Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Life-science applications of the Cambridge Structural Database

Robin Taylor

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England

Correspondence e-mail: taylor@ccdc.cam.ac.uk

Received 29 January 2002
Accepted 25 February 2002

Several studies show that the molecular geometries and intermolecular interactions observed in small-molecule crystal structures are relevant to the modelling of *in vivo* situations, although the influence of crystal packing is sometimes important and should always be borne in mind. Torsional distributions derived from the Cambridge Structural Database (CSD) can be used to map out potential-energy surfaces and thereby help identify experimentally validated conformational minima of molecules with several rotatable bonds. The use of crystallographic data in this way is complementary to *in vacuo* theoretical calculations since it gives insights into conformational preferences in condensed-phase situations. Crystallographic data also underpin many molecular-fragment libraries and programs for generating three-dimensional models from two-dimensional chemical structures. The modelling of ligand binding to metalloenzymes is assisted by information in the CSD on preferred coordination numbers and geometries. CSD data on intermolecular interactions are useful in structure-based inhibitor design both in indicating how probable a protein–ligand interaction is and what its geometry is likely to be. They can also be used to guide searches for bioisosteric replacements. Crystallographically derived information has contributed to many life-science software applications, including programs for locating binding 'hot spots' on proteins, docking ligands into enzyme active sites, *de novo* ligand design, molecular superposition and three-dimensional QSAR. Overall, crystallographic data in general, and the CSD in particular, are very significant tools for the rational design of biologically active molecules.

1. Introduction

Structure-based drug design relies heavily on experimental protein X-ray structures, either proprietary or taken from the Protein Data Bank (PDB; Bernstein *et al.*, 1977; Berman *et al.*, 2000, 2002). Of equal importance is our ability to predict the conformational preferences and non-covalent interactions of putative protein ligands from small-molecule crystal-structure data, such as those compiled in the Cambridge Structural Database (CSD; Allen, 2002). This article reviews the applications of crystal structure data in general, and the CSD in particular, to life-science research.

There are many straightforward applications of the CSD in drug design. Some examples are as follows.

(i) Retrieval of crystallographic data on individual molecules (such as natural products) is useful for checking stereochemistries and finding starting geometries for molecular modelling.

(ii) The CSD is widely used for validating computational methodology, *e.g.* to check the accuracy of computer-generated three-dimensional structures (Sadowski *et al.*, 1994), for testing shape-complementarity algorithms (Leherte *et al.*, 1996) and for validating protein–ligand docking programs such as *DOCK* (Grootenhuis *et al.*, 1994).

(iii) The CSD has been used directly as a searchable three-dimensional database for discovering lead molecules for synthesis and biological testing; some biologically active compounds have been discovered in this way (*e.g.* DesJarlais *et al.*, 1990; Lam *et al.*, 1994; Chowdhury *et al.*, 2001). While the value of the CSD for this purpose is limited because no physical specimens of 'hit' compounds are available for assaying (therefore, they have to be synthesized rather than taken from a company's own compound collection), its exceptional chemical diversity is attractive when there is an emphasis on finding very novel areas of chemistry.

In contrast to these straightforward uses of the CSD, the remainder of the paper focuses on applications which rely on more sophisticated data-mining and knowledge-engineering techniques.

2. CSDS programs

Programs in the Cambridge Structural Database System (CSDS) of particular relevance to the life sciences are: (i) the main search program, *ConQuest*, together with *Mercury*, a program for exploring and visualizing intermolecular interactions in crystal structures (Bruno *et al.*, 2002), (ii) *Vista*, a program for performing statistical analyses of CSD search results (Cambridge Crystallographic Data Centre, 1995), (iii) IsoStar, a database of information about non-bonded interactions that have been culled from the CSD, the PDB and theoretical calculations (Bruno *et al.*, 1997) and (iv) *SuperStar*, a program that uses IsoStar information to predict 'hot spots' where ligand atoms might bind in enzyme active sites (Verdonk *et al.*, 1999). Many of the results discussed below were generated with these programs or with *QUEST* (Allen *et al.*, 1991), the previous search interface to the CSD, now superseded by *ConQuest*.

3. Relevance of small-molecule crystal structures to *in vivo* situations

The relevance of small-molecule crystal structures to *in vivo* situations is a question of obvious importance. Three aspects of this question are addressed here. First, are the molecular conformations observed in crystal structures a good guide to conformational preferences in aqueous solution or at protein-binding sites? Secondly, are metal coordination geometries in the CSD relevant to those occurring in metalloenzymes? Thirdly, are intermolecular interactions in crystal lattices similar to those that occur between proteins and their bound ligands?

3.1. Relevance of CSD data on molecular conformations

The conformation of a molecule in a crystal structure will be affected by its crystal-field environment and cannot be assumed to represent either the global minimum-energy geometry in aqueous solution or the geometry adopted when the molecule binds to a protein. However, if a particular molecular fragment containing a rotatable bond is observed in a series of crystal structures, it is likely that more strained (higher energy) conformations will be observed less often than relatively unstrained (lower energy) geometries. Thus, the observed distribution of torsion angles around the rotatable bond should reflect the potential energy curve for rotation around that bond. Allen *et al.* (1996) studied 12 molecular fragments and compared their torsion-angle distributions obtained from the CSD with potential-energy curves obtained from *ab initio* (6-31G*//STO-3G and 6-31G*//3-21G) calculations on appropriate model compounds. The qualitative complementarity of the torsion-angle histograms and the calculated energy curves was striking. Each fragment was able to adopt two conformers, *anti* and *gauche*, and the relative frequencies with which these two conformers were observed in crystal structures,

$$C = N_{anti}/N_{gauche}$$

could be related to the *ab initio* calculated energy differences

$$\Delta E = E_{gauche} - E_{anti}$$

by the empirical equation

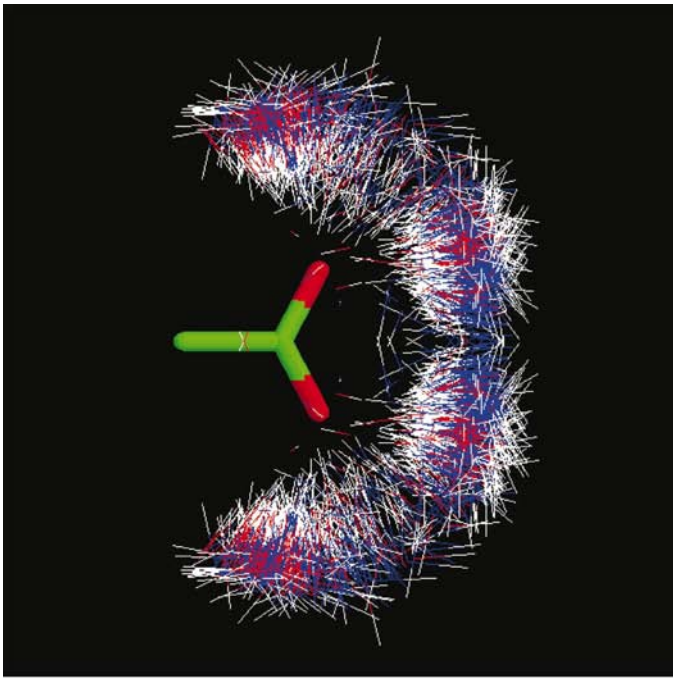
$$\Delta E = 0.20 + 0.23 \ln C$$

with a Pearson correlation coefficient of 0.74 and a statistical significance of 0.006. This showed that, in spite of their being no underlying theoretical justification (Bürgi & Dunitz, 1988), the observed distribution of crystal-structure rotamers approximately follows a Boltzmann distribution. However, analysis of the intercept and coefficient in the above equation suggested that high-energy conformers are somewhat under-represented in crystal structures compared with a room-temperature Boltzmann distribution.

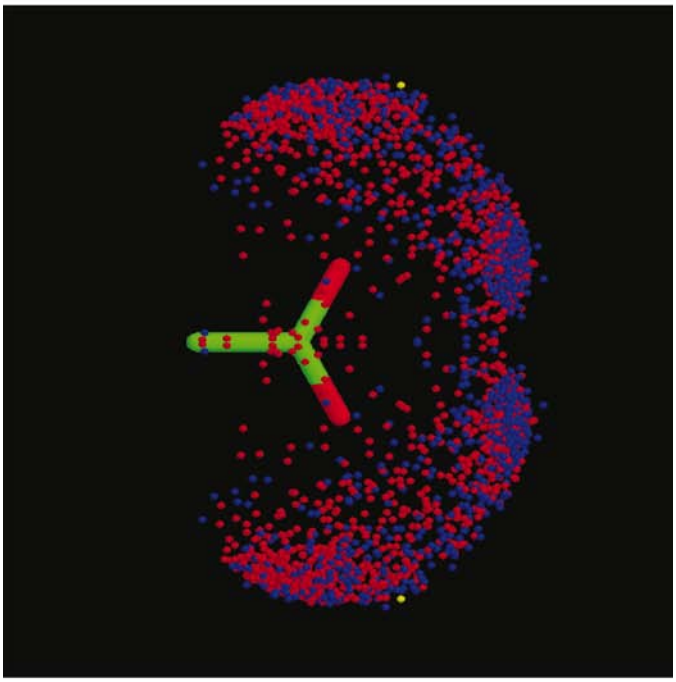
Occasionally, systematic crystal-packing effects cause substantial deviations from Boltzmann-like statistics. For example, the electron-diffraction value for the inter-ring torsion angle of biphenyl is 44° but the molecule is planar in its crystal structure, presumably because this optimizes close packing (Brock & Minton, 1989). In the CSD, the distribution of inter-ring torsion angles in biphenyls with no *ortho*-substituents shows two peaks, one at 0° and one close to the gas-phase value of 44°.

In one respect, crystal-structure conformations are more representative of *in vivo* situations than are theoretical energy calculations. The latter are almost invariably *in vacuo* calculations, which severely limits their value for predicting conformational preferences in aqueous solution. This problem is particularly serious for molecules that are capable of forming intramolecular hydrogen bonds. For such molecules, conformational searches based on *in vacuo* calculations almost always predict the lowest energy conformer to be the one

containing intramolecular hydrogen bond(s). In aqueous solution, this tendency is usually reversed because of the possibility of competing intermolecular hydrogen bonding with water. Crystal-structure geometries are a better guide to aqueous solution conformations than are *in vacuo* calculations, since intermolecular hydrogen bonding can occur in crystal lattices. Bilton *et al.* (2000) have identified common substructural motifs which tend to form intramolecular hydrogen bonds in crystal structures and these should help identify molecules which are likely to form intramolecular bonds in aqeuous solution.



(*a*)



(*b*)

Figure 1
Distribution of hydrogen-bond donor groups around carboxylates in (*a*) the CSD and (*b*) the PDB. Contacts less than (sum of van der Waals radii) − 0.3 Å are shown and data is taken from IsoStar Version 1.4.

The above considerations suggest that geometry distributions in small-molecule crystal structures generally serve as useful guides to geometry distributions in solution. However, the question arises whether there may be some systematic difference between conformational preferences in crystal structures and the conformations adopted by ligands bound to proteins. This may indeed be the case for enzyme substrates, since strain energy in a bound substrate effectively lowers the activation energy of the reaction catalysed by the enzyme and may therefore be favoured by evolution. However, no such argument applies to synthetic ligands, since they have not been selected by evolutionary pressure. Boehm & Klebe (1996) compared torsion distributions of molecular fragments in the CSD with those of the corresponding fragments in protein-bound ligands and found them to be similar. Studies like this are complicated by the fact that ligand electron density is often difficult to fit unambiguously in protein crystallographic studies (although there are, of course, an increasing number of high-resolution protein structures). For this reason, unusual ligand conformations in the PDB should be treated with caution (Boehm & Klebe, 1996); the higher precision data from the CSD have a valuable validation role here.

3.2. Relevance of CSD data on metal coordination

The extent to which metal-ion coordination in small-molecule structures is relevant to coordination in metalloenzymes is a subject worthy of more investigation, especially given the possibility that unusual metal coordination geometries might have evolved for catalytic reasons. In many cases, it does appear that metal coordination in enzymes is the same as that seen in small-molecule structures. Interestingly, however, differences between small molecules and macromolecules have been observed in cation binding to phenolates (Chakrabarti & Hsu, 1994). In both, cations tend to avoid the oxygen sp^2 lone-pair directions. However, cations in small-molecule phenolate complexes tend to be positioned close to the aromatic ring plane, between the sp^2 and C—O vectors, whereas in proteins they tend to lie outside this plane. Experimental data on a wider variety of metalloenzymes are needed before reliable conclusions can be drawn on whether differences like this are rare exceptions or relatively common.

3.3. Relevance of CSD data on intermolecular interactions

As with molecular conformations, there is evidence that the geometrical distribution of non-covalent interactions in small-molecule crystal structures is usually Boltzmann-like, implying that such information can be used to make deductions about geometrical preferences in other phases. Comparison of observed hydrogen-bond geometry distributions with simulated (gas-phase) Boltzmann distributions using an empirical hydrogen-bond potential was performed by Taylor (1981). The empirical potential contained van der Waals and electrostatic

components, together with a Morse term whose (favourable) contribution to the calculated hydrogen-bond energy was attenuated as the O—H···O angle became less linear. The simulated distribution of hydrogen-bond distances, H···O, was found to extend to much longer distances than the observed distribution, presumably because the tendency of molecules to close-pack in crystal structures militates against the occurrence of long hydrogen bonds. All observed distributions of angular parameters (*e.g.* OH···O hydrogen-bond angle and parameters relating to lone-pair directionality) were, however, found to be well reproduced by the Boltzmann simulations.

The geometries of non-bonded contacts in the CSD were comprehensively compared with those between proteins and bound ligands in the PDB by Boer *et al.* (2001). These authors selected >50 different contacts from the IsoStar database (Bruno *et al.*, 1997). For each contact, scatter plots showing the distribution of one group around another were available, based on data from (i) the CSD and (ii) the PDB (for example, Fig. 1 shows the distributions of hydrogen-bond donors around carboxylate groups in the CSD and the PDB). The similarity of each such CSD–PDB pair was computed using the Carbo index (Carbo *et al.*, 1980), which measures the geometrical (*i.e.* shape) similarity of the plots. The average Carbo similarity index was 0.76, indicating high similarity, with very few pairs of scatter plots having Carbo coefficients <0.6. These few exceptions almost always involved scatter plots for which only a limited amount of data was available. Low similarities can also occur when the group in the PDB plot has more than one possible ionization or tautomeric state. This is a troublesome problem when studying hydrogen bonds using data from protein crystallography, since experimental resolution is generally too low to permit location of H atoms and algorithms for deducing H-atom positions are imperfect. In contrast, many (probably most) small-molecule structures in the CSD have experimentally determined H-atom positions.

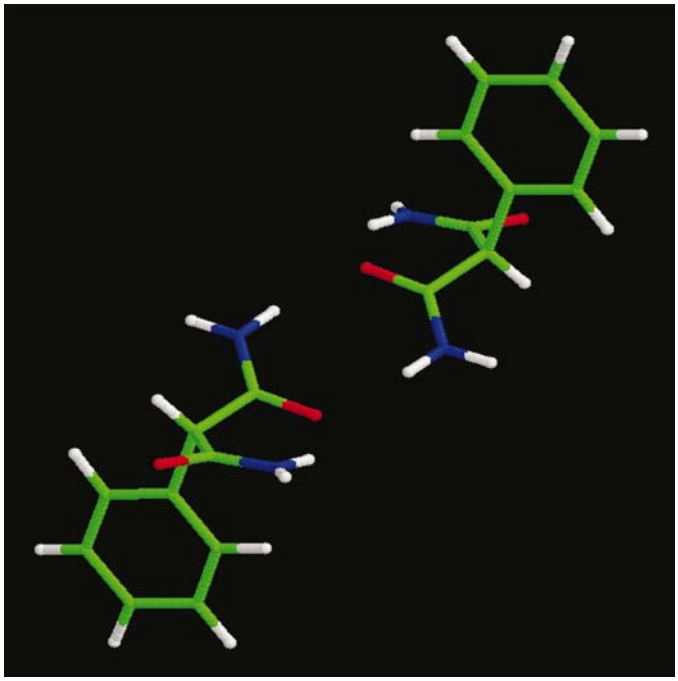


Figure 2
Common dimer motif in small-molecule crystal structures (CSD entry FULSUA; Sakamoto *et al.*, 2000).

As with molecular conformations, there are occasional situations in which systematic crystal packing effects produce biases in non-bonded contact geometry distributions. For example, the dimer motif shown in Fig. 2 is very common in small-molecule amide crystal structures. Consequently, hydrogen bonds to one of the carbonyl O atom lone pairs (the lone pair proximal to the NH_2 group) are favoured over hydrogen bonds to the other lone pair.

Packing effects and biases can also be important in PDB data. In particular, about 10% of PDB protein–ligand binding sites were found to have close contacts to protein or ligand atoms from neighbouring molecules in the crystal-packing environment (Bergner *et al.*, 2002; Nissink *et al.*, 2002). In cases where this association of protein molecules occurs only in the crystal structure (*i.e.* is not maintained *in vivo*), this effect can seriously jeopardize the relevance of the structure for structure-based inhibitor design.

Although there are very few statistically significant differences between the *geometries* of non-bonded contacts in CSD and PDB structures, the same is not true for non-bonded contact *frequencies*. Verdonk *et al.* (1999) determined the radial distributions of C═O···CH_3 and CH_3···CH_3 contacts in (i) CSD structures and (ii) PDB structures, normalized to account for stoichiometric differences in the frequencies of occurrence of C═O and CH_3 groups. Their plots suggested that hydrophobic (C···C) contacts are relatively much more common in the PDB than in the CSD. Presumably, this is because of the hydrophobic effect (Tanford, 1980), *i.e.* the entropically favoured displacement of molecules from protein cavities consequent upon ligand binding. The corresponding effect is presumably less important or absent for small-molecule crystal formation, not least because most organic crystals are grown from non-aqueous solvents. Confirmation of this result was obtained in the work of Boer *et al.* (2001). Although they found the similarity of CSD-based and PDB-based IsoStar scatter plots to be high when judged by the Carbo index (see above), the same was not true if the Hodgkin index (Hodgkin & Richards, 1987) was used. The two indices differ in that the former takes into account only the shape of the distributions while the latter also takes account of the densities (*i.e.* magnitudes) of the distributions. Detailed analysis confirmed that the difference was a consequence of hydrophobic contacts being relatively more frequent in the PDB than in the CSD.

4. Use of the CSD to study intramolecular geometries

A CSD-based compilation of the means, medians and standard deviations of many types of organic bond lengths (Allen *et al.*, 1987) has been very widely used (Redman *et al.*, 2001) for model building *etc*. Crystallographic data has also been crucial in the estimation of van der Waals radii for the non-metallic

elements (Chothia, 1975; Rowland & Taylor, 1996), which are necessary for 'bump checking', force-field parameterization and the determination of protein packing densities (Tsai *et al.*, 1999). Although not always well referenced, it also seems likely that crystal-structure geometries underlie many of the fragment libraries used for building molecules in modelling packages.

In structure-based drug design, the most important intramolecular geometrical parameters are the torsion angles around rotatable bonds, since they influence the overall shapes of molecules far more than do bond lengths and valence angles. Many theoretical techniques for estimating torsional preferences are available, but none is without problems (Leach, 1991). Supplementing energy calculations with a CSD-based conformational analysis therefore increases the confidence with which conclusions may be drawn about molecular conformations. For example, torsional distributions derived from CSD searches for key fragments were used in combination with force field and *ab initio* MO calculations to investigate the conformational preferences of insecticidal pyrethroids (Mullaley & Taylor, 1994). The work resulted in a proposed biologically active conformation for these highly flexible molecules. In another study, Pirard & Durant (1996) used both CSD-derived torsional distributions and molecular-dynamics (MD) simulations to elucidate the conformational properties of, and propose a pharmacophore pattern for, GABA antagonists. In general, the CSD and MD results were in good agreement except for torsional distributions based on small numbers of CSD entries. CSD data on piperidone-ring geometries were used to correct a theoretical model of renin inhibitors and thereby explain previously unaccountable structure–activity relationships (Lunney & Humblet, 1998). Use of CSD data can be especially useful in elucidating the conformational preferences of medium-sized rings (see, for example, Ghose *et al.*, 1995), which are notoriously difficult to determine.

A crucial breakthrough in computational drug design came with the development of programs such as *AIMB* (Wipke & Hahn, 1986), *CONCORD* (Pearlman, 1987) and *CORINA* (Gasteiger *et al.*, 1990) that can convert a two-dimensional chemical structure into a reasonable three-dimensional geometry (for a review, see Sadowski & Gasteiger, 1993). These programs allow the construction of proprietary three-dimensional (modelled) molecular databases which, while not as accurate as the experimental data in the CSD, enable pharmaceutical companies to perform pharmacophoric and other searches on molecules that are physically available for biological screening. The available two-dimensional to three-dimensional structure converters work in different ways, but most rely on CSD data to some extent, either as a source of bond-length data or for torsion-angle distributions that can be used to help ensure that only low-energy conformers are generated.

Programs for two-dimensional to three-dimensional conversion generally produce a single low-energy conformer for each molecule. A step beyond this is to generate all low-energy conformers or provide a means of performing a conformational search on demand from some starting point. A knowledge-based program for doing this, *MIMUMBA*, was written by Klebe & Mietzner (1994). They produced a library of torsion-angle distributions by searching the CSD for 216 molecular fragments, each containing a rotatable bond, and producing histograms of the observed torsion-angle distributions. Conformational analysis is then performed by partitioning the molecule of interest into rings and open-chain fragments. The CSD torsional data is used to assign likely values to each acyclic torsion and the theoretical program *SCA* (Hoflack *et al.*, 1989) used to identify possible ring geometries. Combining this information produces a list of possible conformations for the molecule as a whole which can be empirically ranked and the best subjected to energy minimization. Part of the procedure involves converting the observed torsional distributions into pseudo-energy curves, a procedure first described by Murray-Rust (1982). *MIMUMBA* was shown to be successful in finding the experimentally observed protein-bound conformations of eight ligands taken from the PDB, the most flexible of which had nine rotatable bonds. In one case, however (methotrexate), several hundred conformations had to be generated before the observed geometry was found.

A further improvement in conformer generation was achieved recently in a *tour de force* of knowledge engineering by workers at Merck Research Laboratories (Feuston *et al.*, 2001). Their program, called *et*, differs significantly from *MIMUMBA* in taking better account of correlations between the torsion angles of adjacent rotatable bonds, which are often very strong and can therefore restrict conformational space. The program is based on about 800 substructural fragments, each containing from one to three variable torsions (or more in the case of seven- and eight-membered ring fragments). A subset of about 18 000 diverse organic molecules from the CSD was used to identify the conformational 'bins' into which each fragment can fall. For a given bin, information was stored about the average torsion angle of each rotatable bond in the fragment, together with its standard deviation. The accumulated expertise of Merck computational chemists was used to refine the raw crystallographic data by eliminating bins that in practice gave poor modelling predictions.

Conformational analysis for a molecule proceeds by finding all the fragments in the library that match the molecule. Of these, the largest fragments are chosen, since they will be the ones that best capture the correlations between adjacent torsions and therefore restrict conformational space most effectively. Where possible, overlapping fragments are used; a torsion angle matching two fragments is restricted to values that are consistent with both. When this is not possible (*i.e.* the two overlapping fragments suggest conflicting values for a torsion angle), a smaller fragment is used for that torsion alone. Once the complete molecule is matched and possible torsion angles assigned, the total number of theoretically possible molecular conformations can be computed. If this is too large, some conformations are eliminated. This is achieved by an algorithm that is biased towards more probable torsion angles (*i.e.* those more frequently observed in the CSD) but

also takes account of the diversity of conformers selected. A final step in evaluating conformations is to check for bumps between atoms in the molecule that were matched onto uncorrelated substructural fragments. The program can deal with rings up to size eight; beyond that, the conformational possibilities of the ring become too large and the amount of available crystallographic data too small.

The program was validated by generating conformers for 113 molecules whose protein-bound conformations have been determined and deposited in the PDB. In comparison with a distance-geometry algorithm (Blaney *et al.*, 1990), the knowledge-based approach was found to be faster and more efficient in finding the experimentally observed conformations. For example, when *et* was used to generate 25 conformations for each ligand, a conformation within 1.5 Å RMSD of the observed ligand geometry was found in about 90 of the 113 cases. The corresponding figure for the distance-geometry method was <80.

5. Use of the CSD to study metal coordination

The ability to predict ligand binding to metal ions is important because of the many metalloenzymes of known pharmaceutical relevance. Information derived from the CSD and the PDB can help in several ways. Orpen *et al.* (1989) compiled a large list of average distances for bonds between metals and ligand donor atoms. Later, other authors produced more focused compilations concentrating on the metal ions most commonly found in enzymes, *e.g.* Ca, Mg, Mn, Fe, Cu and Zn. Harding (1999), for example, gives values for bond lengths involving these ions (in various oxidation states, where appropriate) and donor atoms of the types found in amino acids, *viz* carboxylate, alcohol and phenolate O atoms, imidazole N atoms and thiolate S atoms.

Perhaps more important are data on metal coordination numbers and polyhedral geometries. A review by Glusker (1991) includes histograms (based on the CSD) of metal-ion coordination numbers for Mg (most likely coordination numbers: 6, 4), Na (6), Ca (8, 6), K (8, 7, 6), Zn (4, 6), Cd (6), Fe (6), Co (6), Cu (4, 5) and Mo (6). It also includes tables of observed coordination polyhedra and lists the types of atoms most likely to coordinate to given metal ions (which is primarily determined by the hardness or softness of the ion). For some ions, especially the alkali and alkaline earth metals, it can be difficult to define the limiting distances for metal-ligand bond lengths and, in consequence, coordination numbers become uncertain.

Several authors (Glusker, 1991; Harding, 1999 and references therein) have looked at the geometry of metal coordination with respect to the donor-atom group. For example, the interaction of a single metal ion with an isolated carboxylate group can occur at either the *syn* or *anti* oxygen lone-pair positions, or in the 'direct' position symmetrically placed between the two O atoms. Crystal-structure statistics show that the occurrence of these geometries decreases in the order *syn* (62.9%) > *anti* (22.7%) > direct (14.4%) (Glusker, 1991). The metal ion usually lies close to the carboxylate plane, though out-of-plane distortions can occur and are more common for some metals than for others. Information thus derived about the probability of a metal ion binding in a given position around a ligand group can be used to identify metal-binding sites in proteins. For example, Carrell *et al.* (1989) used this method to locate two binding sites in xylose isomerase, one involving three carboxylates, a histidine and a water and the other involving four carboxylates and a water.

Crystal-structure data are particularly useful in studying metal complexation because theoretical calculations can be very difficult indeed. This is well illustrated by a comparison between crystal-structure and computed (*ab initio*) geometries of mono-anionic and di-anionic phosphate groups coordinated to sodium ions (Schneider *et al.*, 1996). Good agreement between calculated and observed geometries was only achieved when polarization functions were included on the basis sets and a sodium counter-ion was added to the model system. More seriously, the experimental data showed that sodium binding to di-anionic phosphates was often mediated by water. The sodium ions tended to cluster in two positions, both lying outside the O═P═O plane and each interacting with only one of the charged O atoms. In contrast, the theoretically predicted position was located symmetrically between the two O atoms and in the O═P═O plane. This discrepancy was resolved only when a water molecule was included in the model system.

6. Use of the CSD to study intermolecular interactions

One of the first studies of non-bonded interactions based on a systematic analysis of crystal-structure data was performed by Kroon *et al.* (1975). They characterized the geometries of OH$\cdots$O hydrogen bonds and showed, *inter alia*, that hydrogen bonds to hydroxyl and ether oxygen acceptors do not preferentially form along the sp^3 lone-pair directions. This conclusion was later supported by an analysis based solely on structures determined by neutron diffraction, in which H atoms were located with high precision (Ceccarelli *et al.*, 1981). In contrast, hydrogen bonds to sp^2-hybridized O atoms were shown to have a statistically significant preference for the lone-pair directions (Taylor *et al.*, 1983; Murray-Rust & Glusker, 1984; see also Tintelnot & Andrews, 1989; Mills & Dean, 1996). Other early CSD-based studies provided the first conclusive evidence for the ability of some C—H groups to hydrogen bond (Taylor & Kennard, 1982) and showed that electrophiles approach divalent S atoms (C—S—C) in a direction different from nucleophiles (the former approach approximately along the normal to the C—S—C plane and the latter approach along the extension of the C—S valence-bond directions; Rosenfield *et al.*, 1977).

A key methodological advance came with the conversion of data to scatter plots and thence to contoured density plots (Rosenfield *et al.*, 1984). In this procedure, the CSD is searched for intermolecular interactions between a functional group, *A*, and an atom, *B*. The various *A*$\cdots$*B* contacts are superimposed by least-squares overlaying of the atoms of *A*. The resulting plot shows the experimentally observed three-

dimensional distribution of *B* atoms around *A*. This can then be embedded in a grid and the number of *B* atoms in each grid cube counted. Contouring on these counts produces a density plot showing the probability distribution of *B* around *A*. This methodology was used sporadically by several workers (*e.g.* Glusker, 1995) to elucidate the geometrical preferences of various interactions and was finally used to create the IsoStar database (Bruno *et al.*, 1997). This is a large compendium of scatter plots and density plots showing the geometrical distributions of thousands of different types of intermolecular interactions, including hydrogen bonds, hydrophobic contacts and electrostatic interactions.

The importance of such information to structure-based drug design lies in the non-covalent nature of most enzyme–inhibitor binding. Successful design therefore requires an understanding of both the preferred geometries of non-bonded contacts and the likelihood of their occurrence. The role of the CSD in elucidating the directional preferences of hydrogen bonds has already been mentioned. IsoStar scatter plots show that other types of interactions are equally directional. For example, nitro group N atoms show a strong tendency to form contacts to O atoms in a direction approximately normal to the plane of the nitro group (Taylor *et al.*, 1990). Even hydrophobic contacts can show strong directional preferences (Cole *et al.*, 1998). Of particular importance (because it is so common) is the interaction between a phenyl group and another aromatic ring. Here, contact geometries that lead to electrostatically favourable quadrupole–quadrupole interactions are preferred (Hunter, 1994; Klebe & Diederich, 1993).

IsoStar contains data on many attractive non-bonded interactions that might be exploited in rational ligand design. However, identifying contacts that are *unlikely* to occur is just as important. For example, crystallographic data show that organically bound fluorine is very unlikely to accept a hydrogen bond (Dunitz & Taylor, 1997), and both crystal-structure studies and *ab initio* calculations indicate that aromatic oxygen is a very weak hydrogen-bond acceptor (Boehm *et al.*, 1996). In consequence, hydrogen bonds to these types of atoms are probably to be avoided in structure-based inhibitor design. Conversely, IsoStar is also useful as a source of novel interactions that might be used to increase novelty in structure-based inhibitor design. For example, an obvious way to effect binding to a tryptophan residue in an enzyme active site is to design a ligand that can hydrogen bond to the indole NH proton. Inspection of IsoStar scatter plots, however, suggests at least four alternative approaches, *viz* including an electron-deficient hydrophobic group in the ligand that can stack with the electron-rich indole ring, including in the ligand an aromatic ring that can form an 'edge-to-face' interaction with the indole ring, forming an NH$\cdots\pi$ bond between the ligand and the indole aromatic ring and forming CH$\cdots$O interactions between ligand O atoms and the CH atoms on the indole ring.

Detailed examination of the CSD, the PDB and IsoStar can sometimes show that a non-covalent interaction has important structural consequences that have previously gone unrecognized. For example, Umezawa *et al.* (1999) suggest that CH$\cdots\pi$ interactions are responsible for many solid-state conformations of cyclic and acyclic peptides. Another good example is the dipolar interaction between proximal carbonyl groups (Allen *et al.*, 1998). It has been shown (Maccallum *et al.*, 1995*a*,*b*) that this interaction between backbone protein carbonyl groups is a stabilizing factor in α-helices, β-sheets and the right-handed twist often observed in β-strands. Further analysis of PDB structures indicated that interactions between the side-chain carbonyls of asparagine and aspartate residues and neighbouring backbone carbonyls may explain the known tendency for these residues to tolerate left-handed α-helical regions more readily than do other non-glycyl amino acids (Deane *et al.*, 1999).

7. Contribution of the CSD to life-science applications programs

Apart from those already mentioned, CSD data has contributed directly or indirectly to a great many life-science applications programs. For example, the program *GRID* (Goodford, 1985) uses empirical energy functions parameterized, in part, to reproduce the geometrical distributions of non-bonded contacts taken from the CSD. The program identifies binding 'hot-spots' in enzyme active sites. A probe atom is moved around the active site and its interaction energy computed as a summation of its pairwise atom$\cdots$atom interactions with the atoms of the protein. The results can be displayed as a contoured energy surface which can be used to aid structure-based inhibitor design. Recently, the programs *X-SITE* (Laskowski *et al.*, 1996) and *SuperStar* (Verdonk *et al.*, 1999, 2001) were developed to achieve the same aim by using a purely knowledge-based approach. The enzyme active site is broken down into its constituent functional groups. The crystallographically observed probability distribution of the chosen probe atom around each functional group is retrieved from IsoStar or an equivalent database and overlaid onto the group in the active site. Overlapping probability distributions are combined, for example by multiplication (this requires that the individual distributions are normalized to the same scale; *X-SITE* and *SuperStar* use slightly different procedures for doing this). The final result is a composite map for the entire active site showing the likely points at which the probe atom will bind. The use of similar methodologies to produce knowledge-based force fields has been explored by several workers, for example Sippl (1995).

The use of protein–ligand docking programs for virtual screening is becoming increasingly important and many of the leading programs exploit crystallographic data. Both *GOLD* (Jones *et al.*, 1997, 1999) and *FlexX* (Rarey *et al.*, 1996) use information derived ultimately from small-molecule crystal-structure data to determine whether a putative protein–ligand hydrogen bond has an energetically favourable geometry. Additionally, both programs use torsional distributions from the CSD. In *GOLD*, these are used to focus the conformational search of the complete ligand molecule into favoured regions of conformer space. In *FlexX*, the ligand is built up

within the binding site by combining the rigid fragments from which it is composed and CSD information is used to determine the torsion angles around the bonds linking these rigid fragments.

One step on from docking is automated *de novo* ligand design, where a putative ligand molecule is designed and built algorithmically within a protein binding site. A pioneering program of this type is *LUDI* (Boehm, 1992*a*,*b*). It works by placing molecular fragments (taken from a library) in a protein binding site in positions where they are expected to interact favourably with the protein. Fragments are connected to one another by single bonds, bridges (*i.e*. two single bonds with a spacer atom in the middle), fusion *etc.* A scoring function estimates the likely binding affinity of the complete ligand. The role of CSD data is to guide the initial placement of fragments. Searches of the CSD were performed to elucidate the angular and dihedral ranges in which polar fragment atoms could form hydrogen bonds. The space enclosed by these ranges is filled with an ensemble of potential interaction sites. Fitting such a site onto a complementary hydrogen-bond atom therefore places that atom and the fragment of which it is part in a position where a geometrically reasonable protein–ligand hydrogen bond is possible. A similar methodology for hydrophobic contacts, together with bump checking to exclude close contacts, completes the placement procedure.

It is frequently the case that drug invention must be achieved without access to the three-dimensional structure of the target protein. In this case, rational design is usually restricted to looking for similarities between known actives and possible synthetic targets. Three-dimensional methods for doing this usually require as a precursor that the molecules are superimposed in some way, which is non-trivial when the molecules are flexible and when there are many of them to compare. Several groups have published superposition methods that utilize CSD-derived information in some way. For example, *MIMUMBA* (see above) has been used to generate likely conformers for molecules prior to superposition (Klebe, Mietzner *et al.*, 1994; Klebe *et al.*, 1999). The *MIMUMBA* torsion-angle database is also used in the *FlexS* program (Lemmen & Lengauer, 1997; Lemmen *et al.*, 1998) which superimposes a flexible molecule onto a rigid molecule by partitioning into fragments. Superposition of a central anchor fragment is performed first. The remaining fragments are then added one by one, the torsion-angle database being used to guide the dihedral geometries around the linking bonds.

Several groups (*e.g*. Jain *et al.*, 1994) have noted that instead of overlaying molecules by least-squares fitting of their atomic positions, superpositions can be performed by matching points in space around the molecules, *e.g*. those at which hydrogen-bond donors and acceptors are likely to be positioned. These 'interaction points' can be determined by using non-bonded-contact information derived from the CSD (*e.g*. Mills *et al.*, 1997, 2001). The method acknowledges the fact that ligand hydrogen-bonding atoms do not need to be superimposed on each other in order to be able to interact with the same complementary receptor atom. It also allows molecules to be overlaid that are ostensibly chemically dissimilar, provided they are able to form the same pattern of intermolecular interactions.

If molecular superposition is fast enough, it can be used as the basis of a three-dimensional similarity search of a database. Even the fastest algorithms, however, are inadequate for searching extremely large databases, perhaps containing many conformers per molecule. Nicholls (2001) described a particularly innovative use of the CSD for speeding up a three-dimensional shape-similarity search by pre-screening. Each entry in the database to be searched is assigned a shape fingerprint consisting of a string of 1000 bits. This is performed by superimposing the molecule optimally on each of 1000 diverse molecules chosen from the CSD. Any comparison resulting in a shape-similarity score over a certain threshold sets the corresponding bit 'on' in the fingerprint. A search of the database can then be pre-screened by generating the query molecule's fingerprint and immediately eliminating from the search any database entry whose fingerprint does not match.

Once overlaid, a set of molecules can be used to derive three-dimensional QSAR relationships, *e.g*. using *CoMFA* (Cramer *et al.*, 1988) or *CoMSIA* (Klebe, Abraham *et al.*, 1994). It has been shown that the use of *SuperStar* fields as descriptors can produce statistically superior correlations compared with those obtainable from the hydrogen-bond descriptors provided in the standard implementation of *CoMSIA* (Boehm & Klebe, 2002). *SuperStar* and IsoStar fields can also be used to generate descriptors for functional group as a means of expressing functional-group similarity (Nissink *et al.*, 2000; Watson *et al.*, 2002).

8. Conclusions

The CSD has contributed profoundly to many aspects of the life sciences. Far from diminishing in importance as alternative, theoretical, techniques become more powerful or as the PDB grows in size, the use of the CSD in the life sciences is increasing. This is partly a consequence of the continuing value of high-precision structural data and also of the increasingly innovative ways in which CSD data are being exploited, often in synergy with other techniques. The extraction of knowledge from the CSD and its encapsulation in derived databases such as IsoStar opens many interesting possibilities for the future, especially when such derived databases are used to 'drive' knowledge-based applications like *SuperStar*. It may increasingly become the case that CSD data will be used 'behind the scenes' to aid life-science molecular modelling, without the end-user necessarily seeing the CSD search interface itself.

All staff of the Cambridge Crystallographic Data Centre, past and present, are thanked for their efforts in building and maintaining the CSD System.

References

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.

Allen, F. H., Baalham, C. A., Lommerse, J. P. M. & Raithby, P. R. (1998). *Acta Cryst.* B**54**, 320–329.

Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. G. (1991). *J. Chem. Inf. Comput. Sci.* **31**, 187–204.

Allen, F. H., Harris, S. E. & Taylor, R. (1996). *J. Comput. Aided Mol. Des.* **10**, 247–254.

Allen, F. H., Kennard, O., Watson, D. G., Brammer, L., Orpen, A. G. & Taylor, R. (1987). *J. Chem. Soc. Perkin Trans. II*, pp. S1–S19.

Bergner, A., Guenther, J., Hendlich, M., Klebe, G. & Verdonk, M. (2002). Submitted.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Ravichanran, V., Schneider, B., Thanki, N., Padilla, D., Weissig, H., Westbrook, J. D. & Zardecki, C. (2002). *Acta Cryst.* D**58**, 899–907.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Bilton, C., Allen, F. H., Shields, G. P. & Howard, J. A. K. (2000). *Acta Cryst.* B**56**, 849–856.

Blaney, J. M., Crippen, G. M., Dearing, A. & Dixon, J. S. (1990). *DGEOM. Distance Geometry Program.* QCPE Program No. 590. Quantum Chemistry Program Exchange, Indiana University, Bloomington, USA.

Boehm, H.-J. (1992*a*). *J. Comput. Aided Mol. Des.* **6**, 61–78.

Boehm, H.-J. (1992*b*). *J. Comput. Aided Mol. Des.* **6**, 593–606.

Boehm, H.-J., Brode, S., Hesse, U. & Klebe, G. (1996). *Chem. Eur. J.* **2**, 1509–1513.

Boehm, H.-J. & Klebe, G. (1996). *Angew. Chem. Int. Ed. Engl.* **35**, 2588–2614.

Boehm, M. & Klebe, G. (2002). Submitted.

Boer, D. R., Kroon, J., Cole, J. C., Smith, B. & Verdonk, M. L. (2001). *J. Mol. Biol.* **312**, 275–287.

Brock, C. P. & Minton, R. P. (1989). *J. Am. Chem. Soc.* **111**, 4586–4593.

Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Cryst.* B**58**, 389–397.

Bruno, I. J., Cole, J. C., Lommerse, J. P. M., Rowland, R. S., Taylor, R. & Verdonk, M. L. (1997). *J. Comput. Aided Mol. Des.* **11**, 525–537.

Bürgi, H.-B. & Dunitz, J. D. (1988). *Acta Cryst.* B**44**, 445–448.

Cambridge Crystallographic Data Centre (1995). *Vista Version 2.0.* Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ.

Carbo, R., Leyda, L. & Arnau, M. (1980). *Int. J. Quant. Chem.* **27**, 1185–1189.

Carrell, H. L., Glusker, J. P., Burger, V., Manfre, F., Tritsch, D. & Biellmann, J.-F. (1989). *Proc. Natl Acad. Sci. USA*, **86**, 4440–4444.

Ceccarelli, C., Jeffrey, G. A. & Taylor, R. (1981). *J. Mol. Struct.* **70**, 255–271.

Chakrabarti, P. & Hsu, B. T. (1994). *Inorg. Chem.* **33**, 1165–1170.

Chothia, C. (1975). *Nature (London)*, **254**, 304–308.

Chowdhury, S. F., Di Lucrezia, R., Guerrero, R. H., Brun, R., Goodman, J., Ruiz-Perez, L. M., Pacanowska, D. G. & Gilbert, I. H. (2001). *Bioorg. Med. Chem. Lett.* **11**, 977–980.

Cole, J. C., Taylor, R. & Verdonk, M. L. (1998). *Acta Cryst.* D**54**, 1183–1193.

Cramer, R. D., Patterson, D. E. & Bunce, J. D. (1988). *J. Am. Chem. Soc.* **110**, 5959–5967.

Deane, C. M., Allen, F. H., Taylor, R. & Blundell, T. L. (1999). *Protein Eng.* **12**, 1025–1028.

DesJarlais, R. L., Seibel, G. L., Kuntz, I. D., Furth, P. S., Alvarez, J. C., Ortiz de Montellano, P. R., DeCamp, D. L., Babe, L. M. & Craik, C. S. (1990). *Proc. Natl Acad. Sci. USA*, **87**, 6644–6648.

Dunitz, J. D. & Taylor, R. (1997). *Chem. Eur. J.* **3**, 89–98.

Feuston, B. P., Miller, M. D., Culberson, J. C., Nachbar, R. B. & Kearsley, S. K. (2001). *J. Chem. Inf. Comput. Sci.* **41**, 754–763.

Gasteiger, J., Rudolph, C. & Sadowski, J. (1990). *Tetrahedron Comput. Methods*, **3**, 537–547.

Ghose, A. K., Logan, M. E., Treasurywala, A. M., Wang, H., Wahl, R. C., Tomczuk, B. E., Gowravaram, M. R., Jaeger, E. P. & Wendoloski, J. J. (1995). *J. Am. Chem. Soc.* **117**, 4671–4682.

Glusker, J. P. (1991). *Adv. Protein Chem.* **42**, 1–76.

Glusker, J. P. (1995). *Acta Cryst.* D**51**, 418–427.

Goodford, P. J. (1985). *J. Med. Chem.* **28**, 849–857.

Grootenhuis, P. D. J., Roe, D. C., Kollman, P. A. & Kuntz, I. D. (1994). *J. Comput. Aided Mol. Des.* **8**, 731–750.

Harding, M. M. (1999). *Acta Cryst.* D**55**, 1432–1443.

Hodgkin, E. E. & Richards, W. G. (1987). *Int. J. Quant. Chem.* **14**, 105–110.

Hoflack, J., DeClercq, P. J. & Cauwberghs, S. (1989). *SCA. Systematic Conformational Analysis.* QCPE Program No. QCMP079. Quantum Chemistry Program Exchange, Indiana University, Bloomington, USA.

Hunter, C. A. (1994). *Chem. Rev.* **94**, 101–109.

Jain, A. N., Dietterich, T. G., Lathrop, R. H., Chapman, D., Critchlow, R. E., Bauer, B. E., Webster, T. A. & Lozano-Perez, T. (1994). *J. Comput. Aided Mol. Des.* **8**, 635–652.

Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. (1997). *J. Mol. Biol.* **267**, 727–748.

Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. (1999). *Rational Drug Design*, edited by A. L. Parrill & M. R. Reddy, pp. 271–291. Washington DC: American Chemical Society.

Klebe, G., Abraham, U. & Mietzner, T. (1994). *J. Med. Chem.* **37**, 4130–4146.

Klebe, G. & Diederich, F. (1993). *Philos. Trans. R. Soc. London Ser. A*, **345**, 37–48.

Klebe, G. & Mietzner, T. (1994). *J. Comput. Aided Mol. Des.* **8**, 583–606.

Klebe, G., Mietzner, T. & Weber, F. (1994). *J. Comput. Aided Mol. Des.* **8**, 751–778.

Klebe, G., Mietzner, T. & Weber, F. (1999). *J. Comput. Aided Mol. Des.* **13**, 35–49.

Kroon, J., Kanters, J. A., van Duijneveldt-van de Rijdt, J. G. C. M., van Duijneveldt, F. B. & Vliegenthart, J. A. (1975). *J. Mol. Struct.* **24**, 109–129.

Lam, P. Y. S., Prabhakar, K. J., Eyermann, C. J., Hodge, C. N., Ru, Y., Bacheler, L. T., Meek, J. L., Otto, M. J., Rayner, M. M., Wong, Y. N., Chang, C.-H., Weber, P. C., Jackson, D. A., Sharpe, T. R. & Erickson-Viitanen, S. (1994). *Science*, **263**, 380–384.

Laskowski, R. A., Thornton, J. M., Humblet, C. & Singh, J. (1996). *J. Mol. Biol.* **259**, 175–201.

Leach, A. R. (1991). *Reviews in Computational Chemistry*, Vol. 2, edited by K. B. Lipkowitz & D. B. Boyd, pp. 1–55. New York: VCH.

Leherte, L., Latour, T. & Vercauteren, D. P. (1996). *J. Comput. Aided Mol. Des.* **10**, 55–66.

Lemmen, C. & Lengauer, T. (1997). *J. Comput. Aided Mol. Des.* **11**, 357–368.

Lemmen, C., Lengauer, T. & Klebe, G. (1998). *J. Med. Chem.* **41**, 4502–4520.

Lunney, E. A. & Humblet, C. (1998). *Structure-Based Ligand Design*, edited by K. Gubernator & H.-J. Boehm, pp. 37–71. New York: Wiley.

Maccallum, P. H., Poet, R. & Milner-White, E. J. (1995*a*). *J. Mol. Biol.* **248**, 361–373.

Maccallum, P. H., Poet, R. & Milner-White, E. J. (1995*b*). *J. Mol. Biol.* **248**, 374–384.

Mills, J. E. J. & Dean, P. M. (1996). *J. Comput. Aided Mol. Des.* **10**, 607–622.

Mills, J. E. J., de Esch, I. J. P., Perkins, T. D. J. & Dean, P. M. (2001). *J. Comput. Aided Mol. Des.* **15**, 81–96.

Mills, J. E. J., Perkins, T. D. J. & Dean, P. M. (1997). *J. Comput. Aided*

Mol. Des. **11**, 229–242.
Mullaley, A. & Taylor, R. (1994). *J. Comput. Aided Mol. Des.* **8**, 135–152.
Murray-Rust, P. (1982). *Molecular Structure and Biological Activity*, edited by J. F. Griffin & W. L. Duax, pp. 117–133. New York: Elsevier.
Murray-Rust, P. & Glusker, J. P. (1984). *J. Am. Chem. Soc.* **106**, 1018–1025.
Nicholls, A. (2001). American Chemical Society Conference, Chicago, USA.
Nissink, J. W. M., Murray, C., Hartshorn, M., Verdonk, M. L., Cole, J. C. & Taylor, R. (2002). Submitted.
Nissink, J. W. M., Verdonk, M. L. & Klebe, G. (2000). *J. Comput. Aided Mol. Des.* **14**, 787–803.
Orpen, A. G., Brammer, L., Allen, F. H., Kennard, O., Watson, D. G. & Taylor, R. (1989). *J. Chem. Soc. Dalton Trans.*, pp. S1–S83.
Pearlman, R. S. (1987). *Chem. Des. Autom. News*, **2**, 1.
Pirard, B. & Durant, F. (1996). *J. Comput. Aided Mol. Des.* **10**, 31–40.
Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. (1996). *J. Mol. Biol.* **261**, 470–489.
Redman, J., Willett, P., Allen, F. H. & Taylor, R. (2001). *J. Appl. Cryst.* **34**, 375–380.
Rosenfield, R. E. Jr., Parthasarathy, R. & Dunitz, J. D. (1977). *J. Am. Chem. Soc.* **99**, 4860–4862.
Rosenfield, R. E. Jr, Swanson, S. M., Meyer, E. F. Jr, Carrell, H. L. & Murray-Rust, P. (1984). *J. Mol. Graph.* **2**, 43–46.
Rowland, R. S. & Taylor, R. (1996). *J. Phys. Chem.* **100**, 7384–7391.
Sadowski, J. & Gasteiger, J. (1993). *Chem. Rev.* **93**, 2567–2581.
Sadowski, J., Gasteiger, J. & Klebe, G. (1994). *J. Chem. Inf. Comput. Sci.* **34**, 1000–1008.
Sakamoto, J., Nakagawa, T., Kanehisa, N., Kai, Y. & Katsura, M. (2000). *Acta Cryst.* C**56**, e485.
Schneider, B., Kabelac, M. & Hobza, P. (1996). *J. Am. Chem. Soc.* **118**, 12207–12217.
Sippl, M. J. (1995). *Curr. Opin. Struct. Biol.* **5**, 229–235.
Tanford, C. (1980). *The Hydrophobic Effect*, 2nd ed. New York: Wiley.
Taylor, R. (1981). *J. Mol. Struct.* **73**, 125–136.
Taylor, R. & Kennard, O. (1982). *J. Am. Chem. Soc.* **104**, 5063–5070.
Taylor, R., Kennard, O. & Versichel, W. (1983). *J. Am. Chem. Soc.* **105**, 5761–5766.
Taylor, R., Mullaley, A. & Mullier, G. W. (1990). *Pest. Sci.* **29**, 197–213.
Tintelnot, M. & Andrews, P. (1989). *J. Comput. Aided Mol. Des.* **3**, 67–84.
Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. (1999). *J. Mol. Biol.* **290**, 253–266.
Umezawa, Y., Tsuboyama, S., Takahashi, H., Uzawa, J. & Nishio, M. (1999). *Bioorg. Med. Chem.* **7**, 2021–2026.
Verdonk, M. L., Cole, J. C. & Taylor, R. (1999). *J. Mol. Biol.* **289**, 1093–1108.
Verdonk, M. L., Cole, J. C., Watson, P., Gillet, V. & Willett, P. (2001). *J. Mol. Biol.* **307**, 841–859.
Watson, P., Willett, P., Gillet, V. J. & Verdonk, M. L. (2002). Submitted.
Wipke, W. T. & Hahn, M. A. (1986). *Applications of Artificial Intelligence in Chemistry*, edited by T. Pierce & B. Hohne, pp. 136–146. Washington DC: American Chemical Society.