bond donors, water, halide ions *etc.* Information is displayed in the form of scatter plots for each interaction. Version 1.1 contains about 12000 scatter plots: 9000 from the CSD and 3000 from the PDB. IsoStar also reports results for 867 theoretical potential-energy minima.

For a given contact between between a central group (*A*) and a contact group (*B*), CSD search results were transformed into an easily visualized form by overlaying the *A* moieties. This results in a 3D distribution (scatter plot) showing the experimental distribution of *B* around *A*. Fig. 22.4.5.1(*a*) shows an example of a scatter plot: the distribution of OH groups around carboxylate anions, illustrating hydrogen-bond formation along the lone-pair directions of the carboxylate oxygens. The IsoStar software provides a tool that enables the user to inspect quickly the original crystal structures in which the contacts occur *via* a hyperlink to the original CSD entries. This is very helpful in identifying outliers, motifs and biases. Another tool generates contoured surfaces from scatter plots, which show the density distribution of the contact groups. A similar approach was first used by Rosenfield *et al.* (1984). Contouring aids the interpretation of the scatter plot and the analysis of preferred geometries. Fig. 22.4.5.1(*b*) shows the contoured surface of the scatter plot in Fig. 22.4.5.1(*a*); the lone-pair directionality now becomes even more obvious.

The fact that carboxylate anions form hydrogen bonds along their lone-pair directions may be well known, although force fields do not always use this information. However, the IsoStar library also contains information on many less well understood functional groups. The interaction between aromatic halo groups and oxygen atoms (Lommerse *et al.*, 1996) is referred to above, and Fig. 22.4.5.3 shows the distribution of oxygen acceptor atoms around aromatic iodine groups. It is clear that the contact O atoms are preferentially observed along the elongation of the C—I bond.

The PDB scatter plots in IsoStar only involve interactions between non-covalently bound ligands and proteins, *i.e* side chain–side chain interactions are excluded. Similar work was presented by Tintelnot & Andrews (1989), but at that time the PDB contained only 40 structures of protein–ligand complexes. The IsoStar library contains data derived from almost 800 complexes having a resolution better than 2.5 Å. Fig. 22.4.5.1(*c*) shows an example of a scatter plot from the PDB (the distribution of OH groups around carboxylate groups). Here, although the hydrogen atoms are missing in the PDB plot, the close similarity between Figs. 22.4.5.1(*c*) (PDB) and 22.4.5.1(*a*) (CSD) is obvious.

22.4.5.9. *Protein–ligand binding*

The reluctance to use data from the CSD because they do not relate directly to biological systems has been noted earlier. However, in principle, the same forces that drive the inclusion of a new molecule into a growing crystal should also apply to the binding of a ligand to a protein. In both cases, molecule and target need to be de-solvated first (although in the first case not necessarily from a water environment) and then interact in the most favourable way.

Nicklaus and colleagues suggested that on average, the conformational energy of ligands in the protein-bound state is 66 (48) kJ mol^{-1} above that of the global minimum-energy conformation *in vacuo* (Nicklaus *et al.*, 1995). This result was based on 33 protein–ligand complexes from the PDB for which the ligand also occurs in a small-molecule structure in the CSD. The same investigation also showed that, although ligand conformations in the protein-bound state are generally different from those observed in small-molecule crystal structures, on average the conformational energy of the ligand in the CSD crystal-structure conformation is 66 (47) kJ mol^{-1} above that of the global minimum-energy conformation *in vacuo*, although Boström *et al.* (1998) have shown that these conformational energies are much lower if calculated in a water environment. The computational work indicates that the forces that affect the conformation of a ligand are of comparable magnitude at a protein binding site to those in a small-molecule crystal-structure environment. Thus, if small-molecule crystal-structure statistics tell us that a given structure fragment can only adopt one conformation, generally there is no reason to believe that a ligand that contains this fragment will adopt a different conformation when it binds to a protein.

In principle, the information on non-bonded interactions derived from the CSD and assembled in the IsoStar library should be very important for the understanding and prediction of interaction geometries. However, in light of the comments above, it is important to know whether these data *are* generally relevant to interactions that occur in the protein binding site. Work by Klebe (1994) indicated that, at least for a limited set of test cases, the geometrical distributions derived from ligand–protein complexes are similar to those derived from small-molecule crystal structures. Since the IsoStar library contains information from both the PDB and the CSD, it provides the ultimate basis for establishing similarities (or not) between the interaction geometries observed in small-molecule crystal structures and those observed in protein–ligand complexes. Comparing CSD scatter plots with their corresponding plots from the PDB is an obvious way of establishing the relevance of non-bonded interaction data from small-molecule crystal structures to biological systems.

A full systematic comparison of PDB and CSD scatter plots or, more accurately, of PDB and CSD *density maps* has recently been performed by Verdonk (1998). He calculated residual densities, obtained by subtracting one density map from the other, for each pair of density maps. It appears that, in general, CSD and PDB plots (and thus interaction geometries) are very similar indeed: the average residual density is only 10 (10)%, indicating that 90% of the density in the PDB map is also observed in the CSD map. In Fig. 22.4.5.4(*a*), the average residual densities of each PDB–CSD comparison are plotted *versus* the average concentration of contact groups in the scatter plot. The filled circles represent comparisons for which the protonation state of the central group is unambiguous (*i.e.* carboxylic acid, imidazole *etc.* were excluded). It appears that the residual density decreases with the amount of data in the plots,

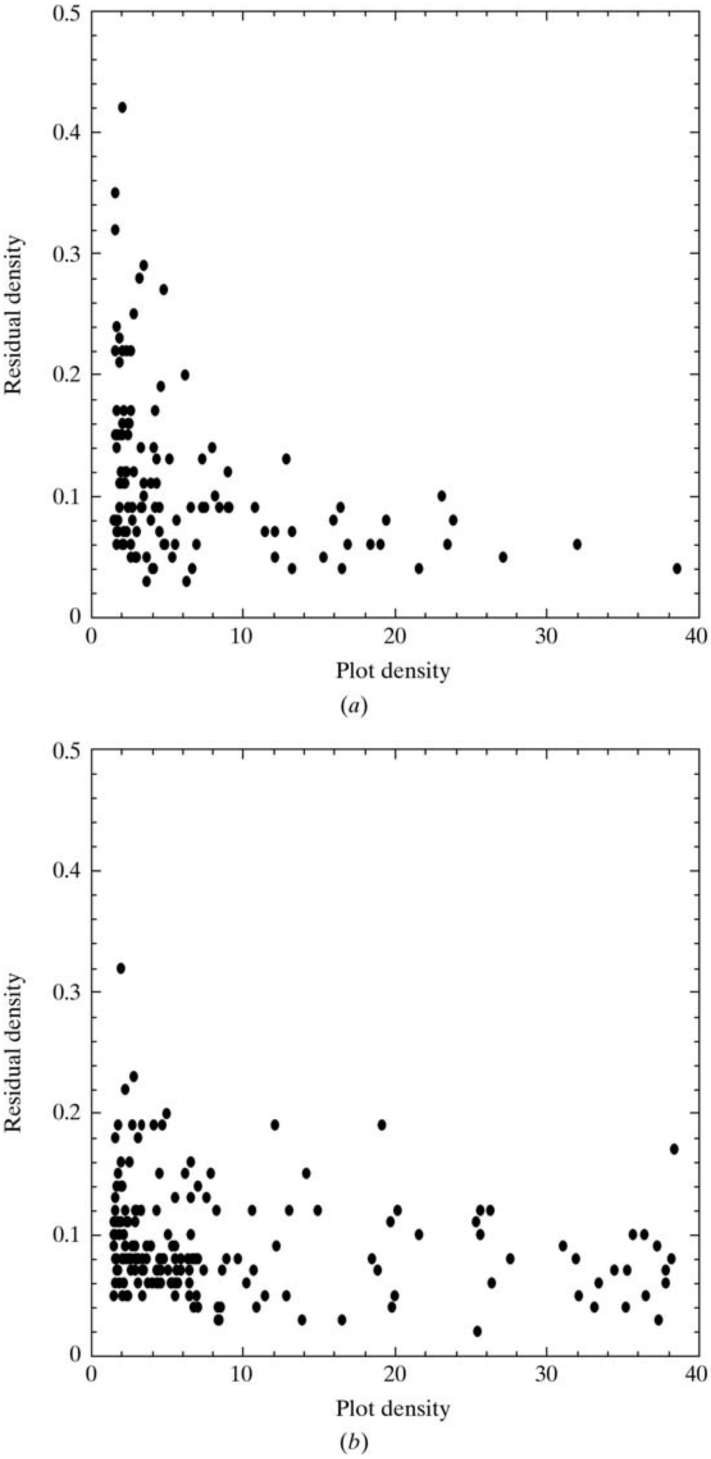

Fig. 22.4.5.4. Pairwise comparison of intermolecular-interaction density maps from the CSD and the PDB. Plots of *residual density* $|\rho(\mathrm{CSD}) - \rho(\mathrm{PDB})|$ *versus plot density*, *i.e.* the average density in the least dense situation (CSD or PDB), for situations where the protonation state of the central group is (*a*) unambiguous, and (*b*) ambiguous.

obviously caused by the more accurate calculation of the residual density. The 'true' residual density seems to be as low as about 6%.

Fig. 22.4.5.4(*b*) shows a similar graph, but now for those density maps in which the protonation state of the central group is ambiguous. As expected, the spread in the calculated residual densities is much higher, even for very dense plots. By comparing the density map from the PDB with the CSD maps for the different protonation states of the central group, the most frequent protonation state of this central group in the protein structures can be predicted. In Table 22.4.5.1, for example, the residual densities for protein carboxylic acid groups are shown, compared with the CSD plots of the neutral carboxylic acid and with those of the charged carboxylate anion. In all cases, the residual density is lower if the PDB map is compared with the CSD map for charged carboxylate anions. This indicates that the majority of glutamate and aspartate side chains are charged, which is consistent with other evidence.

Table 22.4.5.1. *Residual densities for carboxylic acid groups*

The PDB density maps are compared with the CSD maps for uncharged carboxylic acid and for charged carboxylate anions.

	Residual density (CCO_2H)	Residual density ($CCOO^-$)
Any (N,O,S)—H	0.06	0.04
Any N—H nitrogen	0.07	0.05
Any O—H oxygen	0.07	0.05
Non-donating oxygen	0.12	0.04
Carbonyl oxygen	0.13	0.07
Carbonyl carbon	0.12	0.04
Water oxygen	0.07	0.05
Any aliphatic C—H carbon	0.08	0.06

22.4.5.10. *Modelling applications that use CSD data*

Predicting binding modes of ligands at protein binding sites is a problem of paramount importance in drug design. One approach to this problem is to attempt to dock the ligand directly into the binding site. There are several protein–ligand docking programs available, *e.g. DOCK* (see Kuntz *et al.*, 1994), *GRID* (Goodford, 1985), *FLExX* and *FLExS* (Rarey *et al.*, 1996; Lemmen & Lengauer, 1997), and *GOLD* (Jones *et al.*, 1995, 1997). The docking program *GOLD*, developed by the University of Sheffield, Glaxo Wellcome and the CCDC, and which has the high docking success rate of 73%, uses a small torsion library, based on the data from the CSD, to explore the conformational space of the ligand. Its hydrogen-bond geometries and fitness functions are also partly based on CSD data. In the future, we intend to create a more direct link between the crystallographic data and the docking program, *via* IsoStar and the developing torsion library.

Another approach to the prediction of binding modes is to calculate the energy fields for different probes at each position of the binding site, for instance using the *GRID* program (Goodford, 1985). The resulting maps can be displayed as contoured surfaces which can assist in the prediction and understanding of binding modes of ligands. CCDC is developing a program called *SuperStar* (Verdonk *et al.*, 1999) which uses a similar approach to that of the *X-SITE* program (Singh *et al.*, 1991; Laskowski *et al.*, 1996). However, *SuperStar* uses non-bonded interaction data from the CSD rather than the protein side chain–side chain interaction data employed in *X-SITE*. Thus, for a given binding site and contact group (probe), *SuperStar* selects the appropriate scatter plots from the IsoStar library, superimposes the scatter plots on the relevant functional groups in the binding site, and transforms them into one composite probability map. Such maps can then, for example, be used to predict where certain functional groups are likely to interact with the binding site. The strength of *SuperStar* is that it is based entirely on experimental data (although this is also the cause of some limitations). The fields simply represent what has been observed in crystal strucures. We are currently verifying *SuperStar* on a test set of more than 100 protein–ligand complexes from the PDB and preliminary results are encouraging.

Finally, CSD data are used in several *de novo* design programs. These types of programs, *e.g. LUDI* (Böhm, 1992*a*,*b*), predict novel ligands that will interact favourably with a given protein and use hydrogen-bond geometries from the CSD (indirectly) to position their structural fragments in the binding site.

22.4.6. Conclusion

This chapter has summarized the vast range of structural knowledge that can be derived from the small-molecule data contained in the CSD. We have attempted to show that much of this knowledge is directly transferable and applicable to the protein environment. Far from being discrete, structural studies of small molecules and proteins have a natural synergy which, if exploited creatively, will lead to significant advances in both areas. It is therefore unsurprising that some of these CSD studies have been prompted by initial observations made on proteins.

As a result of this activity, it is now very clear that software access to the information stored in the CSD and the PDB must be at two levels: a raw-data level and a derived-knowledge level. The onward development of structural knowledge bases from the underlying data provides for the preservation and storage of the results of data-mining experiments, thus avoiding repetition of standard experiments and providing instant access to complex derivative information. Most importantly, a suitably structured knowledge base can be acted on by software tools that are designed to solve complex problems in structural chemistry (see *e.g.* Thornton & Gardner, 1989; Allen *et al.*, 1990; Bruno *et al.*, 1997; Jones *et al.*, 1997). The availability of knowledge bases derived from experimental observations is likely to be a crucial factor in the solution of those two analogous, and currently intractable, problems in the small-molecule and protein-structure domains: crystal structure and polymorph prediction on the one hand, and protein folding on the other.

References

22.1

Acharya, R., Fry, E., Logan, D., Stuart, D., Brown, F., Fox, G. & Rowlands, D. (1990). *The three-dimensional structure of foot-and-mouth disease virus. New aspects of positive-strand RNA viruses*, edited by M. A. Brinton & S. X. Heinz, pp. 319–327. Washington DC: American Society for Microbiology.

Arnold, E. & Rossmann, M. G. (1990). *Analysis of the structure of a common cold virus, human rhinovirus 14, refined at a resolution of 3.0 Å. J. Mol. Biol.* **211**, 763–801.

Baker, E. N. & Hubbard, R. E. (1984). *Hydrogen bonding in globular proteins. Prog. Biophys. Mol. Biol.* **44**, 97–179.

Bernal, J. D. & Finney, J. L. (1967). *Random close-packed hard-sphere model II. Geometry of random packing of hard spheres. Discuss. Faraday Soc.* **43**, 62–69.

Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1965). *Structure of hen egg-white lysozyme, a three-dimensional Fourier synthesis at 2 Å resolution. Nature (London)*, **206**, 757–761.

Bondi, A. (1964). *van der Waals volumes and radii. J. Phys. Chem.* **68**, 441–451.

Bondi, A. (1968). *Molecular crystals, liquids and glasses*. New York: Wiley.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). *CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem.* **4**, 187–217.

Chandler, D., Weeks, J. D. & Andersen, H. C. (1983). *van der Waals picture of liquids, solids, and phase transformations. Science*, **220**, 787–794.

Chapman, M. S. (1993). *Mapping the surface properties of macromolecules. Protein Sci.* **2**, 459–469.

Chapman, M. S. (1994). *Sequence similarity scores and the inference of structure/function relationships. Comput. Appl. Biosci. (CABIOS)*, **10**, 111–119.

Chothia, C. (1975). *Structural invariants in protein folding. Nature (London)*, **254**, 304–308.

Chothia, C. (1976). *The nature of the accessible and buried surfaces in proteins. J. Mol. Biol.* **105**, 1–12.

Chothia, C. & Janin, J. (1975). *Principles of protein–protein recognition. Nature (London)*, **256**, 705–708.

Connolly, M. (1986). *Measurement of protein surface shape by solid angles. J. Mol. Graphics*, **4**, 3–6.

Connolly, M. L. (1983). *Analytical molecular surface calculation. J. Appl. Cryst.* **16**, 548–558.

Connolly, M. L. (1991). *Molecular interstitial skeleton. Comput. Chem.* **15**, 37–45.

Diamond, R. (1974). *Real-space refinement of the structure of hen egg-white lysozyme. J. Mol. Biol.* **82**, 371–391.

Dunfield, L. G., Burgess, A. W. & Scheraga, H. A. (1979). *J. Phys. Chem.* **82**, 2609.

Edelsbrunner, H., Facello, M. & Liang, J. (1996). *On the definition and construction of pockets in macromolecules*, pp. 272–287. Singapore: World Scientific.

Edelsbrunner, H., Facello, M., Ping, F. & Jie, L. (1995). *Measuring proteins and voids in proteins. Proc. 28th Hawaii Intl Conf. Sys. Sci.* pp. 256–264.

Edelsbrunner, H. & Mucke, E. (1994). *Three-dimensional alpha shapes. ACM Trans. Graphics*, **13**, 43–72.

Eisenberg, D. & McLachlan, A. D. (1986). *Solvation energy in protein folding and binding. Nature (London)*, **319**, 199–203.

Fauchere, J.-L. & Pliska, V. (1983). *Hydrophobic parameters π of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. Eur. J. Med. Chem. Chim. Ther.* **18**, 369–375.

Finkelstein, A. (1994). *Implications of the random characteristics of protein sequences for their three-dimensional structure. Curr. Opin. Struct. Biol.* **4**, 422–428.

Finney, J. L. (1975). *Volume occupation, environment and accessibility in proteins. The problem of the protein surface. J. Mol. Biol.* **96**, 721–732.

Finney, J. L., Gellatly, B. J., Golton, I. C. & Goodfellow, J. (1980). *Solvent effects and polar interactions in the structural stability and dynamics of globular proteins. Biophys. J.* **32**, 17–33.

Fritz-Wolf, K., Schnyder, T., Wallimann, T. & Kabsch, W. (1996). *Structure of mitochondrial creatine kinase. Nature (London)*, **381**, 341–345.

Gelin, B. R. & Karplus, M. (1979). *Side-chain torsional potentials: effect of dipeptide, protein, and solvent environment. Biochemistry*, **18**, 1256–1268.

Gellatly, B. J. & Finney, J. L. (1982). *Calculation of protein volumes: an alternative to the Voronoi procedure. J. Mol. Biol.* **161**, 305–322.

Gerstein, M. (1992). *A resolution-sensitive procedure for comparing surfaces and its application to the comparison of antigen-combining sites. Acta Cryst.* A**48**, 271–276.

Gerstein, M. & Chothia, C. (1996). *Packing at the protein–water interface. Proc. Natl Acad. Sci. USA*, **93**, 10167–10172.

Gerstein, M., Lesk, A. M., Baker, E. N., Anderson, B., Norris, G. & Chothia, C. (1993). *Domain closure in lactoferrin: two hinges produce a see-saw motion between alternative close-packed interfaces. J. Mol. Biol.* **234**, 357–372.

Gerstein, M., Lesk, A. M. & Chothia, C. (1994). *Structural mechanisms for domain movements. Biochemistry*, **33**, 6739–6749.

Gerstein, M. & Lynden-Bell, R. M. (1993*a*). *Simulation of water around a model protein helix. 1. Two-dimensional projections of solvent structure. J. Phys. Chem.* **97**, 2982–2991.