International Tables for Crystallography (2006). Vol. F, Section 22.4.5.9, pp. 565–566.

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,

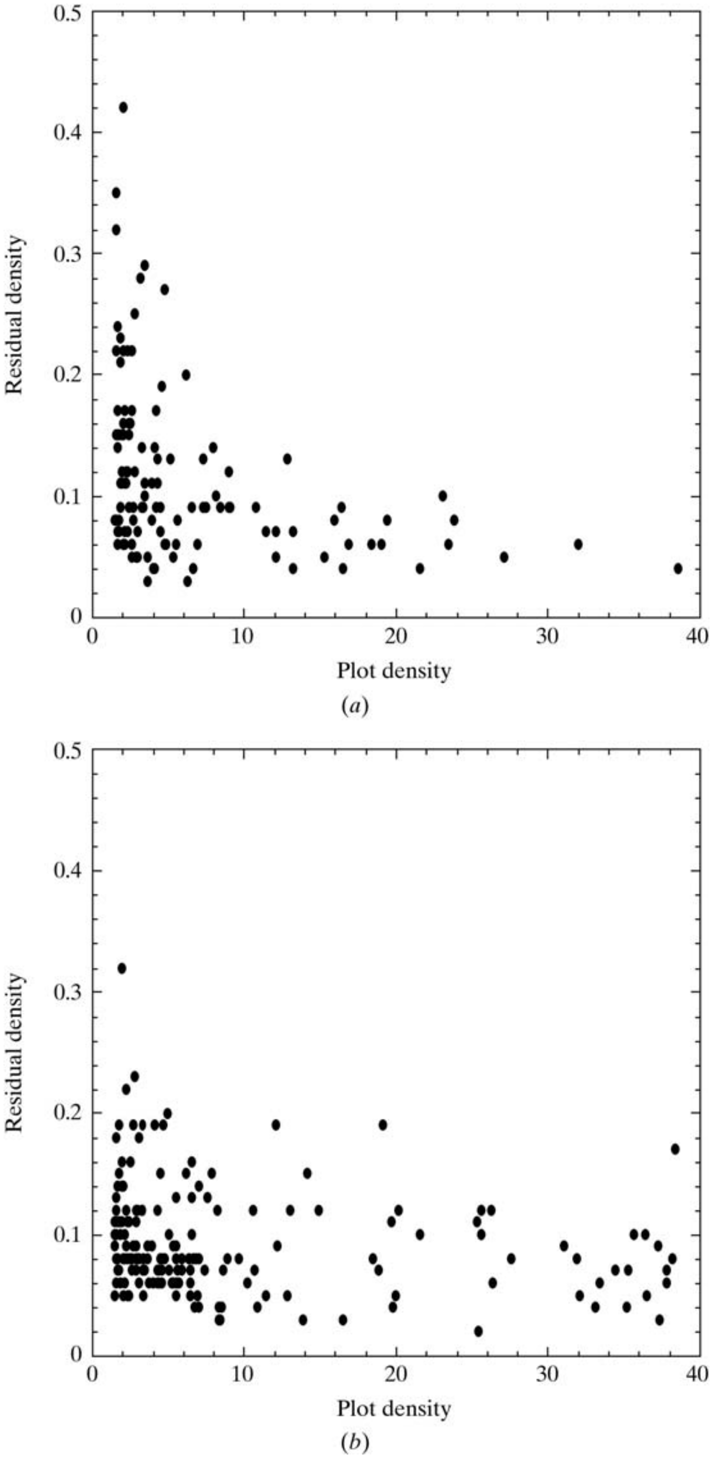

Fig. 22.4.5.4. Pairwise comparison of intermolecular-interaction density maps from the CSD and the PDB. Plots of *residual density* $|\rho(\mathrm{CSD}) - \rho(\mathrm{PDB})|$ *versus plot density*, *i.e.* the average density in the least dense situation (CSD or PDB), for situations where the protonation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