Journal of Computer-Aided Molecular Design, 11 (1997) 525–537.
KLUWER/ESCOM

J-CAMD 418

IsoStar: A library of information about nonbonded interactions*

Ian J. Bruno, Jason C. Cole, Jos P.M. Lommerse, R. Scott Rowland**, Robin Taylor*** and Marcel L. Verdonk

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, U.K.

Received 27 April 1997
Accepted 16 July 1997

Keywords: Nonbonded interactions; Cambridge Structural Database; Brookhaven Protein Databank; Intermolecular perturbation theory

Summary

Crystallographic and theoretical (ab initio) data on intermolecular nonbonded interactions have been gathered together in a computerised library ('IsoStar'). The library contains information about the nonbonded contacts formed by some 250 chemical groupings. The data can be displayed visually and used to aid protein–ligand docking or the identification of bioisosteric replacements. Data from the library show that there is great variability in the geometrical preferences of different types of hydrogen bonds, although in general there is a tendency for H-bonds to form along lone-pair directions. The H-bond acceptor abilities of oxygen and sulphur atoms are highly dependent on intramolecular environments. The nonbonded contacts formed by many hydrophobic groups show surprisingly strong directional preferences. Many unusual nonbonded interactions are to be found in the library and are of potential value for designing novel biologically active molecules.

Introduction

Information about nonbonded interactions is important in many research applications, such as rational drug design [1], supramolecular chemistry [2] and crystal engineering [3]. Scientists working in these areas are interested in the types of interactions that occur between molecules, their energies, and their preferred geometries. In principle, a huge amount of relevant data is available in the literature and in various databases, notably the Cambridge Structural Database (CSD), which contains small-molecule organic and organometallic crystal structures [4], and the Brookhaven Protein Databank (PDB), which contains crystal structures of biological macromolecules [5]. In particular, studies of crystal-structure data have contributed to our knowledge of chemical reaction pathways [6,7], side-chain interactions in protein structures [8–10], protein–ligand binding [11,12], hydrogen-bond geometry [13–15] and crystal packing [16–18], and have highlighted the structural importance of weakly attractive nonbonded interactions [19–21]. A good general review may be found in Ref. 22.

Despite the success of these and other studies, they are difficult to carry out because the data are split over several sources and are not available in a form that is easily assimilated by the chemist and easily read by molecular-modelling programs. The present paper describes an attempt to alleviate these problems. We have developed a computerised library of information about intermolecular nonbonded interactions, based on experimental information from the CSD and PDB and theoretical energy calculations using intermolecular perturbation theory (IMPT) [23]. The library (called 'IsoStar') contains details of the nonbonded interactions formed by about 250 common chemical groupings. For each group, information is held about the types of contacts formed by the group, their geometries and (sometimes) their energies.

Our rationale for using three different types of data is that they are mutually complementary. Crystal structures of small molecules are generally of high experimental

*Dedicated to Professors Jenny Glusker and Jack Dunitz, who have done much to pioneer and further the use of the Cambridge Structural Database.

**Present address: BioCryst Pharmaceuticals Inc., 2190 Parkway Lake Drive, Birmingham, AL 35244, U.S.A.
***To whom correspondence should be addressed.

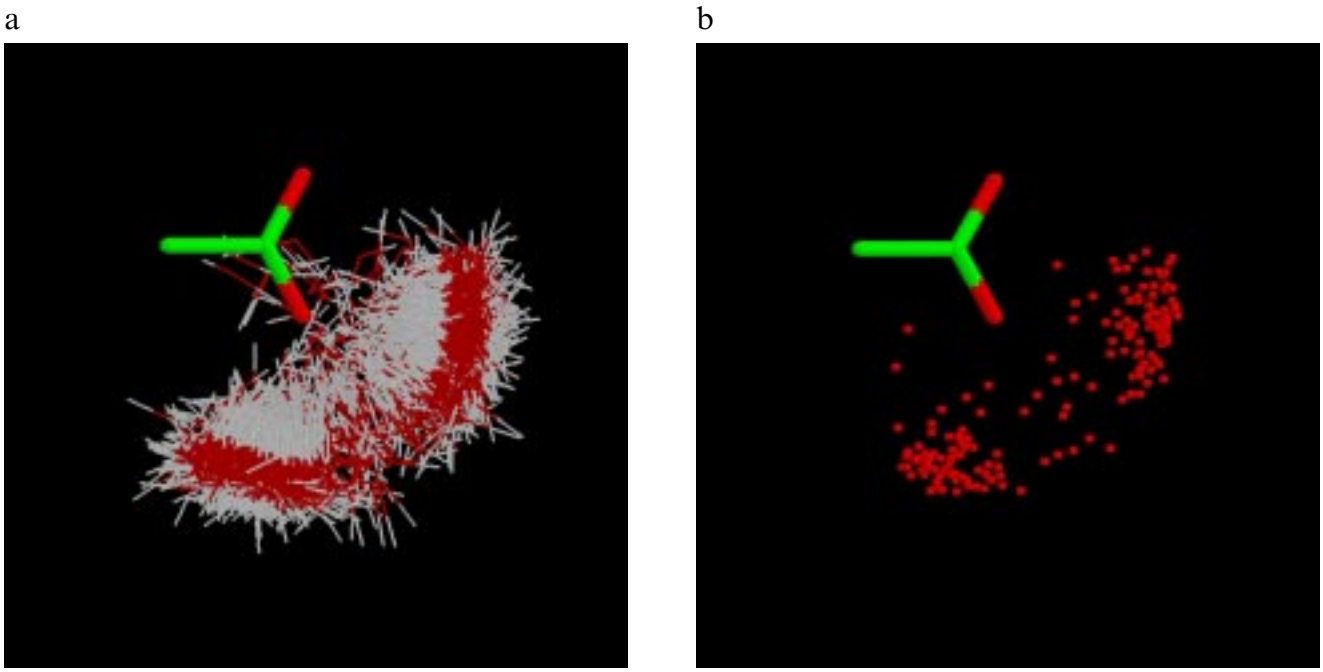


Fig. 1. (a) CSD-based scatterplot of water molecules around carboxylate; (b) PDB-based scatterplot of water oxygens around ionised and unionised carboxylic acid. In these and other scatterplots, except where otherwise stated, all contact groups have at least one atom within V – 0.1 Å of a central-group target atom, where V is the sum of the van der Waals radii of the atoms involved.

precision (e.g. hydrogen atoms are often located) and cover a diverse range of chemical types. However, interactions in a crystal lattice may be biased by systematic crystal-packing forces. Crystal structures of protein–ligand complexes are directly relevant to rational drug design, but are of much lower experimental precision (e.g. hydrogen atoms are almost never located) and, at present, cover a limited range of chemical functionality. Molecular orbital calculations give direct estimates of interaction energies, which are unobtainable from crystal-structure data, but compromises are necessary in choosing basis sets and model compounds. Most high-level molecular

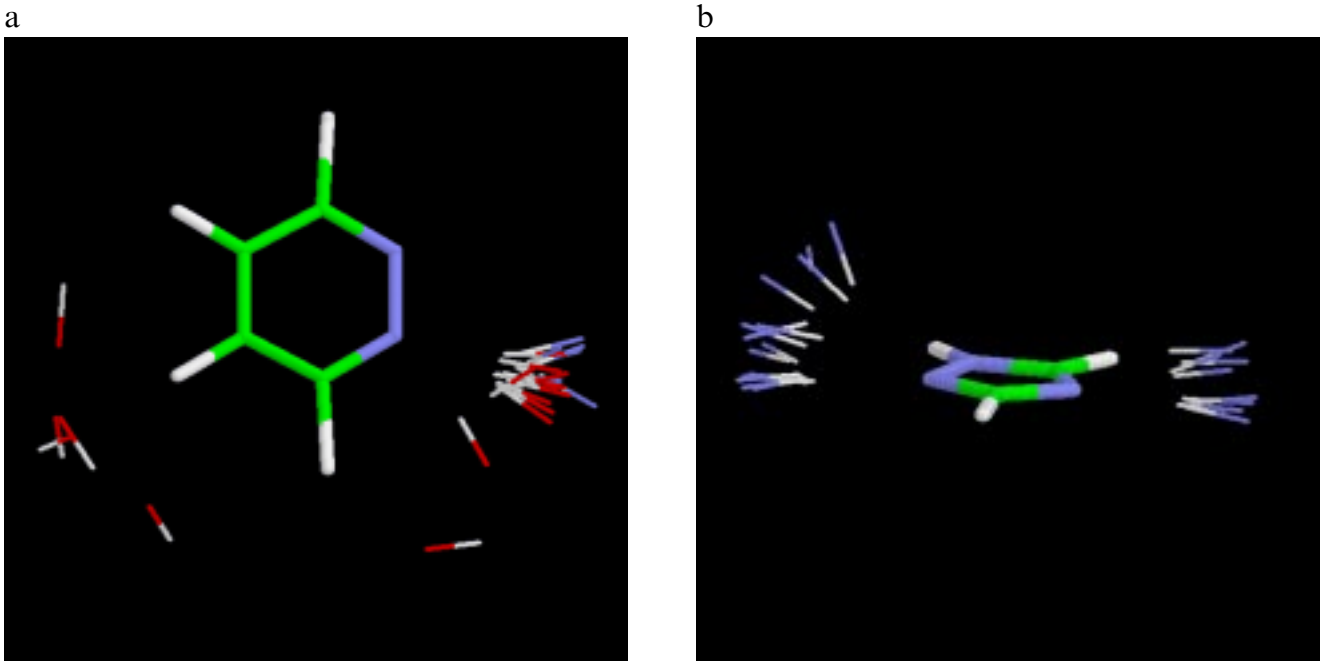


Fig. 2. CSD-based scatterplots of (a) XH groups (X = N,O,S) around pyridazine, and (b) NH groups around 1,2,4-triazole.

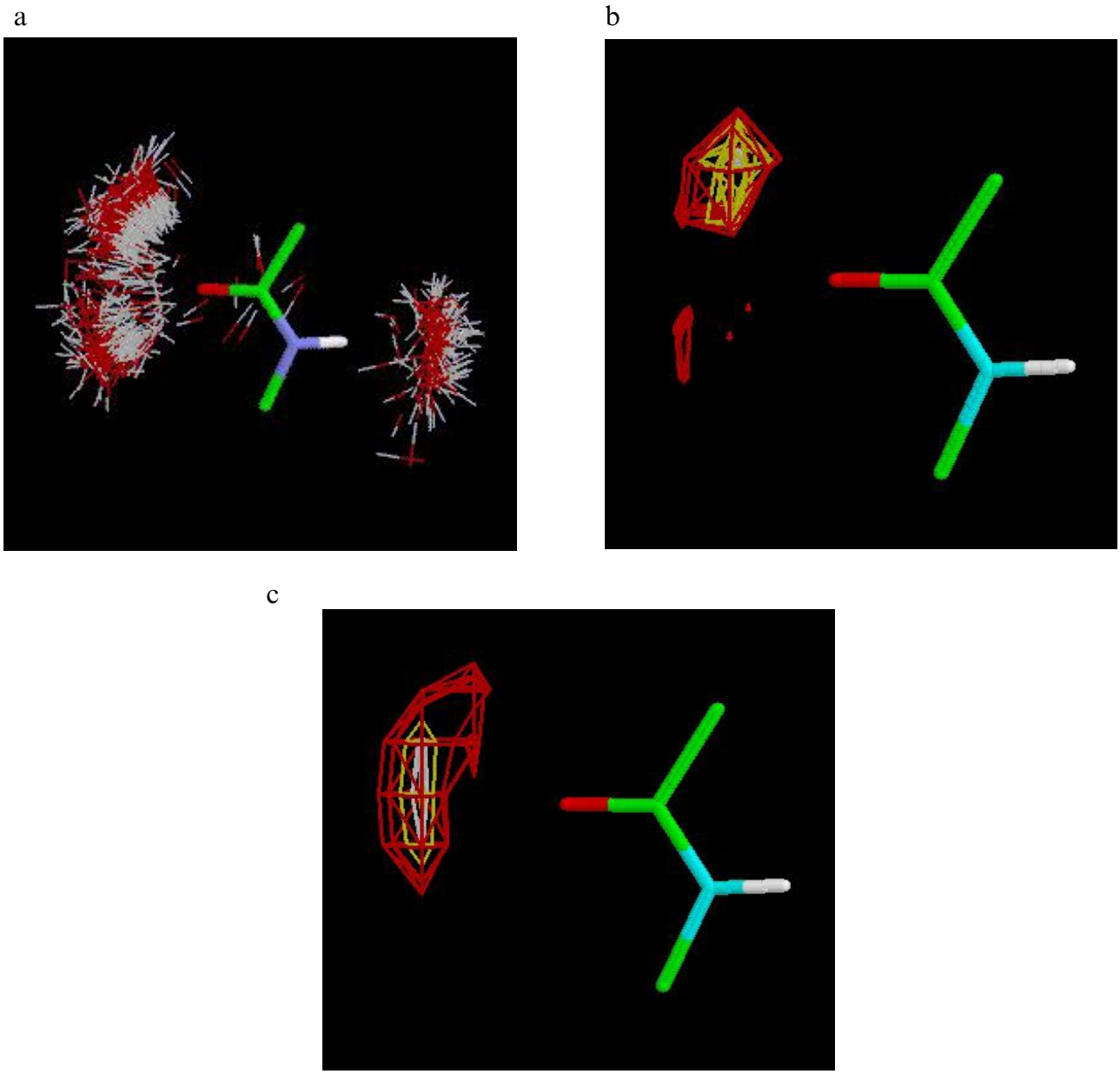


Fig. 3. (a) CSD-based scatterplot of OH around aliphatic amide linkages; (b) contoured density surface computed from scatterplot shown in (a); (c) contoured density surface of NH around aliphatic amide linkages (CSD data).

orbital calculations are so computer intensive that solvation effects cannot be taken into account, i.e. the calculations pertain to an in vacuo situation.

Methods

Searches for nonbonded contacts in small-molecule crystal structures

Nonbonded contacts in small-molecule (i.e. CSD) crystal structures were found with the QUEST package [24], which can be used to search the CSD for intermolecular contacts between any specified pair of chemical groups. Throughout this work, an intermolecular contact was defined as any contact shorter than V + 0.5 Å, where V is the sum of the Bondi van der Waals radii [25] of the atoms involved. In performing searches, X-H covalent bond lengths (X = C,N,O) were 'normalised'. Normalisation of X-H distances involves moving the H-atom along the observed X-H vector until the X-H distance is equal to a standard value (C-H = 1.083 Å, N-H = 1.009 Å, O-H = 0.983 Å) [24,26,27].

All structures in the CSD (v. 5.12) were included in the searches, irrespective of R-factor, etc., except that (i) if a compound occurred more than once in the database (duplicate structure determinations of the same crystal form or determinations of different polymorphs) only one ex-

ample was used (the one whose CSD reference code was last in an alphanumerically sorted listing); and (ii) structures with critical H-atoms missing (i.e. H-atoms belonging to one or the other of the interacting groups) were excluded. These criteria, especially the second, removed many structures from the analysis: of about 135 000 CSD entries with 3D coordinates, only 53 850 contributed to the library.

Searches for nonbonded contacts in protein–ligand crystal structures

Searches for protein-ligand contacts were based on a subset of PDB structures (June 1996 version) satisfying the following criteria: (i) crystallographic determinations only (no NMR structures used); (ii) resolution ≤ 2 Å; (iii) protein–ligand (i.e. protein–small-molecule) complexes only; a 'ligand' was defined as a peptide of up to 10 residues or a non-peptide molecule of at least nine atoms; (iv) DNA and RNA complexes excluded; (v) covalently bonded complexes excluded; and (vi) structures with the words 'mutant', 'mutation' or 'mutated' in their PDB files excluded (in order to avoid near-duplicates in the final data set). The final subset contained 328 entries.

Using the program SYBYL (v. 6.2) [28], the ligand was extracted from each of these entries, together with all protein residues, water molecules and other chemical entities (e.g. cofactors) within 4 Å of any ligand atom. A chemical connectivity (including, as far as possible, correct bond orders) was assigned to each ligand molecule (D.G. Watson, private communication; Ref. 29). However, unambiguous assignment of charges and bond orders was impossible for ligands containing ionisable or tautomeric groups, since hydrogen atoms are invariably absent from PDB structures. The resulting data were converted to QUEST-searchable format, using the program PREQUEST (v. 1.0) [30], and searches were performed for nonbonded contacts between (i) protein residues and ligand molecules; (ii) water and ligand molecules; and (iii) water and protein residues. Nonbonded contacts involving protein atoms only (e.g. an interaction between two neighbouring active-site residues) were not included, as these interactions have already been characterised by previous workers [10]. Nor were contacts between ligands, in cases where two or more small molecules occur in close proximity in a single binding site.

Presentation of contacts from crystal structures

The results of a search for contacts between two groups A and B were transformed into an easily visualised form by overlaying the A moieties. This results in a 3D distribution ('scatterplot') showing the experimental distribution of B (the 'contact group') around A (the 'central group') [31]. For simplicity, the overlaid A moieties were averaged and displayed as a single group. Separate scatterplots were produced from CSD and PDB data. 2D projections of typical scatterplots from the CSD and PDB are shown in Figs. 1a and b.

TABLE 1
SUMMARY OF CENTRAL GROUPS IN LIBRARY

Subsection	Type	No. of entries
Ligand section		
Terminal	C,H only (e.g. vinyl)	4
	N,C,H only (e.g. amino)	19
	O,C,H only (e.g. acetyl)	13
	N,O,C,H only (e.g. nitro)	10
	Si-containing (trimethylsilyl)	1
	P-containing (e.g. phosphato)	4
	S-containing (e.g. thiocyano)	13
	Halo-containing (e.g. bromo)	7
Links	N,C,H only (e.g. amine)	8
	O,C,H only (e.g. ester)	11
	N,O,C,H only (e.g. azoxy)	9
	P-containing (e.g. phosphinate)	4
	S-containing (e.g. sulphone)	18
Rings	Phenyls (e.g. 4-nitrophenyl)	30
	C,H only (e.g. indan)	5
	N,C,H only (e.g. pyrazole)	31
	O,C,H only (e.g. phthalidyl)	14
	N,O,C,H only (e.g. hydantoin)	26
	S-containing (e.g. thiophene)	8
	Nucleic acid bases (e.g. adenine)	7
Solvates	Inorganic (water)	1
	Organic (e.g. chloroform)	14
Protein section		
Terminal	E.g. hydroxy	10
Links	E.g. disulphide	2
Rings	E.g. indole	8

Different ionisation states, conformers, etc., counted separately in computing no. of entries.

Only the asymmetric unit of the scatterplot was computed for symmetrical central groups (e.g. all contacts were reflected into one quadrant for the carboxylate anion). When a central group was found to adopt two or more clearly distinct conformations, each was included in the library as a separate entry. Some central groups were found to have ill-defined conformations, e.g. acyclic disulphide linkages. In such cases, an iterative procedure was used to identify and eliminate outlying structures. An outlier was defined as any structure which did not overlay on the average geometry of the nonoutliers within the following criteria: *all atoms fit within 1 Å; rms deviation <0.5 Å.* Some central groups were found to adopt well-defined conformations except for a small number of hydrogen atoms (e.g. the methyl hydrogens of acetyl). In such cases, the rotationally mobile H-atoms were omitted entirely from the scatterplot display or were shown schematically.

A typical central group consisted of two types of atoms – target atoms and nontarget atoms. Nontarget atoms (e.g.

the carbon atom in the central group $C(sp^3)$-NH_2) were included to help define the chemical environment of the group but were not included in the nonbonded searches. For example, scatterplots for $C(sp^3)$-NH_2 only show contacts to the nitrogen and hydrogen atoms. In deriving scatterplots for ring systems, contacts to atoms, X, directly bonded to the ring were only found for the case X = H (i.e. contacts to ring substituents were ignored).

Theoretical energy calculations

Theoretical energy calculations were performed on 367 model systems. Each system consisted of two molecules that represent a particular interaction in the library. For example, the interaction between the central group $C(sp^3)$-$COC(sp^3)$ (i.e. aliphatic ketone) and the contact group NH(amide) was modelled by the complex between acetone and *N*-methylacetamide.

For each model system, the ab initio package CADPAC (v. 6.0) [32] was first used to calculate the wave functions of the individual monomers and to optimise their geometries at the 6-31G** basis-set level. The wave functions were corrected for electron correlation by the MP2 method [33]. A distributed multipole analysis (DMA) [34,35] was carried out on the monomer wave functions to obtain atomic multipoles up to hexadecapole. A rapid exploration of the potential-energy surface of the dimer was then performed with the program ORIENT (v. 3.2) [36]. This program can efficiently minimise the interaction energy of a given dimer starting-orientation, using a model intermolecular potential. The model potential describes the long-range electrostatic energy by using the DMA multipoles, including all terms of the multipole expansion up to and including r^{-5}. Repulsion and dispersion are described by an empirical, isotropic exp-6 term (FIT) [37], which provides a reasonable first estimate of the intermolecular separations but ignores any anisotropy in the atomic repulsive wall. The methodology is computationally inexpensive and adequate to give a reasonably accurate description of the intermolecular surface, particularly the orientation dependence.

Typically, starting with 100–500 initial orientations, the global and 5–20 local minima were obtained for any particular dimer from the ORIENT calculations. More accurate intermolecular interaction energies were then calculated for selected minima using the IMPT method, as programmed in CADPAC. The monomers (with 6-31G** optimised geometries) were placed in their ORIENT minimum-energy orientations and their IMPT interaction energy was computed at the 6-31G* basis-set level. Due to the limited quality of the empirical repulsion–dispersion parameters in ORIENT, it was necessary to optimise the distance between the model molecules at the IMPT level, maintaining fixed angular orientations from the ORIENT minimisation. The IMPT method is free of basis-set superposition error, a major problem with ab initio supermolecule calculations [38]. It provides an estimate for each of the major contributions to the total interaction energy (first-order electrostatic and exchange-repulsion energies; and the second-order energies due to polarisation, charge transfer and dispersion [39]). The energies of the resulting distance-optimised dimers (both total energy and the separate contributions) and the corresponding dimer orientations were stored in the library. Each minimum can be visualised in 3D and compared with the corresponding scatterplot(s).

Coverage of chemical groups

The library contains 277 central-group entries (Table 1). However, the number of chemically distinct groups is slightly lower (about 250) because some central groups give rise to more than one entry (e.g. if they exist in more than 1 distinct conformational state). The scatterplots for each central group show interactions between that group and up to 28 different contact groups (Table 2). Most contact groups represent chemical functions found in natural amino acids. The central groups cover a much

TABLE 2
CONTACT GROUPS IN LIBRARY

No.	Contact group	No.	Contact group
1	Any C, H, N, O or S atom	15	Water
2	Any NH, OH or SH	16	Imidazole (e.g. as in His)
3	Any NH	17	Guanidinium (e.g. as in Arg)
4	NH in peptides	18	$CONH_2$ (e.g. as in Asn, Gln)
5	Any NH^+	19	Any S
6	$-NH_3^+$ (e.g. as in Lys)	20	Aliphatic thioether (e.g. as in Met)
7	Any OH	21	Any $C(sp^3)$-H
8	Any alcoholic OH	22	Methyl
9	Any phenolic OH	23	Any C(ar)-H
10	$N(sp^2)$ acceptor (e.g. in pyridine)	24	Phenyl
11	O (not OH or H_2O, i.e. acceptor only)	25	Aromatic cationic nitrogen
12	Any carbonyl	26	Chloride
13	Carbonyl in peptides	27	Bromide
14	Carboxylate anion	28	Iodide

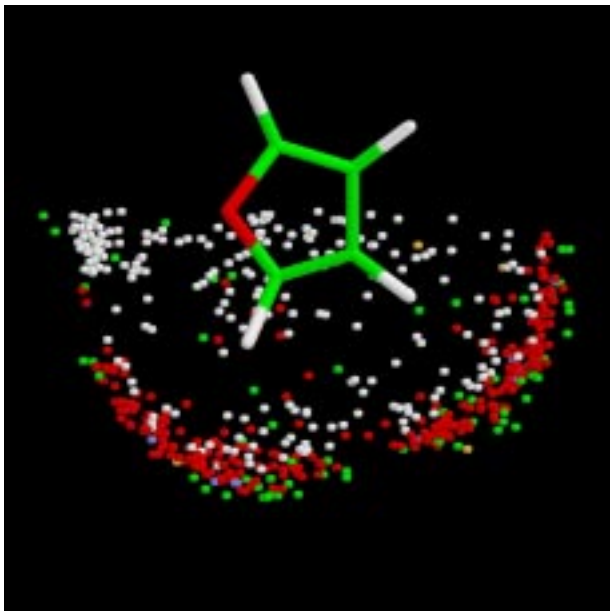


Fig. 4. CSD-based scatterplot of C (green), H (white), O (red), N (blue) and S (yellow) atoms around furan. The overwhelming majority of H-atoms on this plot are bonded to carbon, i.e. are nonpolar.

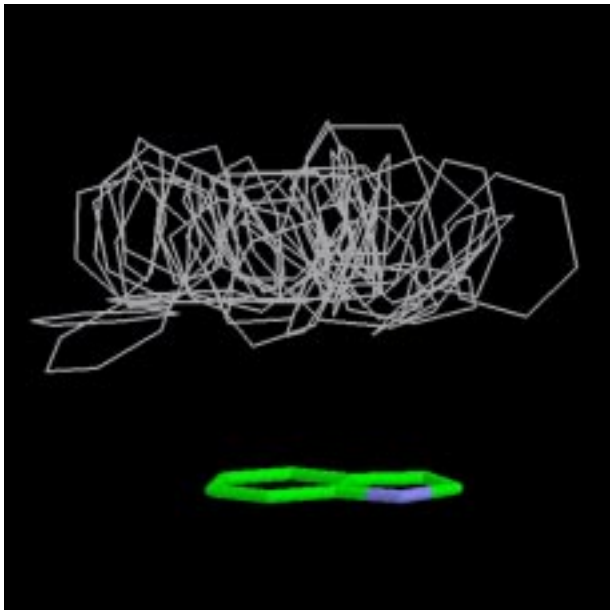
TABLE 3
EXAMPLE VALUES OF THE STATISTIC d_{rel}

Central group	Contact group	d_{rel} (esd)
Carboxylate	$[NH]^+$	8.5 (0.6)
Di-anionic phosphate	OH	5.9 (0.8)
Carboxylate	OH	3.0 (0.1)
Aliphatic hydroxyl	NH_3^+	2.8 (0.4)
Aliphatic–aliphatic ether	OH	2.6 (0.4)
Aliphatic hydroxyl	NH	1.3 (0.1)
Pyridine *N*-oxide	OH	1.3 (0.2)
Pyridine (uncharged)	OH	1.2 (0.1)
Aliphatic cyano	NH	1.0 (0.1)
Acetyl	OH	0.9 (0.1)
Aliphatic–aromatic ether	OH	0.9 (0.1)
Nitro	OH	0.7 (<0.1)
Indole	OH	0.6 (0.1)
Phenyl	OH	0.5 (<0.1)
Fluoro	(N,O,S)H	0.4 (0.1)
Pyridine (charged)	NH	0.1 (<0.1)

wider chemical range. Since each separate central-group–contact-group combination could have both a CSD- and a PDB-based scatterplot, the current library could in theory contain up to 15 512 ($= 277 \times 28 \times 2$) scatterplots. However, many of the theoretically possible plots (especially those based on PDB data) are absent through lack of experimental data. Also, at the time of writing, not all the database searches have been completed. In total, the current version of the library contains 6683 scatterplots: 5296 from the CSD and 1387 from the PDB. It also contains results for 867 theoretical potential-energy minima.

The ionisation states of acidic and basic groups cannot be determined unambiguously from protein crystal structures. In contrast, they are usually determined reliably in CSD structures. Accordingly, separate CSD-based scatterplots are present in the library for, e.g., carboxylate (CO_2^-) and carboxylic acid (CO_2H) but these groups are combined for PDB-based plots (Figs. 1a and b). A similar

a



b

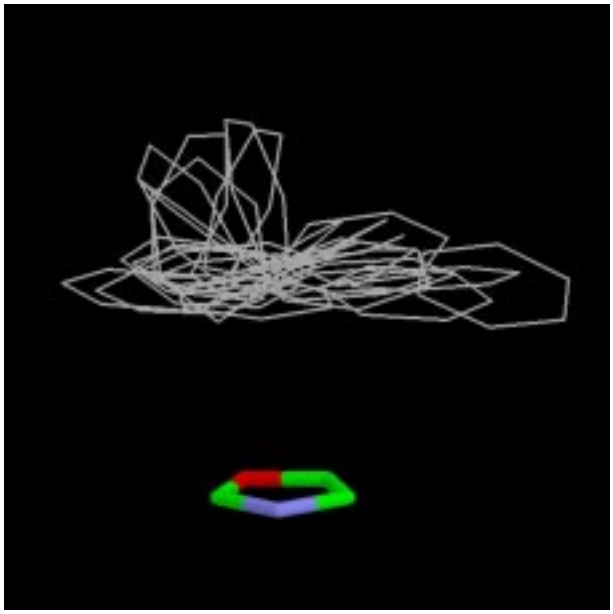


Fig. 5. CSD-based scatterplots of phenyl above (a) indole and (b) oxazole (contacts shorter than the sum of the van der Waals radii +0.5 Å). Phenyl rings around the edges of the indole and oxazole heterocycles are removed for clarity.

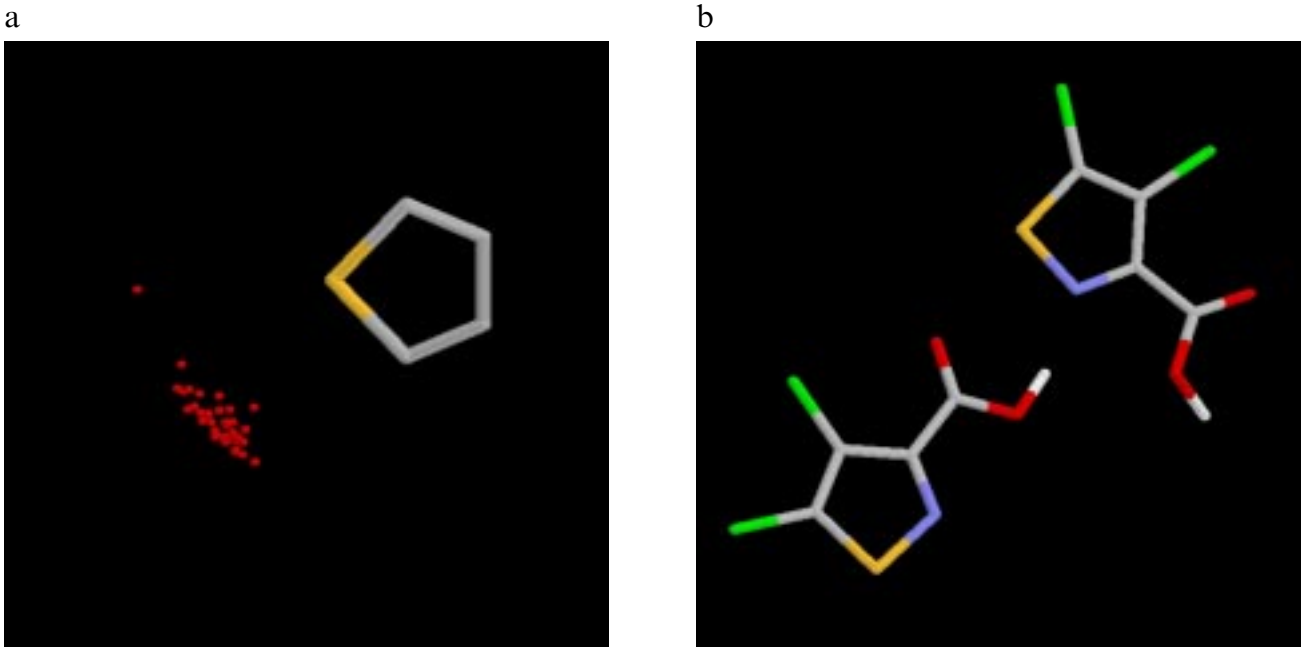


Fig. 6. (a) CSD-based scatterplot of oxygen atoms (excluding H-bond donor oxygens) around aromatic sulphur; (b) the CSD structure PORXAV.

situation is obtained for groups which differ only in the location of a tautomeric hydrogen, e.g. imidazol-4-yl and imidazol-5-yl.

Central groups are divided into two main sections: 'Ligand' and 'Protein'. The distinction is important for the PDB part of the library and is best explained with an example, the central group carbamoyl (-$CONH_2$). This group appears in both the Ligand and Protein sections. The plots in the Ligand section show contacts to carbamoyl groups on ligand molecules. Those in the Protein section show contacts to the carbamoyl groups of asparagine and glutamine side chains in proteins. Strictly speaking, all CSD-based scatterplots should be under the Ligand section, since the CSD contains only small molecules, not proteins. However, CSD data were included in the Protein section as well, so that comparisons can easily be made between nonbonded contacts to protein functional groups and contacts to similar functional groups in the higher precision, small-molecule CSD structures.

Software

The library is implemented under Unix. It is an Intranet application which can be front-ended by Web browsers that support the HTML (Hypertext Markup Language [40]) Frames facility, e.g. Netscape v. 3.0 [41]. Scatterplots, theoretical minima, etc., are selected from a menu and displayed in the molecular visualiser RasMol [42]. Interaction with RasMol displays is achieved by means of a Tcl/Tk [43] control program. Inter alia, this allows the user to measure distances, adjust display characteristics, and focus in on the shortest contacts on a scatterplot by use of a slider bar. Scatterplots are hyperlinked to the CSD or to the reformatted subset of the PDB described above. The user is therefore able to select any contact-group atom on a scatterplot and view, in RasMol, the relevant part of the crystal structure from which the contact was taken. Scatterplots are stored in a common modelling format (mol2 [28]) and can be converted to contoured density surfaces.

Results and Discussion

Hydrogen-bond geometries

An inspection of the scatterplots shows that the tendency for H-bonds to form along lone-pair directions varies greatly according to the nature of the acceptor [44,45]. At one extreme, some acceptor groups show pronounced lone-pair directionality. This is well illustrated by the PDB-based plot of water oxygens around aspartate and glutamate carboxylates (Fig. 1b), which has very distinct clusters in the sp^2 lone-pair directions. Strong lone-pair directionality is also found for many types of aromatic nitrogen, e.g. in pyridine and pyrimidine. This is true even for aromatic rings containing two adjacent nitrogen atoms, such as pyridazine or 1,2,4-thiadiazole. Here, the H-bonds invariably lie close to the nitrogen sp^2 lone pairs, even though it might perhaps have been anticipated that some would fall between the two nitrogens. In fact, the pyridazine plot (Fig. 2a) does show a tendency for the H-bonds to lie on the N2 side of the external C-N1-N2 bisector, but the preference is only slight. There are occasional differences in the directionality of H-bonds to different nitrogens in a single hetero-ring. For example, H-bonds to the N2 atom of 1,2,4-triazole appear to be

less directional (and longer) than those to N4 (Fig. 2b; albeit this observation is based on limited data).

Moderately strong geometrical preferences are exhibited by phosphate groups [46]. In both PDB- and CSD-based scatterplots, H-bonds to terminal phosphate oxygens occur in 'haloes', with P-O···H angles of about 145–150°. There are almost no H-bonds along the extension of the P-O bond itself. Discernible, if weak, directional preferences are seen for many other acceptors. For example, the CSD-based plot of NH and OH groups around aliphatic cyanide shows the highest density of H-bonds along the extension of the CN bond, i.e. in the direction of the nitrogen sp lone pair. The CSD-based plot of OH groups around aliphatic amide linkages shows weak clustering in approximately the carbonyl oxygen sp^2 lone-pair directions. This is revealed more clearly if the scatterplot (Fig. 3a) is converted to a density surface by contouring (Fig. 3b).

At the other extreme, some types of H-bonds show little or no lone-pair directionality. For example, and in contrast with Fig. 3b, the density surface for NH groups around aliphatic amide linkages shows a peak *between* the lone pairs of the carbonyl oxygen (Fig. 3c). H-bonds to aliphatic ether oxygens show a tendency to fall in the plane of the oxygen sp^3 lone pairs but no preference for the tetrahedral directions within that plane. This is consistent with earlier work based on more limited data [13,47].

Although the scatterplots reveal whether geometrical preferences exist, they cast no light on the underlying reasons. For example, a clustering of H-bonds along oxygen sp^2 lone-pair directions might reflect an inherent electrostatic preference for this geometry. Alternatively, it might arise from steric factors, since most sp^2-hybridised oxygens probably accept two hydrogen bonds, which will tend to occupy the positions of least steric interference [48,49]. Theoretical results in the library (e.g. on model systems such as methyl acetate–methanol) indicate that the potential-energy minima lie in approximately the sp^2 lone-pair directions even for isolated gas-phase dimers. Thus, there is an inherent preference for these directions, though it is weak. Possibly, the lone-pair directionality observed in the CSD and PDB is due to a combination of this weak inherent energetic preference together with the steric factors associated with crowding two donor molecules around each acceptor oxygen.

Hydrogen-bond frequencies and strengths

In addition to geometrical data, the library also contains information about the strengths of hydrogen bonds and the frequencies with which they occur in crystal structures. A good example is the CSD-based scatterplot showing the distribution of OH groups around acyclic aliphatic esters. This reveals that the carbonyl oxygen accepts H-bonds commonly but the bridging (C-O-C) oxygen almost never accepts. The same phenomenon occurs for lactones. Curiously, theoretical results in the library suggest that these two observations stem from different causes. In acyclic esters, the bridging oxygen is only a weak acceptor – much weaker, for example, than the oxygen of dimethyl ether (calculated H-bond energies: methyl acetate–methanol, −15.0 kJ/mol (for H-bond to bridging oxygen); dimethyl ether–methanol, −21.1 kJ/mol). This has been ascribed to the unfavourable dipolar interaction that occurs between the ester carbonyl group and any donor group approaching the bridging oxygen [50]. However, the strength of an H-bond to a lactone ring oxygen (calculated as −22.6 kJ/mol with methanol as the donor) is roughly comparable with that of an H-bond to an ether oxygen. Ether oxygens are quite common acceptors. Thus, on the basis of H-bond strengths alone, lactone ring oxygens should also be fairly frequent acceptors. That they are not is probably due to competition for donors from the nearby carbonyl oxygen, which is a stronger acceptor [50].

Scatterplots for isoxazole, oxazole and 1,2,4-oxadiazole indicate that the ring nitrogens of these heterocycles are much more common acceptors than the ring oxygens. Also, the oxygen of furan accepts H-bonds only rarely (i.e. H-bonds to conventional OH and NH donors; short CH···O interactions are sometimes found). The weakness of aromatic oxygen as an acceptor [51,52] is supported by theoretically calculated hydrogen-bond strengths (e.g. furan···methanol −14.4 kJ/mol; isoxazole (ring O)···methanol −17.5 kJ/mol; oxazole (ring O)···methanol −12.6 kJ/mol; cf. dimethyl ether···methanol −21.1 kJ/mol).

In molecular design, information that an atom cannot accept H-bonds is often as important as information about atoms that can. For example, the CSD-based scatterplot of NH and OH groups around thiazole reveals clearly that the sulphur atom of this heterocycle is not an H-bond acceptor. This is in marked contrast to some other types of sulphurs, e.g. in thioamides and thioureas, which are quite common acceptors, as pointed out by previous workers [53]. An inspection of the CSD-based scatterplots of OH groups around aromatic fluoro [54], chloro and bromo substituents reveals only a tiny number of O-H···Hal H-bonds.

Another indication of the strength of an interaction is its distance. Most of the CSD-based scatterplots in the library are accompanied by a statistic, d_{rel} (M.L. Verdonk, unpublished work), defined as

$$d_{rel} = d_{short} / d_{long}$$

where d_{short} is the density of contacts within the sum of van der Waals radii, V, and d_{long} is the density of contacts between V and V + 0.5 Å. The larger d_{rel} is, the greater the tendency for the contacts to be shorter than the sum of van der Waals radii and, by implication, the more attract-

ive they are likely to be. A standard deviation for d_{rel} can be estimated assuming Poisson statistics. Example values from the library are given in Table 3. Preliminary investigations on a range of uncharged donor–acceptor complexes show correlation coefficients of 0.8–0.9 between the logarithm of d_{rel} and calculated hydrogen-bond energies.

Contacts to hydrophobic groups

Contacts to hydrophobic groups are often assumed to be nondirectional, but the scatterplots in the library indicate that this is frequently not so. For example, both the CSD- and PDB-based plots of oxygen atoms (excluding water and hydroxyl oxygens) around phenyl show a strong preference for these contacts to occur around the edges of the phenyl ring, near the (C)H atoms [55,56]. A similar preference is seen in, e.g., the distribution of chloride ions around phenyl (CSD data). However, the opposite preference is seen in the distribution of positively charged aromatic nitrogens around phenyl: the cationic nitrogens tend to be situated above the ring. All of these observations reflect the quadrupole moment of benzene and emphasise the strong directional preferences of nonbonded contacts to this archetypal hydrophobic group [57]. Equally strong directionality is found for other common aromatic rings such as thiophene and furan. For both these systems, there is a preponderance of nonbonded contacts to oxygen atoms around the edges of the ring, except at the position of the hetero-atom. Here, and above and below the ring, contacts to CH hydrogens predominate (e.g. Fig. 4). Consequently, the 1-position of furan and thiophene is effectively more hydrophobic than positions 2 to 5. A similar effect is often seen in substituted phenyls. For example, the edge of the *p*-fluorophenyl group is clearly more hydrophobic in the vicinity of the F-atom than around the ring CH's, based on the distribution of contacts to O and CH.

Even contacts *between* hydrophobic groups can show directional preferences. Thus, CH contacts to electron-rich aromatic systems show rather different geometries from those to electron-deficient aromatics. A particularly clear-cut example is provided by the CSD-based scatterplots of phenyl groups around indole and oxazole (Fig. 5). The phenyl groups above indole tend to be oriented with the edge of the ring pointing towards the indole π cloud, implying T-shaped (edge-to-face) packing. This would be expected, as the H-atoms bonded to aromatic carbons usually carry net positive partial charges and the π-system of indole is particularly electron rich [58]. In contrast, the phenyl groups above oxazole tend to be parallel to the oxazole ring plane, implying face-to-face stacking.

The CSD-based scatterplot of phenyl around thiazole shows an interesting (albeit not statistically significant) feature. Phenyl rings near the thiazole nitrogen tend to be edge-on, with a phenyl CH pointing towards the electronegative nitrogen atom. This might be described as a weak C-H···N hydrogen bond. However, phenyl groups near the thiazole sulphur tend to be face-on. Presumably the sulphur atom is electropositive, being polarised by the ring nitrogen, and therefore prefers to interact with the phenyl π cloud rather than a CH hydrogen.

Unusual interactions

There are probably few nonbonded interactions in the library that are truly novel, i.e. previously unrecognised. However, there are many examples of intermolecular contacts that, while not novel, are unusual or comparatively little known. A good example occurs in the CSD-based scatterplot of oxygen atoms (excluding OH oxygens) around nitro groups. A pronounced cluster of oxygens above the nitro nitrogen is presumably indicative of an attractive N···O interaction that is also highly directional with respect to the nitrogen [59]. Similar tight clusters of oxygen atoms are seen in several other scatterplots, e.g. above the C2 carbonyl atom of uracil, along the extension of the C-I bond in iodo-substituted aromatics [60], and along the extension of the ring X-S bond (X = C,N) in sulphur-containing aromatic heterocycles (Fig. 6a) [7]. The attractive nature of halogen···oxygen and sulphur···oxygen contacts has been confirmed by published theoretical work [61–63].

Hyperlinking to the CSD or PDB often provides insight into the nature of unusual interactions and may suggest novel strategies for molecular design. For example, the CSD structure 4,5-dichloroisothiazole-3-carboxylic acid (CSD code PORXAV [64]), found by hyperlinking from the scatterplot shown in Fig. 6a, suggests a novel way of binding to an unionised carboxylic acid group: an isothiazole ring nitrogen accepts an H-bond from the carboxylic acid OH group and the adjacent ring sulphur forms a short S···O contact (2.88 Å) with the acid carbonyl oxygen (Fig. 6b). Interestingly, this arrangement occurs in preference to the more obvious carboxylic acid dimer motif.

In general, the CSD-based scatterplots are a better source of unusual interactions than the PDB-based plots, simply because they contain more data. However, the PDB-based plots have the advantage that they are more directly relevant to drug design. In the event that a PDB-based plot contains insufficient data to draw firm conclusions, the corresponding CSD-based plot may often be used to obtain supporting evidence. For example, the PDB-based plot of NH nitrogens around tryptophan indole suggests the occurrence of a few NH···π H-bonds, e.g. in a streptavidin–peptide complex (PDB code 1SLE [65]). This tentative conclusion is supported by the CSD-based NH/indole scatterplot, which provides the extra information that most NH···π H-bonds occur to the 6-membered ring of indole rather than the 5-ring. This result is consistent with calculations of electrostatic potentials [46].

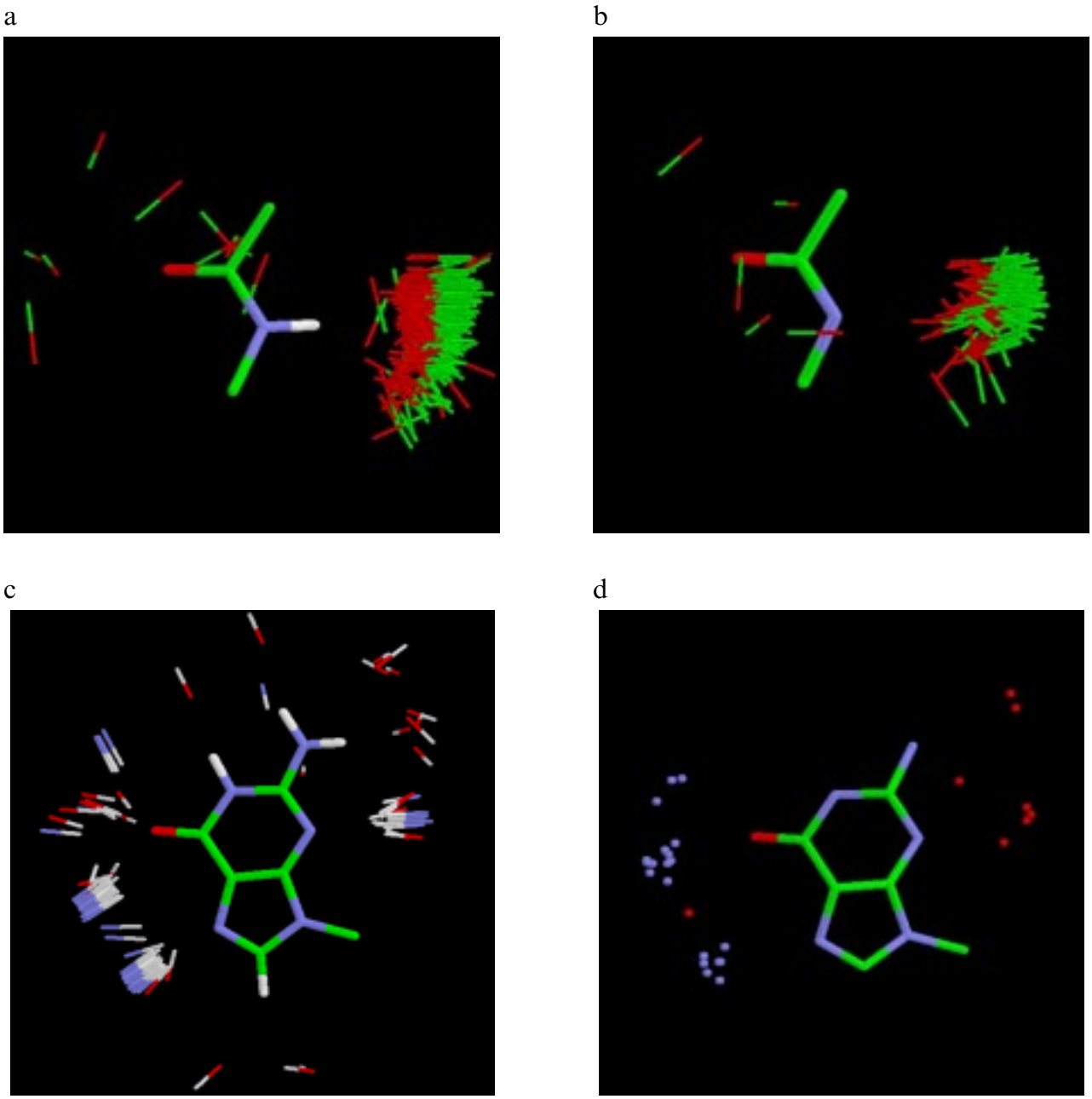


Fig. 7. (a) CSD- and (b) PDB-based scatterplots of carbonyl groups around amide (peptide) linkages; (c) CSD- and (d) PDB-based scatterplots of XH groups (X = N,O) around guanine (H-atoms not visible in the PDB plot).

Comparison of CSD and PDB data

There are relatively few interactions in the library for which sufficient data exist to allow a meaningful comparison of CSD- and PDB-based scatterplots. However, when comparisons can be made, there appears to be good agreement between the two types of plots. This is consistent with previous work based on more limited information [44]. An example has already been shown in Fig. 1 and two further examples are given in Fig. 7, which shows side-by-side comparisons of CSD- and PDB-based plots for peptide/C=O and guanine/XH (X = O,N). A detailed systematic comparison will be published in due course.

Limitations and difficulties of interpretation

(a) *CSD data:* Interactions in the CSD may sometimes be biased by systematic crystal-packing effects. In particular, this occurs when a group can form a strong H-bonded motif with itself. The best example is the cis-amide functionality in heterocycles such as γ-lactam. These frequently form the H-bonded motif shown in Fig. 8a (example taken from CSD structure WEVLOY, 6-chloro-1-cyano-2-

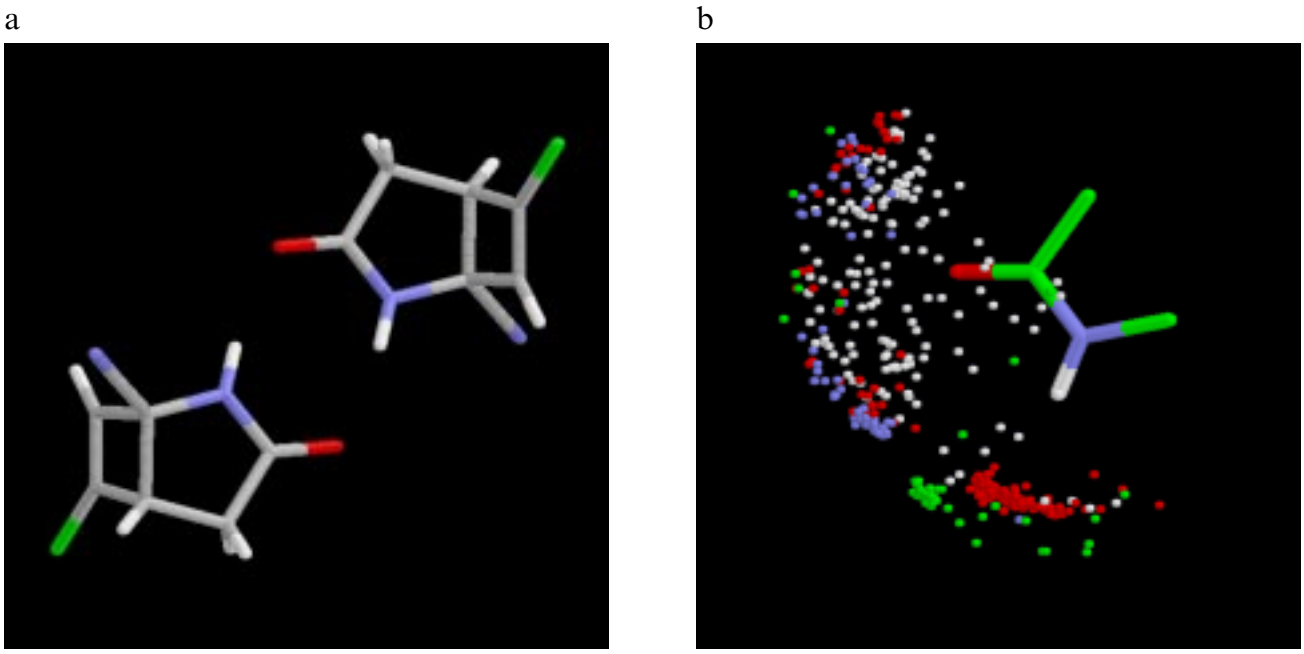


Fig. 8. (a) Common lactam crystal-packing motif in CSD (example taken from CSD structure WEVLOY); (b) CSD-based scatterplot of C,H,N,O,S atoms around γ-lactam (colour scheme as in Fig. 4; only part of the lactam ring is shown).

azabicyclo[3.2.0]hept-6-en-3-one [66]). This produces tight clusters of atoms on the corresponding CSD-based scatterplots, e.g. Fig. 8b. In protein–ligand complexes, the motif shown in Fig. 8a would be of little importance because cis-peptide linkages in proteins are rare (although it could occur to a terminal carbamoyl group of an asparagine or glutamine residue). CSD-based plots may also be misleading if there is a frequently occurring substitution pattern in molecules containing a particular group. For example, the scatterplot of carbonyl groups around aliphatic acetylene groups ($C(sp^3)$-CCH) shows a surprising cluster of oxygens around the tetrahedral carbon. This is due to the fact that many of the CSD structures contain a hydroxyl substituent on this carbon (i.e. C(OH)-CCH), which tends to form OH···O=C hydrogen bonds.

(b) *PDB data:* The absence of H-atoms from PDB-based plots often leads to difficulties in interpretation. For example, the PDB-based scatterplot of NH nitrogens around P-O-P linkages contains a cluster of N-atoms above the P-O-P plane and apparently donating H-bonds to the linking oxygen. The corresponding CSD-based plot, where the H-atoms of the NH groups are visible, clearly shows that most of the NH's in the vicinity of the linking oxygen are not H-bonded to this atom but to nearby terminal oxygens on the phosphorus atoms [46].

The most serious difficulties of interpretation occur with PDB-based plots pertaining to chemical groups that can exist in different ionisation and/or tautomeric states. The imidazole moiety of histidine is illustrative. There is a tautomeric ambiguity (4-yl or 5-yl), an uncertainty in ionisation state (neutral or cationic) and, at poor resolution, the crystallographer may have had difficulty in distinguishing ring nitrogens from ring carbons. This makes PDB-based plots for this ring particularly difficult to interpret. For example, in the plot of OH around imidazole (from histidine side chains), it is not clear which OH groups are donating to ring nitrogens and which are accepting. There are also some short contacts on the plot between hydroxyl oxygens and the imidazole C2 atom. These may be CH···O hydrogen bonds or may be errors resulting from the imidazole ring being fitted to the electron density the wrong way round in the X-ray analysis.

A further problem with PDB-based plots is the existence of closely related protein–ligand complexes in the PDB. They may cause misleading clusters of contact groups on PDB-based plots. For example, the plot showing amide carbonyls around arginine guanidiniums has a tight cluster of contacts lying between the two arginine NH_2 groups. On hyperlinking, it is found that most of these come from a series of closely related thermolysin complexes.

(c) *Theoretical data:* The IMPT calculations are usually performed on the simplest possible model compounds. Consequently, steric effects are likely to be less important in these model systems than in molecules from the CSD and PDB, which are often more highly substituted. For example, the CSD-based scatterplot of OH groups around aliphatic esters shows that most H-bonds form to the carbonyl-oxygen lone pair which is anti to the bridging ester oxygen. However, IMPT calculations indicate that, if anything, the strongest H-bond between methanol and methyl acetate is formed to the other lone pair (calculated

energies: –26.0 kJ/mol to the syn lone pair; –24.7 kJ/mol to the anti). Presumably, many of the esters in the CSD have groups larger than methyl on the ester oxygen, causing increased steric repulsion when H-bonds are formed to the syn lone pair.

As with all molecular orbital data, the accuracy of the results in the library is limited by the basis set used and the methodology employed. However, comparisons between the method used in this work and extended basis-set (triple zeta plus polarisation) supermolecule calculations showed an encouraging degree of concordance (S. Brode and J.P.M. Lommerse, unpublished work). There is, of course, no way of ensuring that all significant minima in a given potential-energy surface are found. Indeed, given the number of systems investigated, it is highly likely that several were missed in the library as a whole.

By far the biggest limitation in the theoretical calculations is that they model an in vacuo situation. They are therefore unsuitable for indicating the gain or loss of free energy that occurs when a group is taken out of solution to form a nonbonded contact to a protein. For example, no account whatsoever is taken of entropic factors. Also, weakly attractive interactions that occur in vacuo are much less likely to be found in condensed phases, where competing interactions are always present. For example, the calculated minimum-energy orientation of fluorobenzene–*N*-methylacetamide involves an N-H···F hydrogen bond. Such an interaction is extremely unlikely to occur in condensed phases, where better hydrogen-bond acceptors will successfully compete for the NH proton [54]. Also, bifurcated interactions are frequently found in the theoretical calculations but are much less likely to occur in condensed phases.

Conclusions

The computerised library of crystallographic and theoretical data described above constitutes a detailed source of information on nonbonded interactions. Results are presented in an intuitive form which allows rapid visual assessment of the types of interactions formed by a group and relatively easy comparison of one group with another. The data can be read in a straightforward manner by external programs. The library should be useful for a variety of purposes, e.g. (i) searching for bioisosteric replacements; (ii) protein–ligand docking and design; (iii) fitting of experimental electron density in protein–ligand crystal-structure analysis; (iv) design of novel molecular recognition systems; (v) 3D QSAR; (vi) design and prediction of crystal structures; and (vii) parametrisation of ligand-design and other molecular-modelling programs.

Results from the library reveal that different types of hydrogen bonds have very different geometrical preferences, ranging from strong lone-pair directionality to almost no lone-pair directionality. The ability of some elements to accept H-bonds depends strongly on their intramolecular environment. In the case of a polarisable element such as sulphur, this is to be expected. More surprisingly, however, some types of oxygen atoms appear to be almost incapable of accepting hydrogen bonds, e.g. the bridging oxygen of esters and the ring oxygen of lactones and aromatic heterocycles such as furan. This is not always because the oxygen atom is an intrinsically weak acceptor. Sometimes, it is due to competition from alternative, stronger interactions (e.g. in lactones). This emphasises that a nonbonded interaction may be attractive and yet still not occur frequently in nature, as pointed out by earlier workers [67].

Other results illustrate the surprisingly strong geometrical preferences of contacts to many hydrophobic systems, including not only phenyl but also aromatic heterocycles such as furan, indole and oxazole. Knowledge of these directional preferences should improve our ability to design and predict the binding orientations of hydrophobic protein ligands. The library contains many examples of intermolecular contacts that, while not necessarily novel, are unusual and comparatively little known, e.g. I···O, S···O and nitro···O. Exploitation of these may increase novelty in drug design.

It is our intention to keep the library updated as the CSD and PDB grow. In addition, we envisage incorporating some entirely new types of data, such as metal-coordination data [68], common crystal-structure motifs (G.P. Shields, F.H. Allen, P.R. Raithby and R. Taylor, unpublished work), and torsion-angle distributions of acyclic bonds commonly found in small organic molecules [69]. Also, we are developing modelling programs that use the library as a ‘knowledge base’. For example, in collaboration with others we are attempting to couple the library directly to an existing program [70] for flexible ligand docking. Given a particular ligand and protein binding site, the program will look up the most appropriate scatterplots in the library and use them to guide docking. An intriguing possibility is to use contoured density surfaces such as those shown in Figs. 3b and c to produce fields on which a 3D QSAR analysis might be based, using the CoMFA technique [71].

Acknowledgements

We are grateful for software and/or advice from Roger Amos and Anthony Stone (University of Cambridge), Sally Price and Irene Nobeli (University College, London), Roger Sayle (Glaxo Wellcome), Stefan Brode (BASF-AG), Frank Allen and David Watson (Cambridge Crystallographic Data Centre), Manfred Hendlich (Merck KGaA), and Ashley Fenwick and Jens Loesel (SmithKline Beecham). Special thanks are due to Gerhard Klebe (formerly BASF-AG, now University of Marburg) for advice on the project from its earliest stages.

References

1 Böhm, H.-J. and Klebe, G., Angew. Chem., Int. Ed. Engl., 35 (1996) 2588.
2 Lehn, J.-M., Supramolecular Chemistry: Concepts and Perspectives, VCH, Weinheim, Germany, 1995.
3 Desiraju, G.R., Crystal Engineering: The Design of Organic Solids, Elsevier, Amsterdam, The Netherlands, 1991.
4 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. and Watson, D.G., J. Chem. Inf. Comput. Sci., 31 (1991) 187.
5 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
6 Bürgi, H.-B., Dunitz, J.D. and Shefter, E., J. Am. Chem. Soc., 95 (1973) 5065.
7 Rosenfield Jr., R.E., Parthasarathy, R. and Dunitz, J.D., J. Am. Chem. Soc., 99 (1977) 4860.
8 Thomas, K.A., Smith, G.M., Thomas, T.B. and Feldmann, R.J., Proc. Natl. Acad. Sci. USA, 79 (1982) 4843.
9 Singh, J., Thornton, J.M., Snarey, M. and Campbell, S.F., FEBS Lett., 224 (1987) 161.
10 Singh, J. and Thornton, J.M., Atlas of Protein Side-Chain Interactions, Oxford University Press, Oxford, U.K., 1992.
11 Tintelnot, M. and Andrews, P., J. Comput.-Aided Mol. Design, 3 (1989) 67.
12 Klebe, G., Habilitationsschrift, Heidelberg, Germany, 1990.
13 Kroon, J., Kanters, J.A., Van Duijneveldt-Van de Rijdt, J.G.C.M., Van Duijneveldt, F.B. and Vliegenthart, J.A., J. Mol. Struct., 24 (1975) 109.
14 Murray-Rust, P. and Glusker, J.P., J. Am. Chem. Soc., 106 (1984) 1018.
15 Taylor, R. and Kennard, O., Acc. Chem. Res., 17 (1984) 320.
16 Jeffrey, G.A. and Mitra, J., Acta Crystallogr., B39 (1983) 469.
17 Gavezzotti, A., J. Am. Chem. Soc., 111 (1989) 1835.
18 Etter, M.C., Acc. Chem. Res., 23 (1990) 120.
19 Murray-Rust, P. and Motherwell, W.D.S., J. Am. Chem. Soc., 101 (1979) 4374.
20 Taylor, R. and Kennard, O., J. Am. Chem. Soc., 104 (1982) 5063.
21 Sarma, J.A.R.P. and Desiraju, G.R., Acc. Chem. Res., 19 (1986) 222.
22 Bürgi, H.-B. and Dunitz, J.D. (Eds.) Structure Correlation, Vols. 1 and 2, VCH, Weinheim, Germany, 1994.
23 Hayes, I.C. and Stone, A.J., Mol. Phys., 53 (1984) 83.
24 CSD User Manual, Cambridge Crystallographic Data Centre, Cambridge, U.K., 1992.
25 Bondi, A., J. Phys. Chem., 68 (1964) 441.
26 Jeffrey, G.A. and Lewis, L., Carbohydr. Res., 60 (1978) 179.
27 Taylor, R. and Kennard, O., Acta Crystallogr., B39 (1983) 133.
28 SYBYL, v. 6.2, Tripos Associates, St. Louis, MO, U.S.A., 1995.
29 Van Aalten, D.M.F., Bywater, R., Findlay, J.B.C., Hendlich, M., Hooft, R.W.W. and Vriend, G., J. Comput.-Aided Mol. Design, 10 (1996) 255.
30 Prequest User Guide, Cambridge Crystallographic Data Centre, Cambridge, U.K., 1996.
31 Rosenfield Jr., R.E., Swanson, S.M., Meyer Jr., E.F., Carrell, H.L. and Murray-Rust, P., J. Mol. Graph., 2 (1984) 43.
32 Amos, R.D., CADPAC 6.0: The Cambridge Analytical Derivatives Package, Issue 6.0, University of Cambridge, Cambridge, U.K., 1996.
33 Møller, C. and Plesset, M.S., Phys. Rev., 46 (1934) 618.
34 Stone, A.J., Chem. Phys. Lett., 83 (1981) 233.
35 Stone, A.J. and Alderton, M., Mol. Phys., 56 (1985) 1047.
36 Stone, A.J., ORIENT, v. 3.2, with contributions from Dullweber, A., Hodges, M.P., Popelier, P.L.A. and Wales, D.J., University of Cambridge, Cambridge, U.K., 1996.
37 Coombes, D.S., Price, S.L., Willock, D.J. and Leslie, M., J. Phys. Chem., 100 (1996) 7352.
38 Stone, A.J., Chem. Phys. Lett., 211 (1993) 101.
39 Stone, A.J., The Theory of Intermolecular Forces, Clarendon Press, Oxford, U.K., 1996, pp. 79–104.
40 Berners-Lee, T.J., Hypertext Markup Language, CERN, Geneva, Switzerland, 1993.
41 Netscape, v. 3.0, Netscape Communications Corporation, Mountain View, CA, U.S.A.
42 Sayle, R.A., Glaxo Wellcome, Stevenage, U.K., 1996.
43 Ousterhout, J.K., Tcl and the Tk Toolkit, Addison-Wesley, Reading, MA, U.S.A., 1994.
44 Klebe, G., J. Mol. Biol., 237 (1994) 212.
45 Mills, J.E.J. and Dean, P.M., J. Comput.-Aided Mol. Design, 10 (1996) 607.
46 Lommerse, J.P.M. and Taylor, R., J. Enzyme Inhib., 11 (1997) 223.
47 Ceccarelli, C., Jeffrey, G.A. and Taylor, R., J. Mol. Struct., 70 (1981) 255.
48 Taylor, R., Kennard, O. and Versichel, W., J. Am. Chem. Soc., 105 (1983) 5761.
49 Mitchell, J.B.O. and Price, S.L., Chem. Phys. Lett., 154 (1989) 267.
50 Lommerse, J.P.M., Price, S.L. and Taylor, R., J. Comput. Chem., 18 (1997) 757.
51 Böhm, H.-J., Brode, S., Hesse, U. and Klebe, G., Chem. Eur. J., 2 (1996) 1509.
52 Nobeli, I., Price, S.L., Lommerse, J.P.M. and Taylor, R., J. Comput. Chem., in press.
53 Allen, F.H., Bird, C.M., Rowland, R.S. and Raithby, P.R., Acta Crystallogr., B53 (1997) 680.
54 Dunitz, J.D. and Taylor, R., Chem. Eur. J., 3 (1997) 89.
55 Rowland, R.S., Allen, F.H., Carson, W.M. and Bugg, C.E., In Bugg, C.E. and Ealick, S.E. (Eds.) Crystallographic and Modeling Methods in Molecular Design, Springer, New York, NY, U.S.A., 1990, pp. 229–253.
56 Flanagan, K., Walshaw, J., Price, S.L. and Goodfellow, J.M., Protein Eng., 8 (1995) 109.
57 Dougherty, D.A., Science, 271 (1996) 163.
58 Pullman, B. and Pullman, A., Proc. Natl. Acad. Sci. USA, 44 (1958) 1197.
59 Taylor, R., Mullaley, A. and Mullier, G.W., Pesticide Sci., 29 (1990) 197.
60 Cody, V. and Murray-Rust, P., J. Mol. Struct., 112 (1984) 189.
61 Lommerse, J.P.M., Stone, A.J., Taylor, R. and Allen, F.H., J. Am. Chem. Soc., 118 (1996) 3108.
62 Burling, F.T. and Goldstein, B.M., J. Am. Chem. Soc., 114 (1992) 2313.
63 Baalham, C.A., M.Phil. Thesis, University of Cambridge, Cambridge, U.K., 1996.
64 Verenich, A.I., Govorova, A.A., Galitskii, N.M., Potkin, V.I., Kaberdin, R.V. and Ol'dekop, Yu.A., Khim. Get. Soedin, SSSR, 1992 399.
65 Katz, B.A., Biochemistry, 34 (1995) 15421.
66 Lamara, K., Redhouse, A.D., Smalley, R.K. and Thompson, J.R., Tetrahedron, 50 (1994) 5515.
67 Mitchell, J.B.O., Nandi, C.L., McDonald, I.K., Thornton, J.M. and Price, S.L., J. Mol. Biol., 239 (1994) 315.
68 Glusker, J.P., Adv. Protein Chem., 42 (1991) 1.
69 Klebe, G. and Mietzner, T., J. Comput.-Aided Mol. Design, 8 (1994) 583.
70 Jones, G., Willett, P., Glen, R.C., Leach, A.R. and Taylor, R., J. Mol. Biol., 267 (1997) 727.
71 Cramer III, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.