Journal of Computer-Aided Molecular Design, 10 (1996) 607–622
ESCOM

J-CAMD 361

Three-dimensional hydrogen-bond geometry and probability information from a crystal survey

J.E.J. Mills* and P.M. Dean

Drug Design Group, Department of Pharmacology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QJ, U.K.

Received 3 May 1996
Accepted 26 June 1996

Keywords: Ligand–receptor interaction; Geometry of intermolecular interactions; Cambridge Structural Database; Drug design

Summary

An extensive crystal survey of the Cambridge Structural Database has been carried out to provide hydrogen-bond data for use in drug-design strategies. Previous crystal surveys have generated 1D frequency distributions of hydrogen-bond distances and angles, which are not sufficient to model the hydrogen bond as a ligand–receptor interaction. For each hydrogen-bonding group of interest to the drug designer, geometric hydrogen-bond criteria have been derived. The 3D distribution of complementary atoms about each hydrogen-bonding group has been ascertained by dividing the space about each group into bins of equal volume and counting the number of observed hydrogen-bonding contacts in each bin. Finally, the propensity of each group to form a hydrogen bond has been calculated. Together, these data can be used to predict the potential site points with which a ligand could interact and therefore could be used in molecular-similarity studies, pharmacophore query searching of databases, or de novo design algorithms.

Introduction

Hydrogen bonds have a crucial rôle in the structure and actions of many bioactive molecules. The hydrogen bond combines the strength of an electrostatic interaction with the directionality of a dipole interaction to produce the most specific and predictable of all biomolecular interactions. In drug design, the coordinates of the hydrogen-bonding atoms in the ligand-binding site dictate the positions at which it would be advantageous to place complementary atoms of a novel drug molecule [1–4]. These atom positions can be used as constraints in structure-generation algorithms that design novel lead compounds [5–7]. Conversely, by analysing the structure of a ligand, it is possible to predict the potential positions of hydrogen-bonding atoms in the binding site (site points). Molecular-similarity studies can use this information to superpose dissimilar ligands on the assumption that the ligands bind to common site points [8,9]. This circumvents the need to superpose ligands on the basis of their common hydrogen-bonding atom positions. The common site points derived can be used to construct a model of the binding site and therefore can be employed as a basis for de novo design algorithms.

Information concerning the prediction of hydrogen bonding can be gleaned from three sources: solution studies, quantum mechanics calculations and crystal surveys. In general, solution studies only give an idea of the propensity of a chemical group to form a hydrogen bond in the form of solute scales [10], although a limited amount of directionality information is available [11]. Quantum mechanics calculations provide much more detailed information about the preferred geometries of hydrogen bonds. However, their use is restricted to small molecules owing to the large computational resources required; other types of model must be used to predict the hydrogen-bonding properties of large and even medium-sized drug molecules [12]. A statistical examination of the hydrogen bonds occurring in crystals provides an unambiguous method for deriving information about hydrogen-bond directionality and strength. Unfortunately, there are not yet enough high-resolution structures of ligand–protein complexes in

*To whom correspondence should be addressed.

TABLE 1
ψ-ϕ HYDROGEN-BONDING GROUPS

Twofold ϕ-symmetry[a]	No ϕ-symmetry[a]
Symmetric C=O (e.g. ketone, urea)	Asymmetric C=O (e.g. ester, amide)
ROR	RCO_2^-
HOH	R_2SO_2
RNR	$R_2PO_2^-$
C=S	NO_2
Symmetric Het_6N	Asymmetric Het_6N
Symmetric Het_5N	Asymmetric Het_5N
Het_5O	

[a] Het_nX represents an sp^2-hybridized X atom in an n-membered ring.

the Brookhaven Protein Databank (PDB) [13] to allow a thorough survey of ligand–protein hydrogen bonds [14]. However, the Cambridge Structural Database (CSD) [15] contains the crystal structures of over 140 000 small molecules; the associated software provides a means by which hydrogen-bond crystal surveys can be rapidly carried out [16–21]. The only assumption made in using data from such surveys to model drug–receptor interactions is that the small-molecule crystal environment does not differ greatly from the ligand–receptor interface, which Klebe [21] asserted to be true. Nevertheless, he found that hydrogen bonds in complexes occur close to the range limits defining hydrogen bonds in small molecules. This could be a consequence of the lower precision inherent in obtaining protein structures from crystals, or there could be a difference between the two environments. Work is currently in progress to ascertain and investigate any differences between the properties of hydrogen bonds in the small-molecule and protein situations.

This paper describes a new crystal survey that extracts valuable information for the drug designer: the 3D distribution of hydrogen bonds about each hydrogen-bonding group and the propensity of each group to form a hydrogen bond (i.e. group-probability values). This allows the prediction of the likely sites of hydrogen-bond formation about any ligand, thereby giving an estimate of the potential site points with which a ligand could interact.

This survey differs from previous surveys in three important aspects. First, the output data pertain to the 3D distribution of hydrogen bonds about each group. Published surveys have usually provided distributions only for individual geometric parameters. If these data are to be used to model the 3D distribution of hydrogen bonds, these parameters must be independent. This has been shown not to be the case for the hydroxyl group [18], implying that 3D data are required to model the hydrogen bond accurately. In this study, the independence of the parameters is tested for other hydrogen-bonding groups. Secondly, the criteria used to define hydrogen bonds in previous surveys have often been arbitrary and certainly not consistent between different surveys. In this survey, the hydrogen-bond data obtained are used to derive the hydrogen-bond criteria for each group. Thirdly, group-probability values are calculated for each hydrogen-bonding group. Such values have not been previously derived from a crystal survey of small molecules.

The crystal survey was carried out in two stages. In the first stage, all intermolecular contacts satisfying very lenient hydrogen-bond criteria were extracted from the CSD. In the second stage, these contacts were filtered using statistical tests such that only precisely defined hydrogen bonds remained. From these results, information about the 3D distribution of hydrogen bonds was obtained, along with the group-probability values. These latter results have been compared with other indices of hydrogen-bonding potential.

TABLE 2
θ-τ HYDROGEN-BONDING GROUPS

Sixfold τ-symmetry		Fourfold τ-symmetry		Twofold τ-symmetry	
Group[a]	D/A[b]	Group[a]	D/A[b]	Group[a]	D/A[b]
R-OH	D+A	Ph-OH	D+A	CO-OH	D
$R\text{-}NH_2$	D+A	$Ph\text{-}NH_2$	D+A	$CO\text{-}NH_2$	D
$R\text{-}N^+H_3$	D	$R_2\text{-}N^+H$	D	CON-H	D
$R_3N^+\text{-}H$	D	$R\text{-}N^+H_2$	D	PhN-H	D
$R_3P{=}O$	A	$Het_5N\text{-}H$	D	$R_2(H)N^+\text{-}H$	D
$R_3N\text{-}lp$	A			R=N-H	D
				$R_2N\text{-}H$	D
				HO-H	D
				$O_2(R)P\text{-}O$	A
				$O_2(R)S\text{-}O$	A
				$R_2S{=}O$	A
				$R_2(H)N\text{-}lp$	A
				RC≡N	A

[a] Het_5N represents a trigonal planar N atom in a five-membered ring. R represents any organic group, Ph a phenyl ring and lp a lone pair. For each group, the last bond listed is the axis about which θ (or υ) is measured.
[b] D denotes a donor group and A an acceptor group.

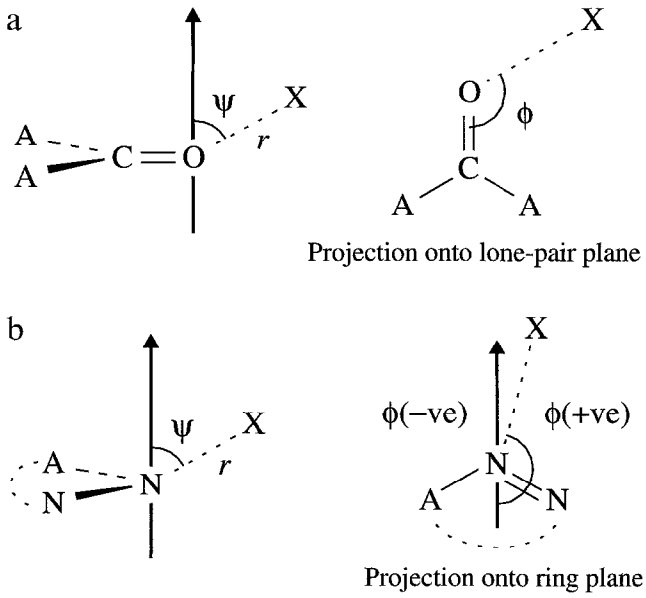


Fig. 1. Two examples of ψ-ϕ groups. X represents a heavy donor atom and A is an arbitrary atom. (a) Symmetric carbonyl group; (b) asymmetric heterocyclic aromatic nitrogen group.

Hydrogen-bonding groups

In this study, a donor group is defined to be an electronegative atom bonded to a hydrogen atom. The electronegative atom is either a nitrogen or oxygen atom. Despite the fact that the C-H group has been shown to form hydrogen bonds in crystal structures [22,23], it is not considered as a hydrogen-bonding group in this study for two reasons. First, the proportion of C-H groups present that form hydrogen bonds is small. C-H hydrogen bonds generally only occur in crystals when few, if any, other hydrogen-bonding donor groups are present. Hydrogen bonds in crystals have been proposed to form in a hierarchy, the strongest hydrogen bonds forming first [24]. C-H hydrogen bonds are weak, so if they do occur, their geometries are often distorted to fit in with the optimal geometry of the hydrogen bonds that form in preference. Secondly, in this study, the crystal data are used to generate hydrogen-bond probability maps for ligands. These maps are superposed to predict the hydrogen-bonding site points common to a series of ligands (see the Discussion section). Most drug molecules contain many C-H groups, so false positive results would occur if they were to be included in the generation and superposition of the maps. C-H groups would dominate the description of the hydrogen-bonding nature of the ligands despite the fact that they are probably the least important of the hydrogen-bonding groups.

Acceptor groups are defined by the presence of a lone pair on an electronegative atom, either a nitrogen, oxygen or sp^2-sulphur atom. The nitrogen atoms belonging to amide groups and attached to aromatic rings (with the exception of the aniline NH_2 group) are not considered as hydrogen-bond acceptors because their lone pairs are known to conjugate with the carbonyl group and aromatic ring, respectively.

After preliminary studies, hydrogen bonds involving sp^3-sulphur atoms and halogens were not considered for two reasons. First, there were very few hydrogen bonds to any of the groups, certainly not enough to derive representative 3D distributions. Secondly, the propensity of each of the groups to form hydrogen bonds was much lower than for any of the other groups surveyed. Both types of group have been previously shown to form very weak and nondirectional hydrogen bonds [25,26].

All of the hydrogen-bonding groups surveyed are shown in Tables 1 and 2. For the purposes of this study, hydrogen-bonding groups are divided into two classes on the basis of the parameters used to describe the geometries of the bonds they form. For the first class (ψ-ϕ groups; Table 1), the spherical coordinate parameters r, ψ and ϕ, as originally used by Taylor et al. [16], describe the position of a complementary hydrogen-bonding atom (see Fig. 1). This requires a plane from which to measure the

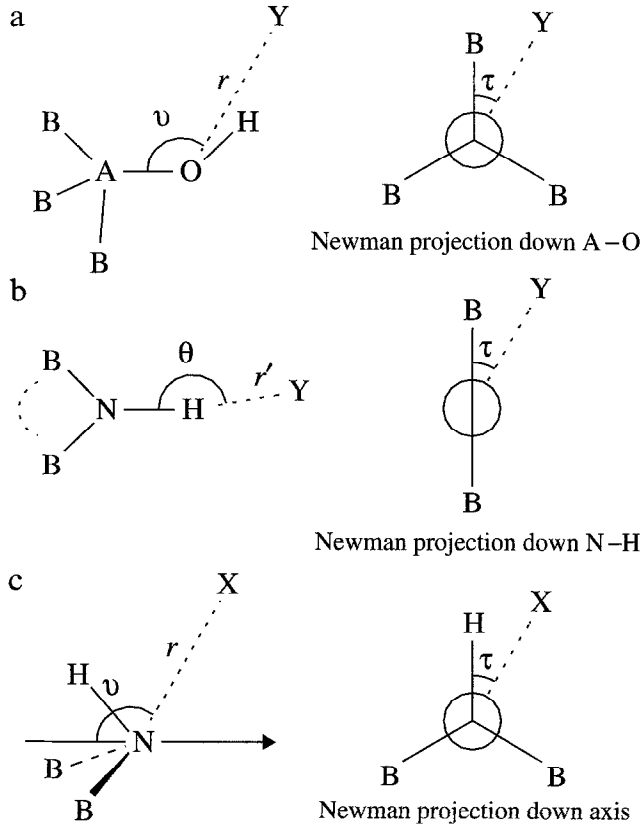
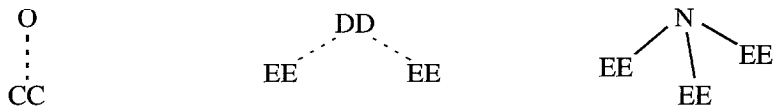


Fig. 2. Three examples of θ-τ groups. X represents a heavy donor atom and Y a heavy acceptor atom. A and B are arbitrary atoms. (a) Rotatable hydroxyl group, for which there is sixfold symmetry in τ-space; (b) heterocyclic nitrogen donor group, for which there is fourfold symmetry in τ-space; (c) secondary amine group, for which there is twofold symmetry in τ-space. In this case, υ is measured using an axis that passes through the nitrogen atom and the predicted lone-pair direction, i.e. the negative bisector of the three bonds to the nitrogen atom.



carbonyl
phosphinyl
phosphate
sulphone

ether
water
heterocycle N
hydroxyl
heterocycle O

amines

Fig. 3. The query structures for acceptor groups used to ascertain the presence of a hydrogen bond. A dotted line signifies a bond that can be single, double or aromatic. CC represents the elements C, S or P. DD represents the elements N, O or S. EE represents the elements C, N, O, S, P or H.

angle ψ and an axis within this plane from which to measure ϕ. The hydrogen bonds of the second class of group (θ-τ groups; Table 2) are described by parameters analogous to the standard r and θ parameters used in previous surveys to describe the geometry about donor groups. However, θ is usually associated with an X-H···Y angle, so in this work, if the angle is not subtended at a hydrogen atom, it is termed υ. In addition, there is a third parameter, τ, measuring the degree of torsion about the axis from which θ or υ is measured; there is either six-, four- or twofold symmetry in τ-space, as shown in Fig. 2. For the majority of hydrogen-bonding groups, distances are measured between the heavy (i.e. non-hydrogen) atoms. However, when the position of the donor hydrogen atom is known because the group is rigid (e.g. secondary amide), the distance is measured from the hydrogen atom. When the distance is measured between the hydrogen and acceptor atoms, the parameter is termed r'.

Methods

Searching the CSD

The first stage of the search identifies close intermolecular contacts, with r values of less than 4.5 Å and θ (and

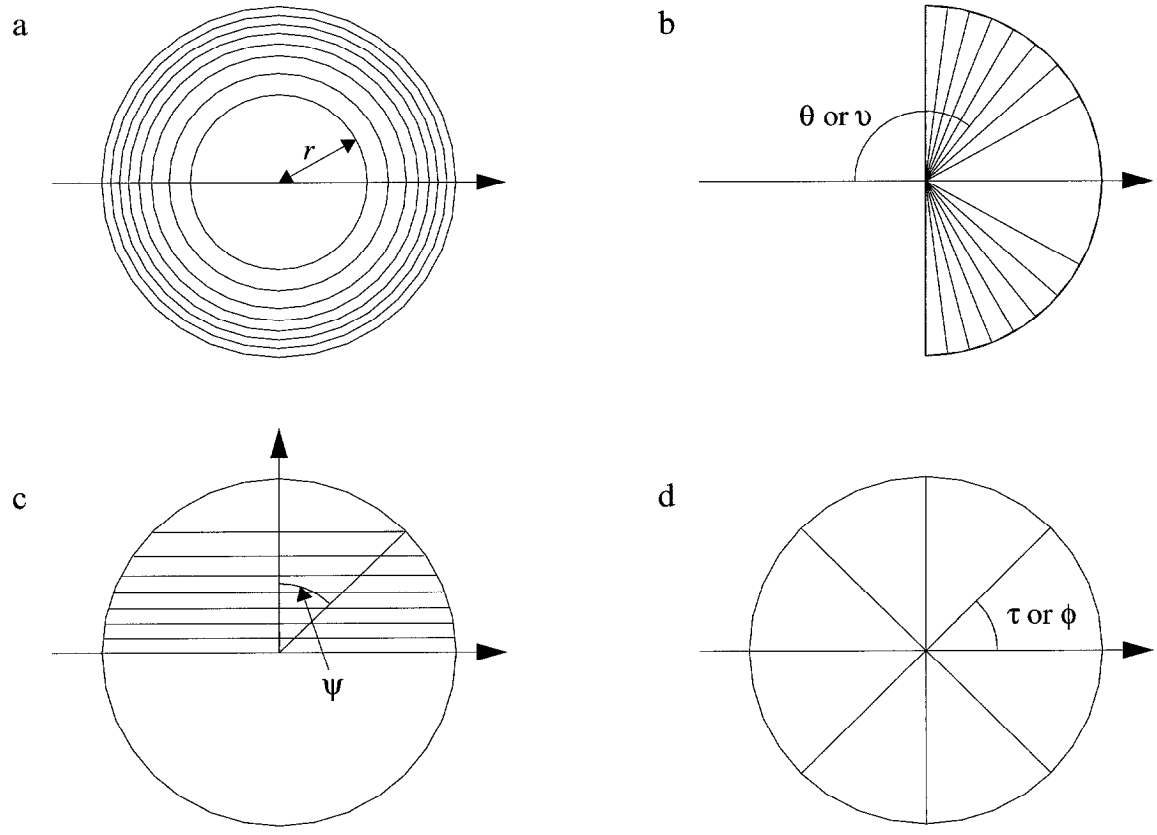


Fig. 4. Cross-sections through the centre of a sphere showing the contour lines for hydrogen-bond parameters delineating eight partitions of equal volume: (a) r contours; (b) hemisphere, viewed from the equator, showing θ (or υ) contours; (c) hemisphere viewed from the equator, showing ψ contours; (d) view from above the poles, showing ϕ contours, or view down the θ (or υ) axis, showing τ contours.

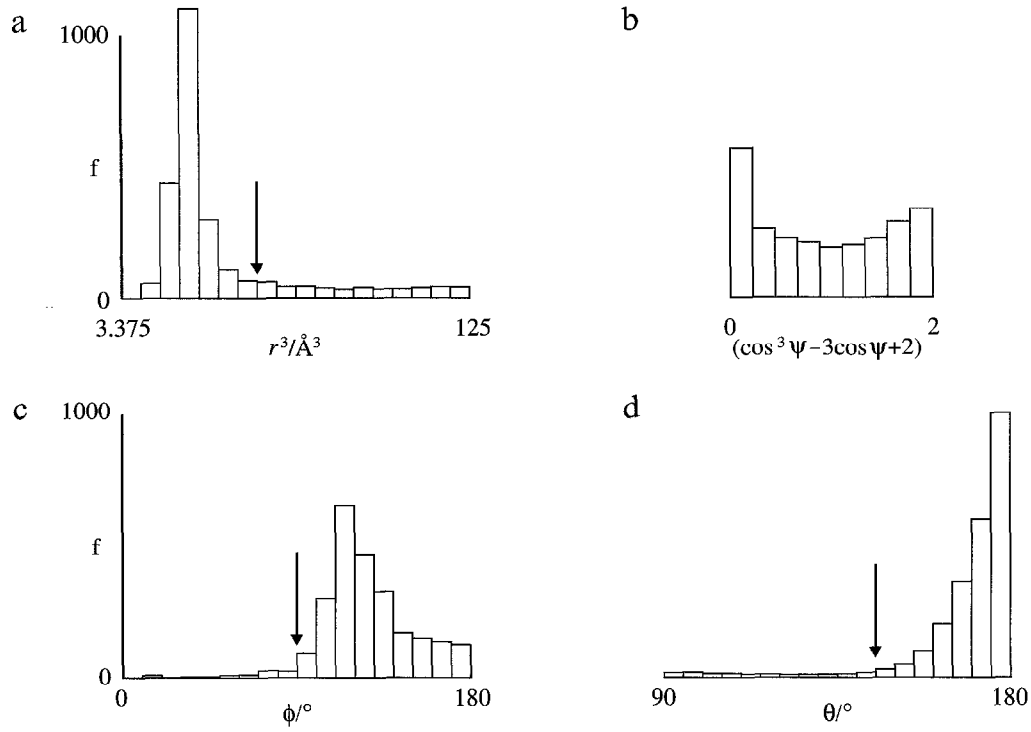


Fig. 5. Frequency distributions for the intermolecular contacts of hydrogen-bond donors with the *syn* lone pair of the carboxylate oxygen atom: (a) r^3; (b) cos ψ; (c) ϕ; (d) normalized distribution for θ. The arrows indicate the cutoff points.

ϕ, where appropriate) values of more than 90°. The crystals must also satisfy a number of standard criteria in that they must be error-free, contain no disorder, contain all atomic coordinates and have an R factor of less than 0.1. Only organic, nonpolymeric crystals are considered. These criteria reduce the number of crystals examined from 140 000 to about 48 000. Additionally, the position of the hydrogen atom attached to a nitrogen or oxygen atom is normalized according to the procedure devised by Taylor and Kennard [27] to remove discrepancies obtained on pooling data from neutron- and X-ray-diffraction studies.

When identifying the hydrogen bonds about acceptor groups, all the complementary donor groups can be represented by X-H, where X is either N or O. There is no obvious way of representing all complementary acceptor groups in the instruction file for a donor group. Instead, three separate searches are carried out, using the three complementary groups shown in Fig. 3.

Hydrogen-bond criteria

The hydrogen bonds are sifted from the contacts output from the first stage by generating frequency distributions for each of the geometric parameters, such that the volume represented by each interval is equal (volumes vary with r^3, ϕ, τ, $\cos\theta$, $\cos\upsilon$ or $\cos^3\psi - 3\cos\psi + 2$). The intervals therefore appear as shown in Fig. 4. This ensures that the number of observations in each bin is representative of the density of points and therefore of the probability of observing a hydrogen bond in that bin. The resultant frequency distributions for the *syn* lone pair of the carboxylate group are shown in Fig. 5. In the cases of the distributions for r, ϕ and θ, it can be seen that cutoffs can be applied, beyond which the values of the frequency distribution appear to be no more than background noise. However, the distributions are not normal distributions, so hydrogen bonds cannot be sifted by determining the confidence intervals for each of the parameters. Two statistical tests are used to ascertain the positions of the cutoffs. Each involve moving from the 'non-hydrogen-bond' end (i.e. high r, low ϕ and low θ) of the distribution in steps of one interval, looking for a significant change. In the first test, the variances of successive blocks of three intervals are compared using the F-test [28]. This test has the disadvantage that if the actual cutoff appears before the fourth interval, it cannot be identified. Hence, in the other test, the z-test, the current value is tested to see if it belongs to the normal distribution created from the previous values. The first cutoff is applied to the parameter for which there is the lowest value of P as determined by the F-test. In the carboxylate example, this is at $\theta = 145°$. The frequency distributions are then recalculated and the same procedure is followed. In the carboxylate example, the cutoffs are calculated to be $r =$ 3.30 Å and $\phi = 90°$. This procedure was adopted for all the hydrogen-bonding groups listed in Tables 1 and 2. In the case of donor groups, separate criteria were derived for the hydrogen bonds to each of the acceptor types shown in Fig. 3.

TABLE 3
GROUP-PROBABILITY VALUES (p) FOR ACCEPTOR GROUPS

Group[a]	N[b]	p	log K_β	m
Terminal phosphate	195	0.90	3.2[c]	
Anti phosphinyl	572	0.74		
Syn carboxylate	2686	0.68		1.69[c]
Water	3088	0.65	1.2	
Asymmetric Het_6N	646	0.54	1.5–2.5	
Trans acid O	2119	0.54		
Syn phosphinyl	572	0.53		
Symmetric Het_3N	797	0.48	1.8–3.7	0.81
Phosphine	131	0.47	3.9[c]	
Anti carboxylate	2686	0.45		1.69[c]
Sulphoxide	151	0.40	2.4–3.0[c]	
Primary amine	121	0.40	2.8	
Asymmetric Het_5N	172	0.39	1.1–2.7	
Thiocarbonyl	591	0.39	1.2–2.0[c]	
Rotatable hydroxyl	10 623	0.38	1.4	0.88
Symmetric Het_6N	299	0.38	1.0–3.5	
Terminal sulphate	1029	0.35		
Tertiary amine	917	0.32	2.8	
Amide C	6170	0.32	2.5–3.1[c]	1.16[c]
Amide N	6170	0.32	2.5–3.1[c]	1.16[c]
Trans carbamate N	593	0.30	2.0–2.4[c]	
Cis carbamate N	77	0.30	2.0–2.4[c]	
Cis ester O	1028	0.27	1.7[c]	
Cis acid O	227	0.25		
Cis carbamate O	77	0.22	2.0–2.4[c]	
Nitrile	521	0.21	0.9–1.2	
Imine	1514	0.19	1.1–1.5	
Ketone	3240	0.18	1.4–1.6[c]	
Secondary amine	490	0.18		
Trans carbamate O	593	0.17	2.0–2.4[c]	
Cis ester C	1028	0.17	1.7[c]	
Trans acid C	2119	0.15		
Cis acid C	227	0.14		
Phenol OH	3012	0.13	0.2	0.52
Trans ester C	3082	0.12	1.2–1.4[c]	
Ether	9655	0.11	1.3–1.7	
Phenyl NH_2	1540	0.08	~1.0	
Trans ester O	3082	0.07	1.2–1.4[c]	
Syn nitro	3338	0.05	0.7[c]	
Het_5O	261	0.04		
Anti nitro	3338	0.03	0.7[c]	
Syn sulphone	2536	0.02	1.0–1.6[c]	
Anti sulphone	2536	0.01	1.0–1.6[c]	

[a] Het_nX denotes an X atom in an n-membered heterocycle.
[b] N is the number of appearances of the group in the database for which a complementary group is also present.
[c] The log K_β and m values are likely to be overestimates because they allow for hydrogen bonding to both lone pairs.

Group-probability values

The group-probability value is calculated as the number of hydrogen bonds involving the group divided by the number of appearances of the group in the database for which there are also complementary groups present in the crystal. This removes the possibility of underestimating the value on account of there being crystals containing either only hydrogen-bond donors or acceptors, although it takes no account of the number of complementary groups present in the crystal, which could feasibly affect the results obtained.

Distribution of hydrogen bonds

In order to characterize the 3D properties of the hydrogen bonds, 3D bins are established by merging the partitions shown in Fig. 4 to produce a number of pseudo-spherical cubes each with equal volume. The number of observations in each of these bins is directly proportional to the density of complementary atoms, and therefore representative of the probability of finding a hydrogen-bonding atom in that bin. Comparisons between the 3D distributions of two sets of hydrogen bonds can then be made by carrying out χ^2-tests [28] on the number of observations in each equivalent bin.

Results

Group-probability values

The group-probability values derived from this study are compared with two equivalent parameters derived from other studies of hydrogen bonds. The first parameter is an equilibrium constant for hydrogen-bond formation (log K_β for acceptor groups and log K_α for donor groups) derived from a solute study [29]. Equilibrium constants (related to free energy) are used in the comparison rather than spectroscopic shifts (related to enthalpy) because the former is more representative of the probability of hydrogen-bond formation. Enthalpy changes do not take account of, in particular, steric effects; they simply

TABLE 4
GROUP-PROBABILITY VALUES (p) FOR DONOR GROUPS

Group[a]	N[b]	p	log K_α	m
Het_5NH	646	0.89	1.0–3.6	0.81
sp^2-N^+H	616	0.82		0.93
Acid OH	2346	0.82	2.0	
sp^2-N^+H_2	165	0.75		1.76[c]
sp^3-N^+H_3	851	0.74		2.64[d]
sp^3-N^+H_2	341	0.73		
Rotatable OH	10 623	0.68	0.9–1.5	1.40
Water	3088	0.65	1.2	
Primary amide NH_2	750	0.62	1.5[c]	1.35[c]
Phenyl NH	1454	0.58	0.6–1.0	
Secondary amide NH	6012	0.54	0.7–1.3	0.94
Phenyl NH_2	1906	0.46	0.6–1.0[c]	
sp^3-N^+H	443	0.34		
Imine NH	52	0.33		
Phenyl OH	3012	0.30	1.0–3.1	1.06
Primary amine NH_2	121	0.26	~0[c]	
Secondary amine NH	490	0.24		

[a] Het_nX denotes an X atom in an n-membered heterocycle.
[b] N is the number of appearances of the group in the database for which a complementary group is also present.
[c] This value is likely to be overestimated because log K_α and m allow for bonding to two hydrogen atoms.
[d] This value is likely to be overestimated because m allows for bonding to three hydrogen atoms.

provide a measure of the strength of the hydrogen bond. The second parameter, *m*, the mean number of hydrogen bonds per group, is derived from a PDB crystal survey [30] of hydrogen bonds occurring within protein structures.

The group-probability values (*p*) for acceptor groups, per lone pair, are listed in Table 3, and for donor groups, per hydrogen atom, in Table 4. The *p* values for donor groups are generally higher than those for acceptor groups. This is because, of all the hydrogen-bonding groups present in the CSD, about 70% are acceptor groups and 30% are donor groups (as can be seen by adding up the *N* values in Tables 3 and 4).

Generally, the *p* values from this crystal survey and the equilibrium constants from the solute studies rank the hydrogen-bonding groups similarly. However, water and both sets of hydroxyls are ranked as better acceptors by the crystal survey. This could be because both lone pairs of the oxygen atom are available to accept hydrogen bonds, whereas *p* was calculated on the assumption that only one lone pair per oxygen atom is involved in hydrogen bonding. Alternatively, the *p* value could be increased by the cooperative effect, which is less likely to occur in solute studies because the solute is present in such small amounts. The ranking of the phenol group is not as exaggerated because one of the oxygen lone pairs is conjugated with the π-system of the aromatic ring and the cooperative effect is therefore not as prominent. No attempt has been made to correlate the parameters *p* and log *K* for a number of reasons. First, a linear relationship between equilibrium constant and relative amount of reaction product is not expected. Secondly, in the case of groups with more than one lone pair, the solute data take no account of which lone pair is forming the hydrogen bond. Whether one lone pair is favoured over the other greatly affects the expected relationship between the two parameters. Finally, the fact that many of the log *K* values are given as wide ranges makes any correlation (even by ranks) between log *K* and *p* difficult to evaluate.

The group-probability values for the protein and small-molecule surveys rank the groups in the same order if the correction for the number of lone pairs or number of hydrogen atoms in the group is made. There are two exceptions, the first of which is the heterocyclic nitrogen acting as an acceptor or donor, for which the ranking value is lower in the protein survey. This is because there are no data relating to the protonation state of the imidazole nitrogen atoms in the histidine residues of proteins. The *m* value was derived by dividing the number of hydrogen bonds by the number of heterocyclic nitrogen atoms, when in fact about half are protonated (and therefore can only donate, but not accept, hydrogen bonds) at physiological pH. Therefore, the probability values from the protein survey should be approximately doubled to allow a fair comparison with the group-probability values derived from this survey. The second exception is the

TABLE 5
HYDROGEN-BOND CRITERIA FOR ψ-ϕ HYDROGEN-BONDING GROUPS, LISTED IN DESCENDING ORDER OF GROUP PROBABILITY

Group	r (Å)<	ϕ (°)>	θ (°)>	Previous criteria
Anti $R_2PO_2^-$	3.03	110	135	r' < 2.1 Å [38]; r < 3.2 Å [39]
Syn $R_2PO_2^-$	3.03	110	145	
Syn $RCOO_2^-$	3.17	90	145	r' < 2.35 Å, θ > 90° [19]; r < 3.2 Å [39]
Anti $RCOO_2^-$	3.17	110	145	
HOH	3.03	120	145	
C=S	3.73	90	145	
C=O O (acid)	3.17	110	130	
C=O N	3.30	110	135	
C=O C (amide, ketone)	3.30	110	130	
C=O C (acid, ester)	3.17	110	135	
C=O O (ester)	3.17	110	135	
Symmetric Het_6N	3.17	150	145	r < 3.25 Å, θ > 90° [31]; r < 3.1 Å [40]
Symmetric Het_5N	3.30	140	155	
Asymmetric Het_6N	3.30	140	140	
Asymmetric Het_5N	3.30	150	135	
ROR	3.42	140	140	r < 3.0 Å [32]
Het_5O	3.42	150	135	
Syn NO_2	3.42	90	145	θ > 90°, N-O···X > 90° [41]
Anti NO_2	3.53	110	130	
RNR	3.42	140	140	
Syn R_2SO_2	3.30	110	130	r < 3.25 Å, θ > 90° [31]; r < 3.2 Å [39]
Anti R_2SO_2	3.30	120	135	

The upper half of the table represents groups for which the ϕ-axis is a bond and the lower half represents groups for which the ϕ-axis passes through the acceptor atom and the midpoint of its neighbouring atoms (Figs. 1a and b, respectively).

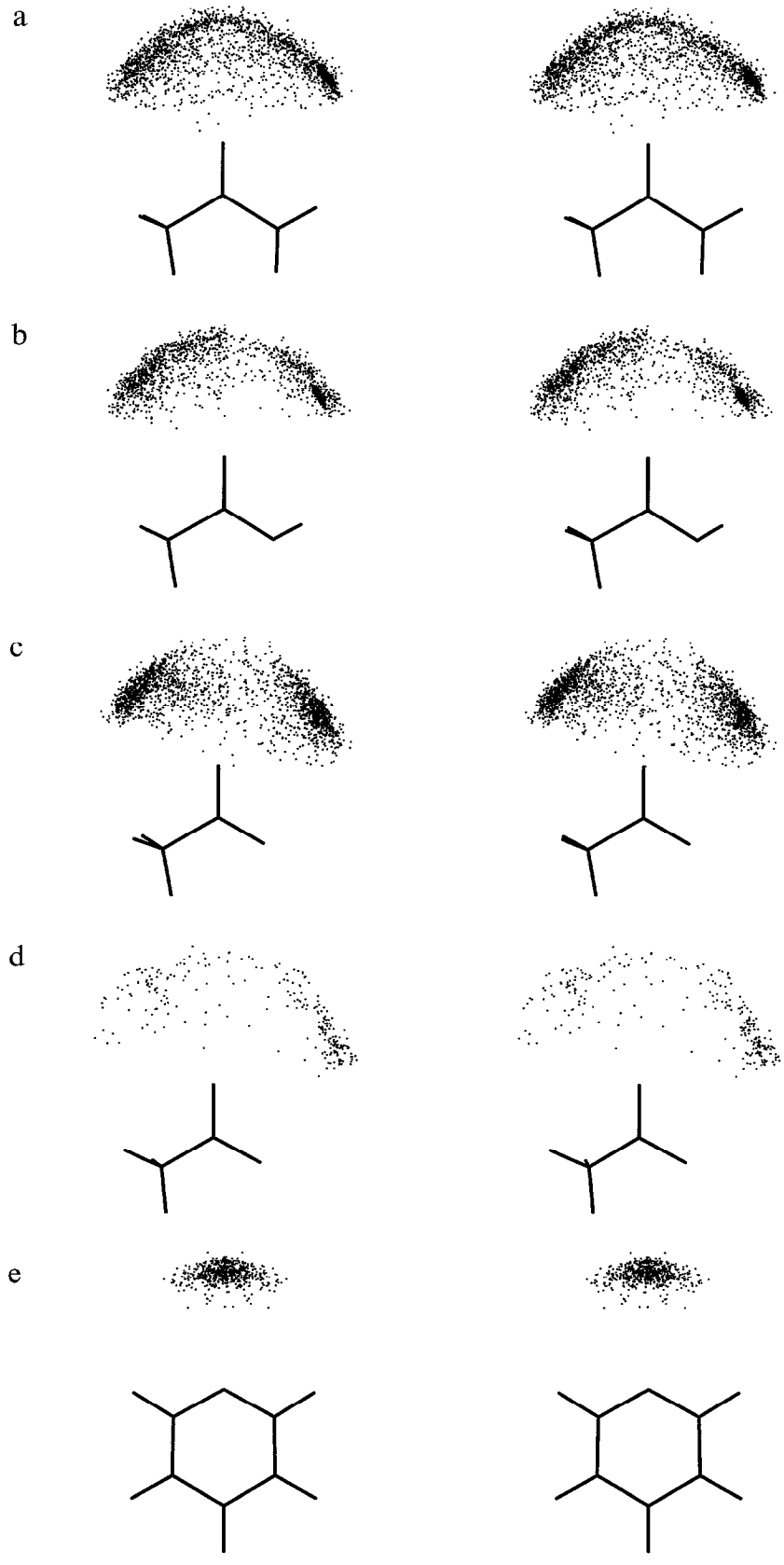


Fig. 6. Stereoviews of distributions of complementary heavy atoms about (a) amide; (b) carboxylic acid; (c) carboxylate; (d) nitro; (e) symmetric six-membered heterocycle nitrogen atom; (f) asymmetric six-membered heterocycle nitrogen atom; and (g) ether hydrogen-bonding groups. In each case, only one hemisphere is shown, as there is mirror symmetry about the plane of the group.

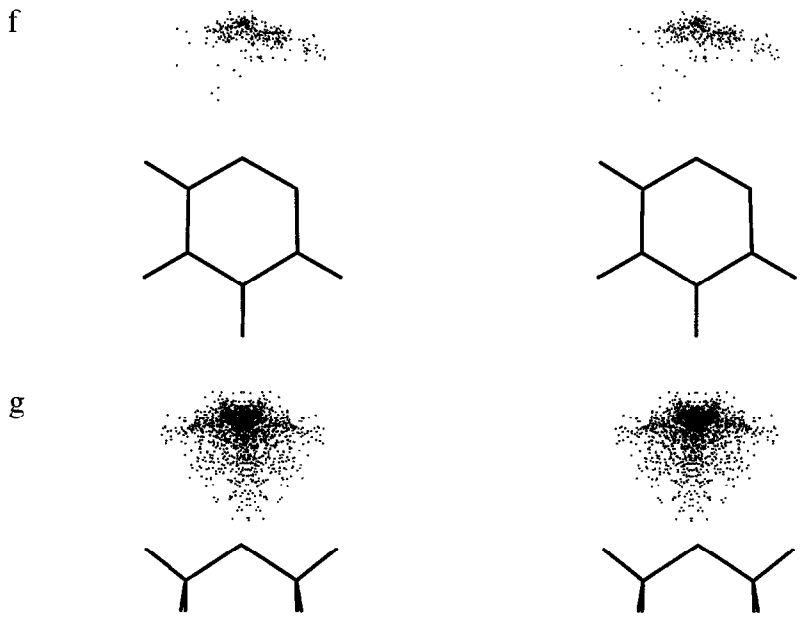


Fig. 6. (continued).

rotatable hydroxyl group, for which the ranking value is higher than expected in the protein survey. Again, this could be a consequence of not knowing the hydrogen-atom positions and also the fact that the group is amphiprotic. Two amphiprotic residues in close proximity would each be registered as both a hydrogen-bond donor and acceptor, when only one of these situations is possible.

The probability values derived from the protein study are higher than those from this survey. There are two reasons for this. First, the criteria for defining hydrogen bonds in the protein survey (H···A < 2.5 Å, where A is the acceptor atom and $\theta > 90°$) were much more lenient than those for this small-molecule survey. Secondly, unlike this survey, the protein survey included intramolecular hydrogen bonds.

3D distribution of hydrogen bonds

The hydrogen bonding to ψ-ϕ groups is illustrated by the examples shown in Fig. 6. In each case, there is a preference for hydrogen bonds to occur in the theoretical lone-pair positions, though there is a great deal of variation about these positions. Hydrogen bonds to the N lone pair (see Fig. 7 for the nomenclature of lone pairs to carbonyl groups) of the amide (Fig. 6a) and the O lone pair of the acid (Fig. 6b) can be seen to be much more directional than to the equivalent C lone pairs, probably owing to dimer formation in the crystal environment. Hydrogen bonds to the C lone pair of the acid appear to be more directional than to the equivalent lone pair of the amide. The thiocarbonyl group gives similar results to the carbonyl group (data not shown). In the case of the carboxylate group (Fig. 6c), the directionality is much more pronounced about the *syn* lone pair, in terms of distribution both about and within the lone-pair plane, as previously observed [19]. In the isosteric but weaker nitro group (Fig. 6d), this directionality is much less pronounced. The phosphinyl and sulphonyl moieties do not display any notable directionality (data not shown), presumably because the lone pairs are not well localized. In the case of the heterocyclic nitrogen atom (Figs. 6e and f), it can be seen how the presence of a neighbouring electronegative atom acts to drag the distribution of complementary donor groups towards itself, as has been observed previously [31]. Similar observations are made for

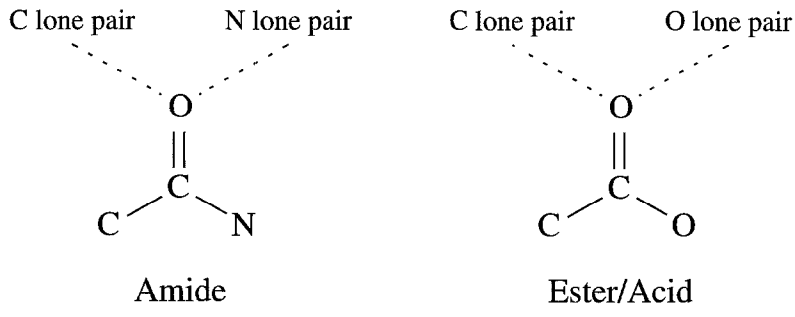


Fig. 7. Convention for labelling the lone pairs of carbonyl groups.

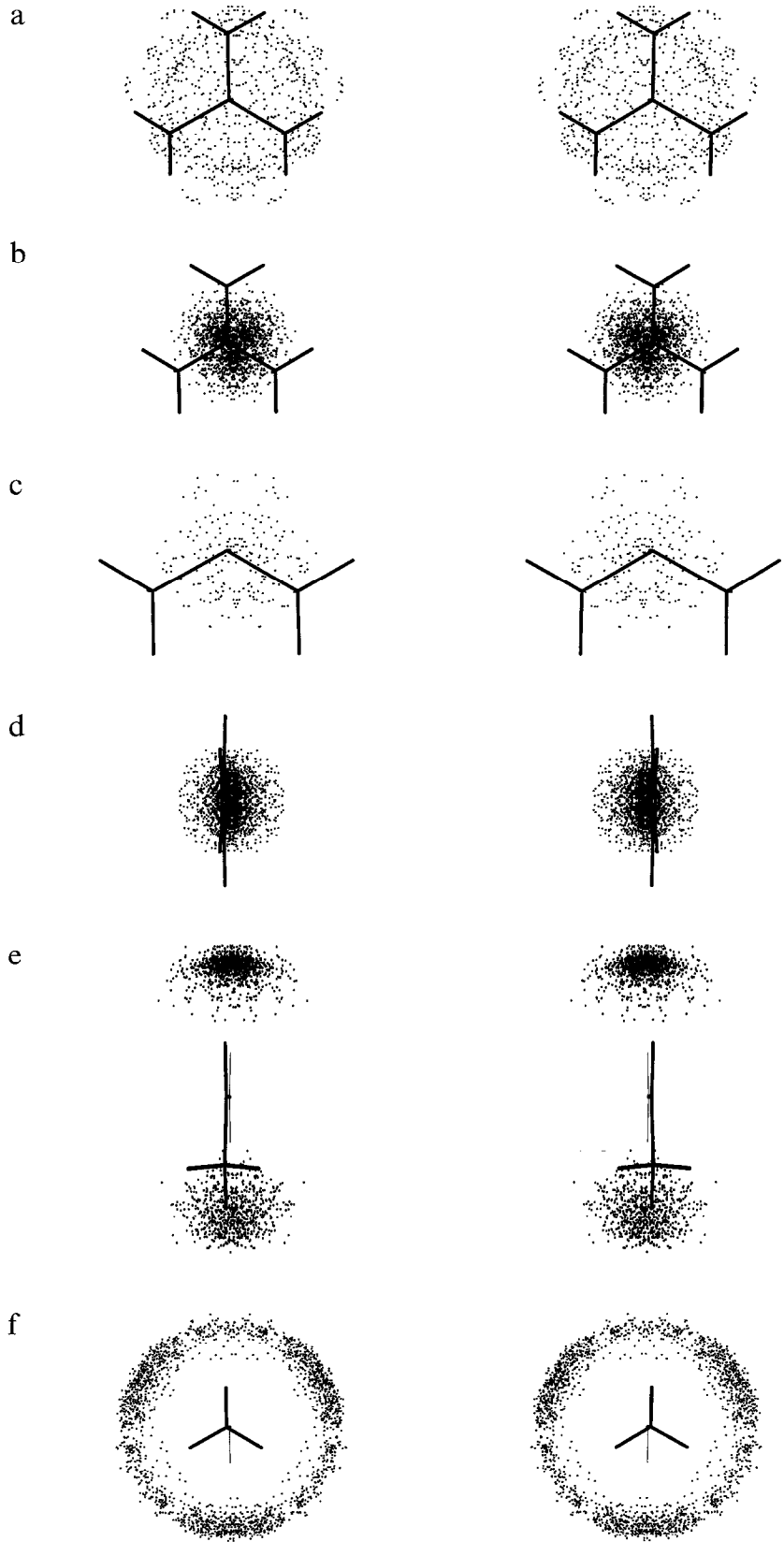


Fig. 8. Stereoviews of distributions of complementary heavy atoms about (a) phosphine; (b) R_3N^+H; (c) R_2NH (acting as an acceptor); (d) heterocyclic NH; (e) primary amide (acting as a donor); (f) rotatable hydroxyl (donor); (g) primary amine (donor); (h) rotatable hydroxyl (acceptor); (i) primary amine (acceptor); (j) sulphate; and (k) phosphate groups. In each case, the view is down the bond from which θ or υ is measured.

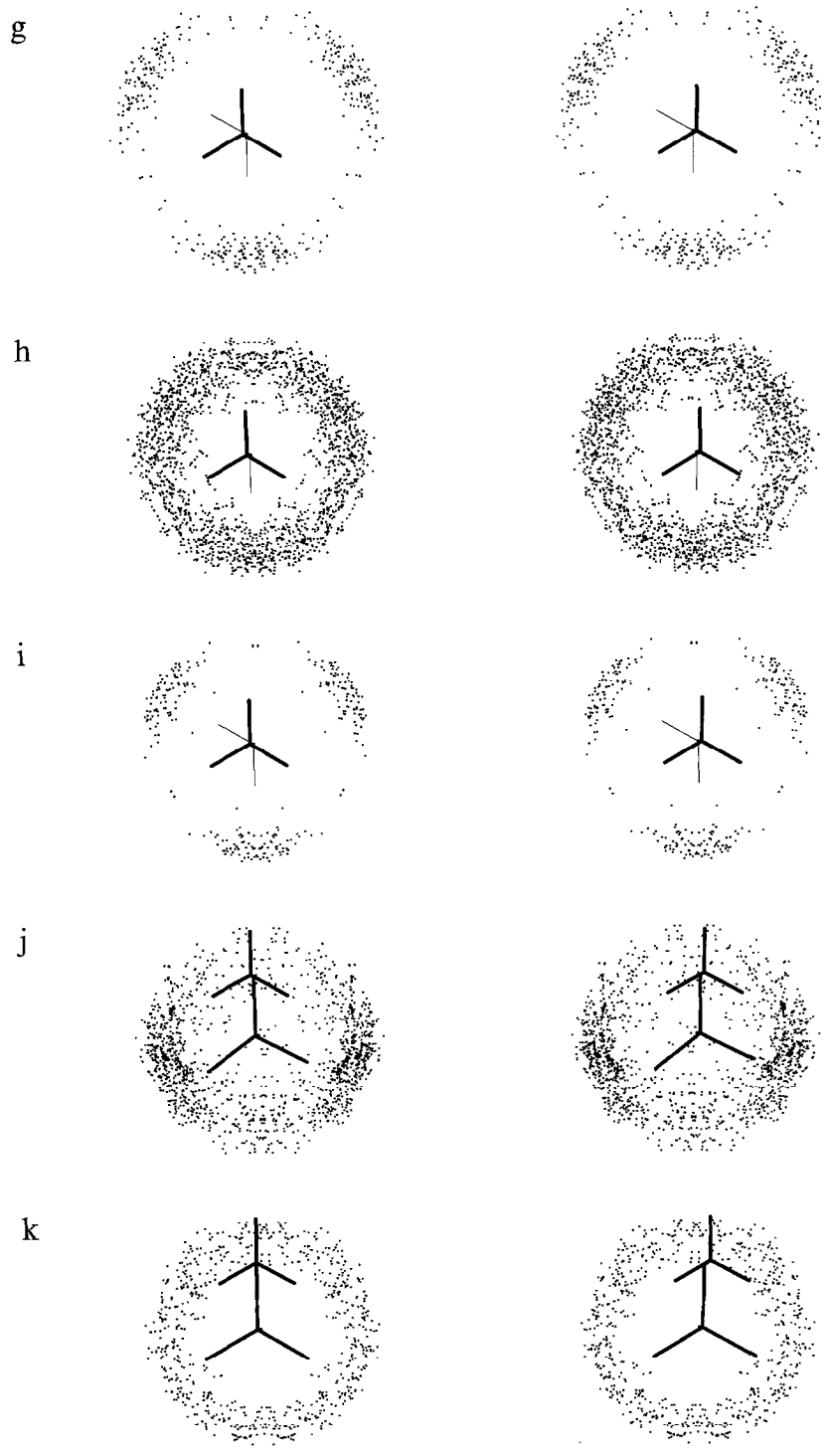


Fig. 8. (continued).

the nitrogen atoms in five-membered heterocycles. As previously observed [32], hydrogen bonds to ether groups (Fig. 6g) are observed in the plane presumed to contain the lone pairs but there is no notable directionality within this plane.

The hydrogen bonding to θ-τ groups is illustrated by the examples shown in Fig. 8. The apparent lack of directionality of hydrogen bonds to the phosphine group (Fig. 8a) implies that the lone pairs of the oxygen atom are not localized. Similar results are obtained with the sulphoxide group (data not shown). The tertiary amine group (Fig. 8b) shows a strong preference for lone-pair directionality. There is little directionality in terms of τ, presumably because steric interference by the nitrogen substituents is minimal at the high values of υ normally seen in the hydrogen bonds. Most groups display this

TABLE 6
HYDROGEN-BOND CRITERIA FOR θ-τ HYDROGEN-BOND ACCEPTOR GROUPS, LISTED IN DESCENDING ORDER OF GROUP PROBABILITY

Group	r (Å)<	υ (°)	θ (°)>	Previous criteria
$O_2(R)P$-O	3.03	103–146	145	r < 3.2 Å [39]
R_3P=O	3.03	109–180	145	
R_2S=O	3.03	100–180	145	
R-NH_2	3.03	103–180	155	
R-OH	3.03	100–161	145	
$O_2(R)S$-O	3.17	103–180	150	r < 3.2 Å [39]
R_3N-lp	3.17	153–180	145	
RC≡N	3.30	141–180	150	
$R_2(H)N$-lp	3.30	153–180	150	
Ph-OH	3.17	100–153	150	
$PhNH_2$	3.42	90–137	140	

kind of pattern, i.e. a preferred linear geometry with approximate rotational symmetry about this axis (e.g. secondary amide and amine, nitrile, imine and water; data not shown). The rotational symmetry is not apparent in secondary amines (Fig. 8c), where donor groups avoid a secondary interaction with the amine hydrogen atom. Rotational symmetry is also not observed with the heterocyclic nitrogen donor group (Fig. 8d), where acceptor groups prefer to lie in the plane of the ring. This could be caused by secondary electrostatic interactions with the relatively electropositive hydrogen atoms attached to the ring. This effect is also seen for the R_2N^+H group (data not shown). The preference towards hydrogen bonding to the *Z* hydrogen atom of the primary amide (Fig. 8e; 64% of the hydrogen bonds are to the *Z* atom) again reflects the tendency for this group to dimerize (see also Fig. 6a), as was also observed in carboxylic acids. Indeed, on further examination, 55% of the hydrogen bonds between *Z* atoms and sp^2-oxygen acceptor atoms were observed to be dimer interactions. As expected, acceptor groups positioned about the rotatable hydroxyl group (Fig. 8f) prefer a staggered position because the hydrogen atom preferentially adopts a staggered orientation to minimize 1,4 steric interactions. This effect is more pronounced in the primary amine group (Fig. 8g) because there are two hydrogen atoms, and therefore potentially twice the amount of unfavourable steric interaction in the eclipsed orientation. Complementary donor groups are much less localized about the hydroxyl group (Fig. 8h) because the oxygen atom has two lone pairs, which are themselves not localized (as seen with the isosteric ether oxygen atom in Fig. 6g). Similarly, with the phenol group, complementary acceptor atoms are preferably positioned in the plane of the ring whereas donor atoms approach the ring from almost any angle (data not shown). In contrast, the single lone pair of a nitrogen atom is much more localized and, consequently, the hydrogen bonds to the primary amine (Fig. 8i) and aniline (data not shown) groups are more directional. From a drug designer's perspective, hydrogen bonding to primary amines is therefore more predictable than to rotatable hydroxyl groups. A comparison of the sulphate (Fig. 8j) and phosphate groups (Fig. 8k) shows that only the former encourages a secondary interaction with one of the other terminal oxygen atoms in the group. The phosphate group has a higher *p* value, so it is likely that all three oxygen atoms form hydrogen bonds, sterically precluding secondary electrostatic interactions.

Hydrogen-bond criteria

The hydrogen-bond criteria for acceptor groups are shown in Tables 5 and 6, along with cutoffs used in previously published CSD surveys. The criteria for donor groups are shown in Tables 7 and 8. For comparison, the only previous surveys listing criteria for donor groups were the two general surveys of Klebe [21] (i.e. 2.0 Å < r < 3.2 Å, 1.4 Å < r' < 2.2 Å, θ > 130°) and Görbitz [33] (r' < 2.4 Å, θ > 110°). It can be seen by examining the tables that, generally, the stronger hydrogen-bonding groups (in terms of group-probability value) require stricter criteria to define their hydrogen bonds. The apparently anomalous high r cutoff observed in the case of the thiocarbonyl group is explained by the increased van der Waals radius

TABLE 7
HYDROGEN-BOND CRITERIA FOR θ-τ HYDROGEN-BOND DONOR GROUPS, LISTED IN DESCENDING ORDER OF GROUP PROBABILITY

Group	sp^2-O acceptor		sp^3-O/sp^2-N acceptor			sp^3-N acceptor		
	r (Å)<	θ (°)>	r' (Å)<	θ (°)>	φ (°)>	r' (Å)<	θ (°)>	υ (°)>
Het_5 N-H	2.09	146	2.09	141	140			
R_2-N^+H	1.92	146	1.92	153	150			
R_2-N^+H_2	2.09	141	2.09	141	155			
HO-H	2.36	146	2.23	146	140	2.23	146	150
PhN-H	2.23	136	2.23	141	150			
CON-H	2.23	146	2.23	146	140	2.36	141	145
R_3N^+-H	1.92	146	2.09	146	140			
R=N-H	2.23	132	2.36	146	140			
R_2N-H	2.36	141	2.48	146	150	2.48	153	155

of the sulphur acceptor atom (1.80 Å) compared with that of the nitrogen (1.55 Å) and oxygen (1.52 Å) atoms [34]. The other main exception is the phenol group, for which the criteria are strict in comparison with its apparently low group-probability value. This is explained by the fact that the hydrogen bonds that are formed by the phenol group are directional, but steric influences of *ortho* substituents lower the group-probability value. When only phenol groups without *ortho* substituents are considered, the accepting group-probability value changes from 0.13 to 0.32, moving Ph-OH up two places in Table 6, and the donating probability changes from 0.30 to 0.70, moving Ph-OH up three places in Table 8 (data not shown). In each case, the criteria then correspond with the group-probability value.

Independence of geometric parameters

The relationship between geometric parameters is investigated for the hydrogen bonds to the *syn* lone pair of the carboxylate group and the rotatable hydroxyl group. The hydrogen-bond distribution is divided into a number of pools on the basis of r values in order to observe more clearly the trends for the angular parameter distributions. The results are shown in the form of cumulative distribution functions (cdf's) in Fig. 9. In the case of the carboxylate group (a and b), as the r value decreases, the distributions of ψ and ϕ become more directional in nature. In the cdf's, this is seen as a much steeper line at the parameter values $\psi \approx 90°$ and $\phi \approx 120°$. As the hydrogen bond becomes shorter, it also becomes more directional. Kolmogorov–Smirnov tests [28] on the cdf's show that all the curves are significantly different (at the 5% level) from each other, with the exception of the ϕ curves (2) and (3), and (3) and (4), and the ψ curves (4) and (5). Similarly, hydrogen bonds with high ψ values tend to be shorter and with ϕ values close to 120°, and as ϕ approaches

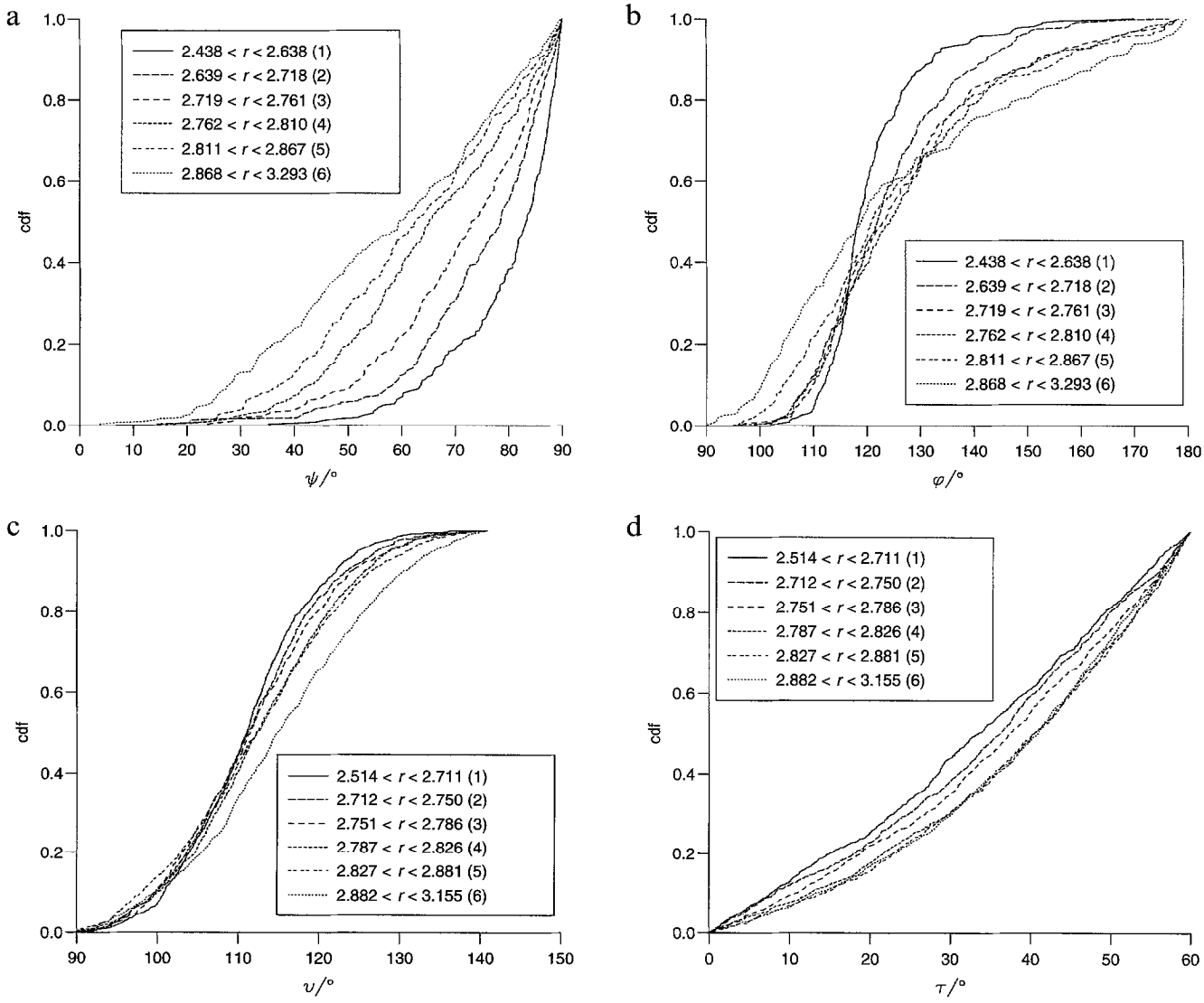


Fig. 9. The cdf's showing the dependence of (a) ψ and (b) ϕ on r for hydrogen bonds to the *syn* lone pair of the carboxylate group and the dependence of (c) υ and (d) τ on r for hydrogen bonds to the rotatable hydroxyl group.

TABLE 8
HYDROGEN-BOND CRITERIA FOR υ-τ HYDROGEN-BOND DONOR GROUPS, LISTED IN DESCENDING ORDER OF GROUP PROBABILITY

Group	sp^2-O acceptor				sp^3-O/sp^2-N acceptor					sp^3-N acceptor				
	r (Å)<	υ (°)	θ (°)>	τ (°)	r (Å)<	υ (°)	θ (°)>	τ (°)	φ (°)>	r (Å)<	υ (°)	θ (°)>	τ (°)	υ_N (°)>
CO-OH	2.87	103–128	140	<30.0	2.87	103–128	155	<30.0	150	2.87	103–128	140	<30.0	150
R-N^+H_2	3.17	90–146	140	<30.0	3.17	90–146	140	<22.5	150					
R-N^+H_3	3.17	90–136	140	>20.0	3.17	90–132	150	>25.0	130	3.17	90–124	155	>15.0	150
R-OH	3.17	90–141	130		3.03	90–141	135		140	3.03	100–132	145		150
CO-NH_2	3.30	103–153	135	<45.0	3.30	109–141	140	<30.0	140	3.30	113–161	125	<30.0	145
Ph-NH_2	3.30	100–146	145	<37.5	3.30	103–146	140	<37.5	130					
Ph-OH	3.03	103–136	145	<37.5	3.03	103–136	150	<45.0	140	2.87	109–128	145	<52.5	150
R-NH_2	3.30	90–128	150	>30.0	3.42	90–124	155	>25.0	130	3.42	90–132	155	>20.0	140

120°, the hydrogen bonds are observed to be shorter and closer to the lone-pair plane (data not shown). Similar results are obtained with the other hydrogen-bonding groups shown in Tables 1 and 2, in that any directionality is emphasized by shorter r values (data not shown).

The one notable exception is the rotatable hydroxyl group acting as a hydrogen-bond donor (Figs. 9c and d) and acceptor (data not shown), for which shorter hydrogen bonds are generally more directional in terms of υ (curve (4) differs significantly from (3), and (6) and (5) also differ), but less directional in terms of τ (curves (1) and (2), and (3) and (4) differ significantly), i.e. there is less preference for the staggered conformation. This is probably because the higher energy of the shorter hydrogen bond more effectively compensates for the energy penalty paid by the group adopting an eclipsed conformation. That is to say, the hydrogen-bond directionality is normally determined by steric factors, which can be overcome if a hydrogen bond is strong enough. This phenomenon is not seen with the rotatable amine groups because the increased steric interactions cannot be overcome by hydrogen-bond formation.

Discussion and Conclusions

This paper describes a comprehensive crystal survey of hydrogen-bond geometry, covering all the hydrogen-bonding groups of interest to the drug designer, with a view to the prediction of the hydrogen-bonding properties of ligands. Many hydrogen-bond crystal surveys for individual groups have appeared in the literature. Apart from the fact that hydrogen-bond data relating to groups such as the amine and nitrile groups have yet to be published, the data derived from previous surveys were considered unsuitable for use in hydrogen-bond prediction for a number of reasons.

First, and most importantly, all previously published data (with the exception of Ref. 16) only list frequency distributions for individual hydrogen-bond parameters (i.e. r, φ, ψ or θ). If these data were used for the calculation of hydrogen-bond probability values, the probability, P, of finding a hydrogen-bond donor atom at a position given by the spherical polar coordinates (r,ψ,ϕ) in relation to, for example, a carbonyl acceptor atom would be calculated as

$$P = P(r) \times P(\psi) \times P(\phi) \quad (1)$$

where $P(r)$ is the probability of a hydrogen bond occurring at a distance of r, and $P(\psi)$ and $P(\phi)$ are the probabilities of a hydrogen bond occurring with angles ψ and φ, respectively. This approach was used in the calculation of hydrogen-bond probability values by the algorithm HSITE [1] and in the calculation of hydrogen-bond interaction energy values by GRID [4]. It relies on the assumption that the distributions of r, ψ and φ are independent. This has been shown to be an inaccurate assumption by previous work [18] and by work presented here, even for groups for which hydrogen-bond directionality is not very pronounced. Furthermore, for each group, geometric probability values have been calculated from the 3D data and compared (using the χ^2-test; significance level 0.01) with the values calculated using Eq. 1, based on the same data. For every group, the probability distribution over 3D space was significantly different (data not shown). In this work, the geometric data were collected to represent accurately the 3D distribution of complementary atoms about each hydrogen-bonding group. These data have already been used in deriving a hydrogen-bond similarity score for two superposed molecules [35]. The points on the common surface of the two molecules were scored as common hydrogen-bonding points if the hydrogen-bond probability values of the points were nonzero for each molecule.

Secondly, the use of data collected from many different sources introduces inconsistencies in terms of the parameters used to describe the geometry of a hydrogen bond. For example, hydrogen-bond distances are sometimes measured using the two heavy atoms (r in this paper) and sometimes using the donor-hydrogen atom and the acceptor atom (r' in this paper). Also, when considering the geometries of hydrogen bonds to donor groups, only dis-

tributions of r and θ are given. Any probability evaluation using these data would require complete rotational symmetry about the axis from which θ is measured. This cannot be assumed, especially when considering planar groups such as heterocycles and carboxylic acids.

Thirdly, previous surveys were inconsistent in terms of the geometric criteria used to define whether an interaction was to be considered as a hydrogen bond. The idea that a hydrogen bond is defined when the distance between the two heavy atoms is less than the sum of their van der Waals radii is open to question [36]. The geometric criteria used to define hydrogen bonds in other surveys have been varied and so would not provide an accurate representation when making comparisons between different groups. The recent survey by Klebe [21] used the same geometric criteria to define hydrogen bonds to all groups. In the survey described in this paper, the geometric criteria are defined independently and rationally for each hydrogen-bond group. In terms of distance cutoffs, the criteria are similar, though not identical, to those used previously. The current survey differs most from previous surveys in that angle criteria must be satisfied at the complementary group as well as at the group of interest. It is also notable that, in general, the criteria become more strict (lower r, higher θ) as the groups become stronger, i.e. stronger hydrogen bonds are more directional. In a sense, this justifies the derivation of independent criteria for different groups. Not only do the criteria from previous surveys differ in terms of distance and angle thresholds, but the inclusion of multicentre and intramolecular hydrogen bonds varies over different surveys. In this study, only intermolecular contacts are considered, since it is the intermolecular interaction between ligand and protein that is being modelled. The geometries of multicentre interactions between N-H and C=O groups are known to differ from simple two-centred interactions [37]. In this survey, the 3D data for multicentred interactions are only included if they do not affect the overall geometric distribution. The hydrogen-bond criteria derived from this survey can be used in future surveys of the CSD, or in surveys of the PDB (though possibly with a tolerance to allow for the larger errors in structure determination from protein crystals). They could also be used in the form of rapid look-up tables to determine whether or not a position (or vector) in space about a molecule could be occupied by a hydrogen-bonding group. This may be useful in database searches in which ligands binding to a given site of known structure are sought.

Fourthly, the survey has been the first small-molecule survey to be used in the derivation of group-probability values, which give a measure of the propensity of a group to form a hydrogen bond. These values were found to compare well with previously published equivalent parameter values derived from solute studies and from a crystal survey of the PDB.

Group-probability values can be used in conjunction with the 3D distribution data in the construction of hydrogen-bond probability maps that can be used as a basis for molecular superposition (manuscript in preparation). The maps are a representation of all the possible positions of complementary receptor hydrogen-bonding atoms, which are assumed to be constant for a series of ligands known to bind to the same site. Hence superposition of the maps is a more reliable technique than superposition of the ligand hydrogen-bonding atoms, which are not required to be coincident in order to allow binding to the same receptor atom. It is also more realistic than superposing ligands on the basis of one or two projected site points per hydrogen-bonding atom (i.e. points positioned in the positions of formation of the ideal hydrogen bond) because there is clearly a great deal of deviation about the optimum direction of a hydrogen bond. The crystal-survey data provide the foundations for a novel approach to molecular superposition and could therefore prove useful in molecular-similarity studies, database searching and de novo drug design.

Acknowledgements

We wish to thank Rhône-Poulenc Rorer (J.E.J.M.) and the Wellcome Trust through the PRF scheme (P.M.D.) for personal financial support. We also wish to acknowledge the use of EPSRC's Chemical Database Service at Daresbury. Part of this work was carried out in the Cambridge Centre for Molecular Recognition, funded by the BBSRC.

References

1 Danziger, D.J. and Dean, P.M., Proc. R. Soc. London, B236 (1989) 101.
2 Danziger, D.J. and Dean, P.M., Proc. R. Soc. London, B236 (1989) 115.
3 Goodford, P.J., J. Med. Chem., 28 (1985) 849.
4 Boobbyer, D.N.A., Goodford, P.J., McWhinnie, P.M. and Wade, R.C., J. Med. Chem., 32 (1989) 1083.
5 Böhm, H.-J., J. Comput.-Aided Mol. Design, 6 (1992) 61.
6 Böhm, H.-J., J. Comput.-Aided Mol. Design, 6 (1992) 593.
7 Eisen, M.B., Wiley, D.C., Karplus, M. and Hubbard, R.E., Proteins, 19 (1994) 199.
8 Kato, Y., Inoue, A., Yamada, M., Tomioka, N. and Itai, A., J. Comput.-Aided Mol. Design, 6 (1992) 475.
9 Martin, Y.C., Bures, M.G., Danaher, E.A., DeLazzer, J., Lico, I. and Pavlik, P.A., J. Comput.-Aided Mol. Design, 7 (1993) 83.
10 Abraham, M.H., Chem. Soc. Rev., 22 (1993) 73.
11 Leahy, D.E., Morris, J.J., Taylor, P.J. and Wait, A.R., J. Chem. Soc. Perkin Trans. II, (1992) 705.
12 Lindroos, J., Peräkylä, M., Björkroth, J.-P. and Pakkanen, T.A., J. Chem. Soc. Perkin Trans. II, (1992) 2271.
13 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.

14 Tintelnot, M. and Andrews, P., J. Comput.-Aided Mol. Design, 3 (1989) 67.
15 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.
16 Taylor, R., Kennard, O. and Versichel, W., J. Am. Chem. Soc., 105 (1983) 5761.
17 Taylor, R., Kennard, O. and Versichel, W., Acta Crystallogr., B40 (1984) 280.
18 Ceccarelli, C., Jeffrey, G.A. and Taylor, R., J. Mol. Struct., 70 (1981) 255.
19 Görbitz, C.H. and Etter, M.C., J. Am. Chem. Soc., 114 (1992) 627.
20 Görbitz, C.H. and Etter, M.C., J. Chem. Soc. Perkin Trans. II, (1992) 131.
21 Klebe, G., J. Mol. Biol., 237 (1994) 212.
22 Taylor, R. and Kennard, O., J. Am. Chem. Soc., 104 (1982) 5063.
23 Desiraju, G.D., Acc. Chem. Res., 24 (1991) 290.
24 Etter, M.C., J. Phys. Chem., 95 (1991) 4601.
25 Rosenfield Jr., R.E., Parthasarathy, R. and Dunitz, J.D., J. Am. Chem. Soc., 99 (1977) 4861.
26 Murray-Rust, P., Stallings, W.C., Monti, C.T., Preston, R.M. and Glusker, J., J. Am. Chem. Soc., 105 (1983) 3206.
27 Taylor, R. and Kennard, O., Acta Crystallogr., B39 (1983) 133.
28 Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P., Numerical Recipes in FORTRAN, 2nd ed., Cambridge University Press, Cambridge, U.K., 1992, pp. 613–620.
29 Abraham, M.H., Duce, P.P., Prior, D.V., Barratt, D.G., Morris, J.J. and Taylor, P.J., J. Chem. Soc. Perkin Trans. II, (1989) 1355.
30 McDonald, I.K. and Thornton, J.M., J. Mol. Biol., 238 (1994) 777.
31 Vedani, A. and Dunitz, J.D., J. Am. Chem. Soc., 107 (1985) 7653.
32 Murray-Rust, P. and Glusker, J.P., J. Am. Chem. Soc., 106 (1984) 1018.
33 Görbitz, C.H., Acta Crystallogr., B45 (1989) 390.
34 Bondi, A., J. Phys. Chem., 68 (1964) 441.
35 Perkins, T.D.J., Mills, J.E.J. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 479.
36 Jeffrey, G.A. and Saenger, W., Hydrogen Bonding in Biological Structures, Springer, Berlin, Germany, 1991.
37 Taylor, R., Kennard, O. and Versichel, W., J. Am. Chem. Soc., 106 (1984) 244.
38 Alexander, R.S., Kanyo, Z.F., Chirlian, L.E. and Christianson, D.W., J. Am. Chem. Soc., 112 (1990) 933.
39 Pirard, B., Baudoux, G. and Durant, F., Acta Crystallogr., B51 (1995) 103.
40 Llamas-Saiz, A.L., Foces-Foces, C., Mo, O., Yañez, M. and Elguero, J., Acta Crystallogr., B48 (1992) 700.
41 Panunto, T.W., Urbánczyk-Lipkowska, Z., Johnson, R. and Etter, M.C., J. Am. Chem. Soc., 109 (1987) 7786.