Strong and Weak Hydrogen Bonds in the Protein–Ligand Interface

Sunil K. Panigrahi and Gautam R. Desiraju*
School of Chemistry, University of Hyderabad, Hyderabad 500 046, India

ABSTRACT The characteristics of N—H···O, O—H···O, and C—H···O hydrogen bonds and other weak intermolecular interactions are analyzed in a large and diverse group of 251 protein–ligand complexes using a new computer program that was developed in-house for this purpose. The interactions examined in the present study are those which occur in the active sites, defined here as a sphere of 10 Å radius around the ligand. Notably, N—H···O and O—H···O bonds tend towards linearity. Multifurcated interactions are especially common, especially multifurcated acceptors, and the average degree of furcation is 2.6 hydrogen bonds per furcated acceptor. A significant aspect of this study is that we have been able to assess the reliability of hydrogen bond geometry as a function of crystallographic resolution. Thresholds of 2.3 and 2.0 Å are established for strong and weak hydrogen bonds, below which hydrogen bond geometries may be safely considered for detailed analysis. Interactions involving water as donor or acceptor, and C—H···O bonds with Gly and Tyr as donors are ubiquitous in the active site. A similar trend was observed in an external test set of 233 protein–ligand complexes belonging to the kinase family. Weaker interactions like X—H···π (X = C, N, O) and those involving halogen atoms as electrophiles or nucleophiles have also been studied. We conclude that the strong and weak hydrogen bonds are ubiquitous in protein–ligand recognition, and that with suitable computational tools very large numbers of strong and weak intermolecular interactions in the ligand–protein interface may be analyzed reliably. Results confirm earlier trends reported previously by us but the extended nature of the present data set mean that the observed trends are more reliable. Proteins 2007;67:128–141. © 2007 Wiley-Liss, Inc.

Key words: hydrogen bond; protein data bank; force field; water; interaction

INTRODUCTION

The three-dimensional architecture of proteins is stabilized to a substantial degree by hydrogen bonds. Because of their strength these interactions are specific, with conserved orientation.[1,2] Because of their weakness, however, they are also made and broken rapidly during complexation, conformational change, and folding.[3] Accordingly, hydrogen bonds in biomolecules may be switched on or off with energies that are within the range of thermal fluctuations. This is one of the prime factors that facilitates ligand binding in the active site, and biological activity. Effectively, the dual strong/weak nature of hydrogen bonds is exploited by Nature to achieve specificity of both structure and function. The importance of weak interactions also varies with the type of biomolecules in the same way that molecules of structural importance might differ from molecules of enzymatic importance.[4,5]

The literature on hydrogen bonding in biomolecules is voluminous. A seminal review by Hubbard and Baker in 1984 was followed in 1991 by the book of Jeffrey and Saenger which provides much valuable information.[6,7] The subject was reviewed in depth by Glusker in 1995.[8] Work by Sundaralingam on nucleic acids and Derewenda on globular proteins in the mid-to-late 1990s widened the scope of this field.[9–11] Since 2000, there have been a number of papers that have attempted to analyze the systematics of hydrogen bonds in biological structures.[12–18]

We have previously examined the characteristics of strong (N—H···O, O—H···O) and weak (C—H···O) hydrogen bonds in a group of 28 high resolution crystal structures of protein–ligand complexes[19] from the protein data bank (PDB)[20] and have compared them with interactions found in small molecule crystal structures from the Cambridge structural database (CSD).[21] It was found from this study that both strong and weak hydrogen bonds are involved in ligand binding. We suggested that because of the prevalence of extensive multifurcation, the restrictive geometrical criteria set up for hydrogen bonds in small molecule crystal structures may need to be relaxed in macromolecular structures. We also showed that the formation of C—H···O hydrogen bonds is enhanced by the activation of the C_α—H atoms and by the flexibility of the side chain atoms. In contrast to small-molecule structures, anticooperative geometries were found to be common in the 28 macromolecular structures studied by us and there is a

gradual lengthening as the extent of furcation increases. Among other conclusions of our study, it was found that C—H···O bonds formed by Gly, Phe, and Tyr residues are noteworthy. Also, the numbers of hydrogen bond donors and acceptors agree with Lipinski's rule-of-five that predicts drug-like properties. Hydrogen bonds formed by water were also seen to be relevant in ligand binding and ligand C—H···O_w interactions are abundant when compared to N—H···O_w and O—H···O_w.

Among the limitations of our earlier study was the fact that the number of crystal structures examined was relatively small, only 28. Furthermore, several proteins in this group of structures are homologous to each other and so the number of truly independent observations is even smaller. We felt therefore that the conclusions of our previous study should be re-evaluated with a larger set of structures. This would necessitate the use of a computer program to evaluate hydrogen bond geometries (the smaller set of 28 crystal structures was analyzed manually). In the present work, a large dataset of 251 protein–ligand complexes, from the PDB, was analyzed with respect to intermolecular interactions with a new in-house computer program, hydrogen bond analysis tool (HBAT). The analysis was further extended to an external test set of 233 X-ray crystal structures of protein–ligand complexes in the kinase family. As an extension of our previous study, we also examined other new interactions involving halogen atoms (both as electrophiles and nucleophiles), π-acceptors and sulphur-atom acceptors. Eventually, we hope to evaluate the extent to which these weak interactions (hydrogen bonds and others) influence the protein–ligand interface in terms of behavior and function. The initial aim in this study is to document these interactions reliably in a representative group of protein–ligand crystal structures.

MATERIALS AND METHODS

A set of 251 X-ray crystal structures of protein–ligand complexes from the PDB was used in this study. The dataset comprises 27 structures from our earlier study and 224 from Nissink et al.[22] Because of the wrong assignment of ligand, PDB ID 1G2Y.pdb from our previous study is not included in the present study. The external kinase test set of 233 protein–ligand complexes was taken from the PDB [supplementary material (SM) Table VIII]. This was done to assess the general applicability of important conclusions derived from the analysis of the 251 X-ray crystal structures. In general, the available macromolecular crystallographic data are prone to two types of errors: (1) systematic errors caused by biases during the structure determination and refinement procedure, and (2) random errors which affect the precision of the model. Additionally, the quality of the structure varies in different regions, due to higher local conformational and thermal disorder in certain parts.[23] In this respect, the present dataset is free from any major abnormality. The active site was defined by selecting amino acid residues within a 10 Å radius of the ligand molecule.
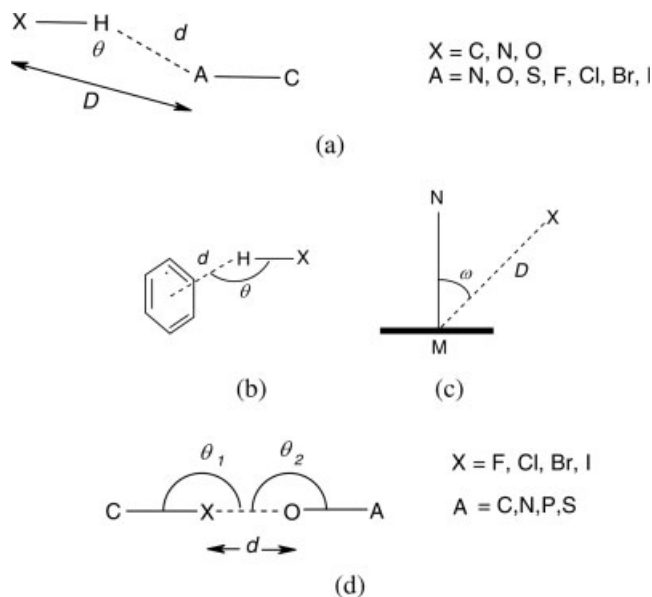

Scheme 1. (**a**) Representative hydrogen bond. A–C is a single or double bond; (**b**) and (**c**) parameters for X—H···π interactions; (**d**) parameters for halogen···O interactions. O–A is a double bond.

The standard H-bonding criteria were set as d (H···A) ≤ 3.0 Å and θ (X—H···A) $\geq 90°$. For other weak interactions, the criteria are mentioned in the respective sections. A schematic description of the various interactions is given in Scheme 1.

Macromolecular crystal structures rarely contain H-atom positional data with the precision required to properly evaluate hydrogen bond geometry. Therefore a method must be found to add or modify all the H-atom positions. H-atoms were added to the protein, water and ligand using the program MOE.[24] The H-atom positions were then refined (energy minimization) keeping the position of the non-H atoms fixed using the MMFF94x force field. It is important to note here the basis of selecting this force field for optimization of the H-atom positions. Initially four different types of force field, namely CHARMM22, AMBER96, OPLSAA, and MMFF94x, were used to derive standard hydrogen bond geometries (d and θ) in the 28 crystal structures studied by us earlier.[24,25] It was found that MMFF94x outperformed all other force fields with respect to optimization of the protein geometry. Programs like REDUCE from the Richardson group and HGEN from the CCP4 package could have been used for protein H-atoms but these programs are not efficient in generating H-atoms for the ligand and water.[26,27]

The calculated H-atom positions (MMFF94x optimized) were benchmarked against an experimental neutron crystal structure, namely 6RSA.pdb. In addition to the assessment of protein–ligand geometries, this benchmarking was also expected to be useful in the validation of H-atom positions in the water molecules. Fixing H-atom positions in water molecules has been a long-standing problem in macromolecular crystallography and in this regard we
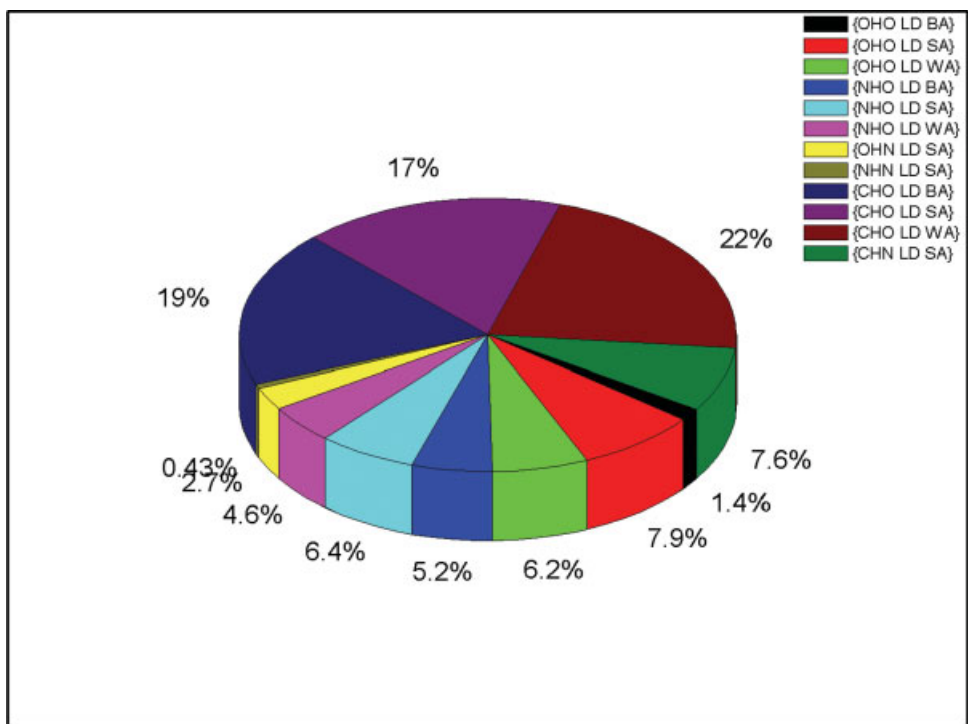
(a)
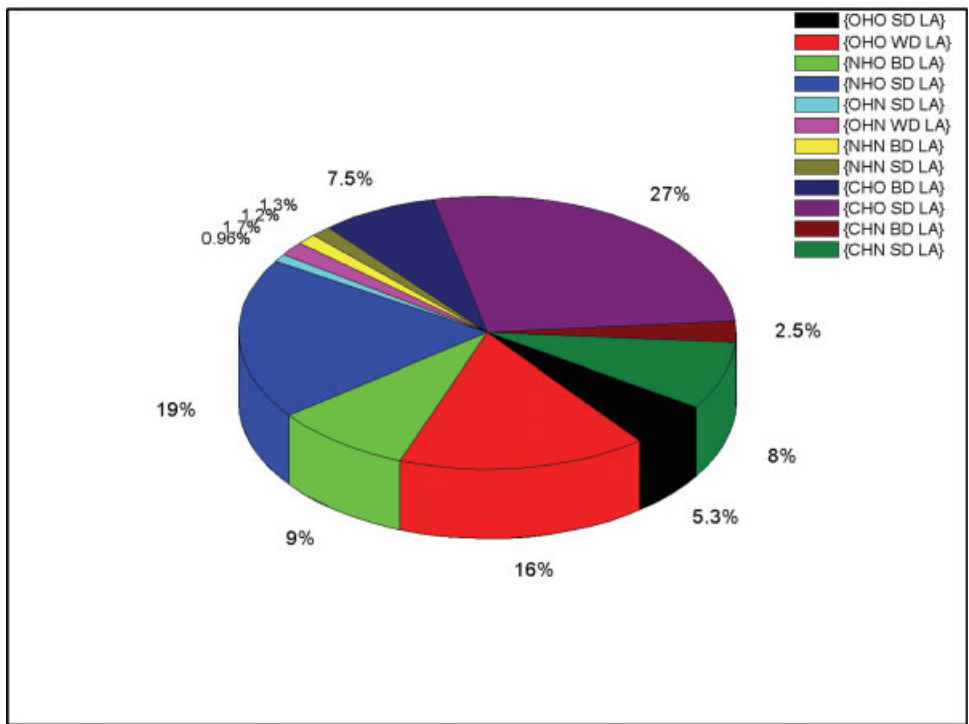
(b)

Fig. 1.   Distribution of all possible of hydrogen bond types in the active sites of protein–ligand complexes (**a**) ligand as donor (**b**) ligand as acceptor.
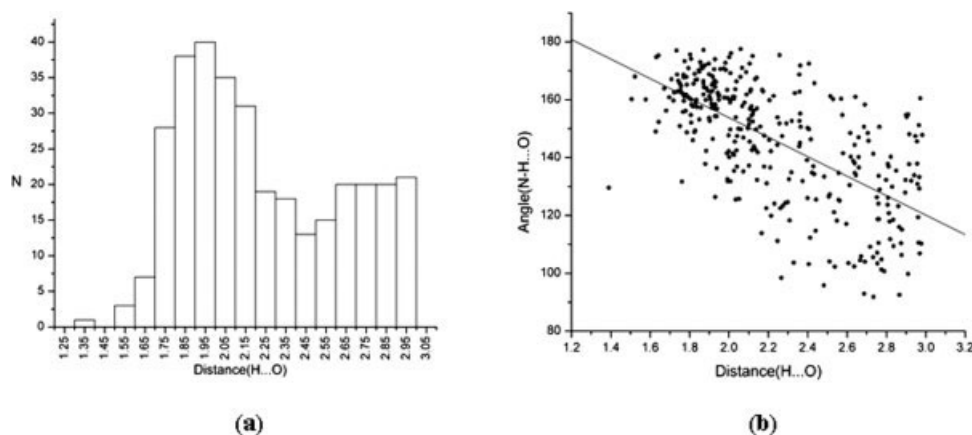
Fig. 2. Distance H···O histogram and d–θ scatterplot ($R^2 = -0.65$) for {NHO BD LA} hydrogen bonds in the active sites of the 251 protein–ligand complexes considered in this study.

believe that the use of MMFF94x achieves reasonable accuracy and is of general utility.

The large number of structures and more personalized requirements led us to design our own in-house software (HBAT) to carry out this study. There exist many programs for interaction analysis like HBPLUS, HBEXPLORE, CONTACT from CCP4 and web-based servers like LPC and NCI in the public domain.[27–31] These programs did not fit our requirements for several reasons, notably the fact that weak interactions like C—H···O and C—H···π are not considered. The advantage of HBAT over the above-mentioned programs is its compactness in delivering all possible interactions in a single package, thus avoiding server-based applications. Among other advantages, an MSOFFICE Excel compatible output file for statistical analysis provides distance–angle distributions across various geometry ranges, while tabulation of frequencies for each residue, ligand, water, and also nucleic acids can be done easily for any kind of interactions. The program is written using PERL and TK languages. It is a user-friendly desktop tool, which offers the freedom to choose several parameters. A detailed description of the application and utility of the program is being submitted shortly. To evaluate the accuracy of our program, we used HBAT to reproduce the geometries in our earlier manual study of 28 complexes[19] and also in other recent articles.[32–35] We obtained excellent agreement. This exercise has given us enough confidence to carry out the present study of the 251 protein–ligand complexes.

RESULTS AND DISCUSSION

Hydrogen Bond Geometry: Lengths and Angles

In any database analysis, evaluation of large data sets provides a more unbiased identification of a chemical signal in the presence of crystallographic noise.[36,37] A data set consisting of more than 100 representatives is ideal for the study of interaction geometry in crystal structures.[38,39] However analysis of hydrogen bonds in macromolecules is still difficult and requires classification based on backbone,

side chain, ligand, and water. The geometries observed for these various situations could be different in terms of their lengths, angles, and scatter.[40] The involvement of many types of hydrogen bond donors and acceptors increases the overall complexity at the protein–ligand interface.

Considering all this, it was felt that a classification of hydrogen bonds based on the participating groups and/or residues would better address the geometrical issues. The interacting partners at the interface are protein, ligand and water. All are able to donate and accept hydrogen bonds. Further, the donors and acceptors fall into different classes based on the strength/weakness of the resulting hydrogen bonds. The percentage contribution of various types of hydrogen bonds in the total protein–ligand interfaces in our 251 crystal structures is shown in Figure 1(a,b). The hydrogen bond abbreviation consists of three parts: hydrogen bond type, donor, acceptor. B stands for backbone, S is side chain, L is ligand, W is water, D is donor, and A is acceptor. For example {NHO BD LA} signifies an N—H···O hydrogen bond involving a backbone N—H donor and a ligand O-atom acceptor.

When the ligand is a donor, the percentage of strong hydrogen bonds is 34% while that of weak hydrogen bonds is 65%. This is reversed when the ligand is the acceptor, with 54 and 46% strong and weak hydrogen bonds, respectively. These numbers are reasonable: not only is the number of strong donors in ligand small but ligands also generally have more acceptors than donors. The data also show that if there are more acceptors than can form hydrogen bonds with good donors from the ligand, donors from the amino acids and water are used. The population of interactions at the protein–ligand interface from the main chain and the side chain is 32 and 68% respectively. The types of strong hydrogen bonds observed in the protein–ligand interface are {NHO LD BA}, {NHO SD LA}, {OHO LD BA}, and {OHO SD LA}. The median H···O distances, d, in all the above cases are <2.0 Å and the hydrogen bonds are linear (see Fig. 2). (See SM.)

It was suggested by us earlier[19] based on PDB cone-corrected angular distributions, that there are small deviations from linearity for both N—H···O and O—H···O
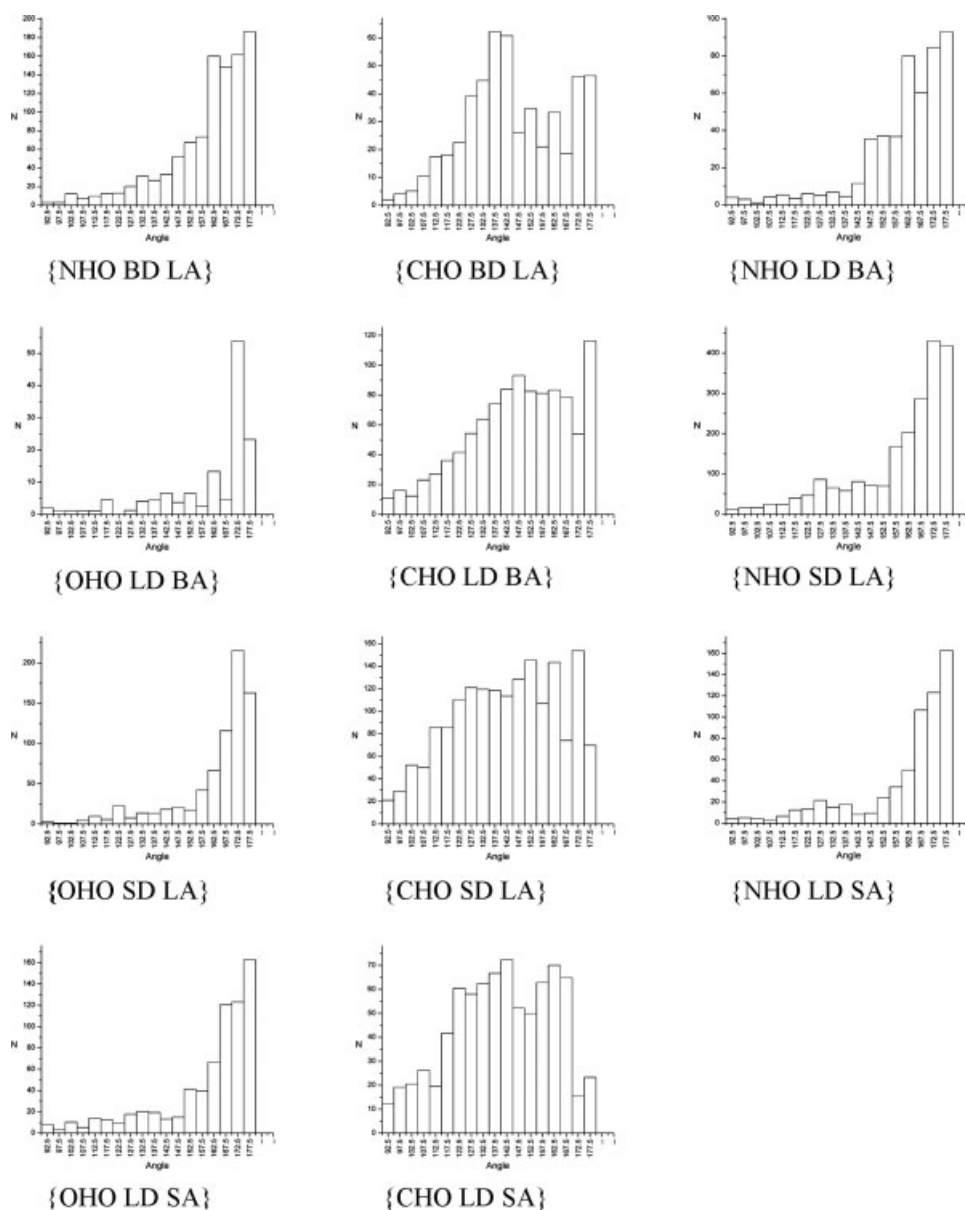
Fig. 3.   Histograms showing cone-corrected angular distribution for strong and weak hydrogen bonds in active sites of 251 protein–ligand complexes.

interactions. This observation does not hold good in the larger dataset of the present study. In both {NHO LD BA} and {NHO BD LA} cases, the cone-corrected angular maxima occur at 180°. The cone-corrected angular distributions for {OHO LD BA} and {OHO SD LA} are similar with maxima in the range 175–180°. The inverse length–angle correlations are also well behaved in all these cases (see Fig. 3). These observations are reassuring and show that the fundamental property of hydrogen bonds, namely linearity, holds by and large for all categories of strong hydrogen bond in macromolecular structures. Of course, the main chain hydrogen bond might be slightly more linear than the side chain interactions but, all in all, the geometries for strong hydrogen bonds observed in
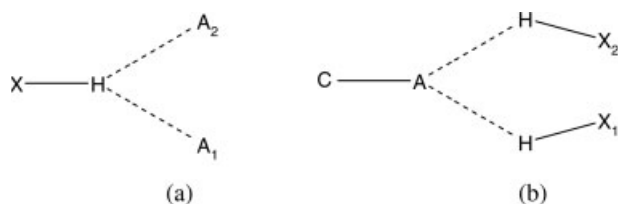
Scheme 2.   Notations for bifurcated hydrogen bonds.

protein–ligand interfaces are comparable to what is observed in small molecules. Baker and Hubbard have discussed hydrogen bond nonlinearity in their 1984 re-

TABLE I. Acceptor and Donor Furcations in the Active Sites of 251 Protein–Ligand Complexes

Furcation level	Furcated acceptors in active site			Furcated donors in active site		
	In protein and water	In ligand	Total	In protein and water	In ligand	Total
Bifurcated	12316	438	12754	21452	541	21993
Trifurcated	8672	321	8993	7852	160	8012
Tetrafurcated	6545	211	6756	2383	54	2437
Pentafurcated	3210	94	3304	605	10	615
Hexafurcated	1191	40	1231	235	7	242
Total	31934	1104	33038	32527	772	33299
Nonfurcated	17076	511	17587	41681	1321	43002

TABLE II. Strong and Weak Hydrogen Bonds for Ligands at Various Levels of Donor and Acceptor Furcation

Furcation level	Furcated acceptor			Furcated donor		
	O—H⋯O	N—H⋯O	C—H⋯O	O—H⋯O	N—H⋯O	C—H⋯O
Bifurcated	158	193	283	193	157	562
Trifurcated	175	265	337	68	103	238
Tetrafurcated	176	258	350	7	79	104
Pentafurcated	94	154	192	–	14	22
Hexafurcated	32	97	99	1	21	–
Nonfurcated	127	55	103	178	131	786

view.[6] However, and based on the present observations, we can assert that O—H⋯O and N—H⋯O hydrogen bonds tend to linearity in all macromolecular crystal structures.

The C—H⋯O interactions include {CHO BD LA}, {CHO LD BA}, {CHO SD LA}, and {CHO LD SA}. For {CHO BD LA} the angle distribution has a maximum at 170–180°. There is another maximum at 135–150° which corresponds generally to multifurcated geometries. Also similar is {CHO LD BA} with a maximum of 175–180°. For {CHO SD LA} the maximum is still around 170–175°. For {CHO LD SA} the favored angle is around 140–145° with a narrow range of linearity. Unlike strong hydrogen bonds, the cone-corrected weak C—H⋯O geometries show two distinct maxima at 130–150° and 170–180°. While the weakest {CHO LD SA} have variable geometry, the metrics of the other C—H⋯O bonds are surprisingly consistent. This is especially true of bonds donated by main chain C—H groups (see Fig. 3). Also seen are C—H⋯O$_w$ bonds formed by main chain and side chain C—H groups to water as acceptor. These are discussed later.

Hydrogen Bond Geometry: Furcation

Hydrogen bond furcation is a ubiquitous phenomenon in macromolecular structures. A donor can interact with several acceptors simultaneously or an acceptor can interact simultaneously with many donors. The terms bifurcated and trifurcated are commonly used to describe these arrangements. A bifurcated geometry can also be termed three-centered, and a trifurcated geometry can be termed four-centered (Scheme 2). In this study, the H-bonding criteria for furcated geometries were set as $d \leq 3.0$ Å and $\theta \geq 90°$. These furcated geometries (bifurcated, trifurcated, etc.) constitute independent sets in the sense that the trifurcated geometries do not implicitly include the bifurcated ones and so on. Multifurcation was first discussed in the 1960s and 1970s, in small molecule crystal structures. In the modern context, furcation would include all kinds of hydrogen bonds, strong (O—H⋯O, N—H⋯O) and weak (C—H⋯O) and the general idea is that the weak hydrogen bonds fill out or complete the hydrogen bond potential of an acceptor (or donor) which has a small number of strong hydrogen bonds.[41]

In the active sites of protein–ligand complexes, the level of furcation ranges from bifurcated to hexafurcated. Table I shows that donor and acceptor furcation occur roughly to the same extent (33,299 furcated donors and 33,038 furcated acceptors in the entire data set). This conveys that furcation is an inherent characteristic of macromolecular crystal structures. It does not arise—as it generally does in small molecule crystal structures—because of a donor–acceptor imbalance. If the analysis is restricted to the ligand, the frequency of furcated acceptors (1104) is more than that of furcated donors (772). This is in accordance with the fact that there are more acceptor atoms in ligands than donor atoms. It could also be due to steric reasons. Furcation levels higher than three are possible in principle, but are rarely found in practice because they require very high spatial densities of atoms and groups, especially when the donor is furcated. The above are number of furcated donors and acceptors. The total number of interactions is naturally much higher. Also notable are the numbers of nonfurcated geometries. Nonfurcated donors are more numerous than nonfurcated acceptors because: (1) there are more donors overall (C—H is included as a donor) and (2) acceptors are furcated more easily than donors for steric reasons.

We emphasize that C—H⋯O interactions are more common than the strong N—H⋯O and O—H⋯O hydrogen bonds in the furcated geometries (Table II). These
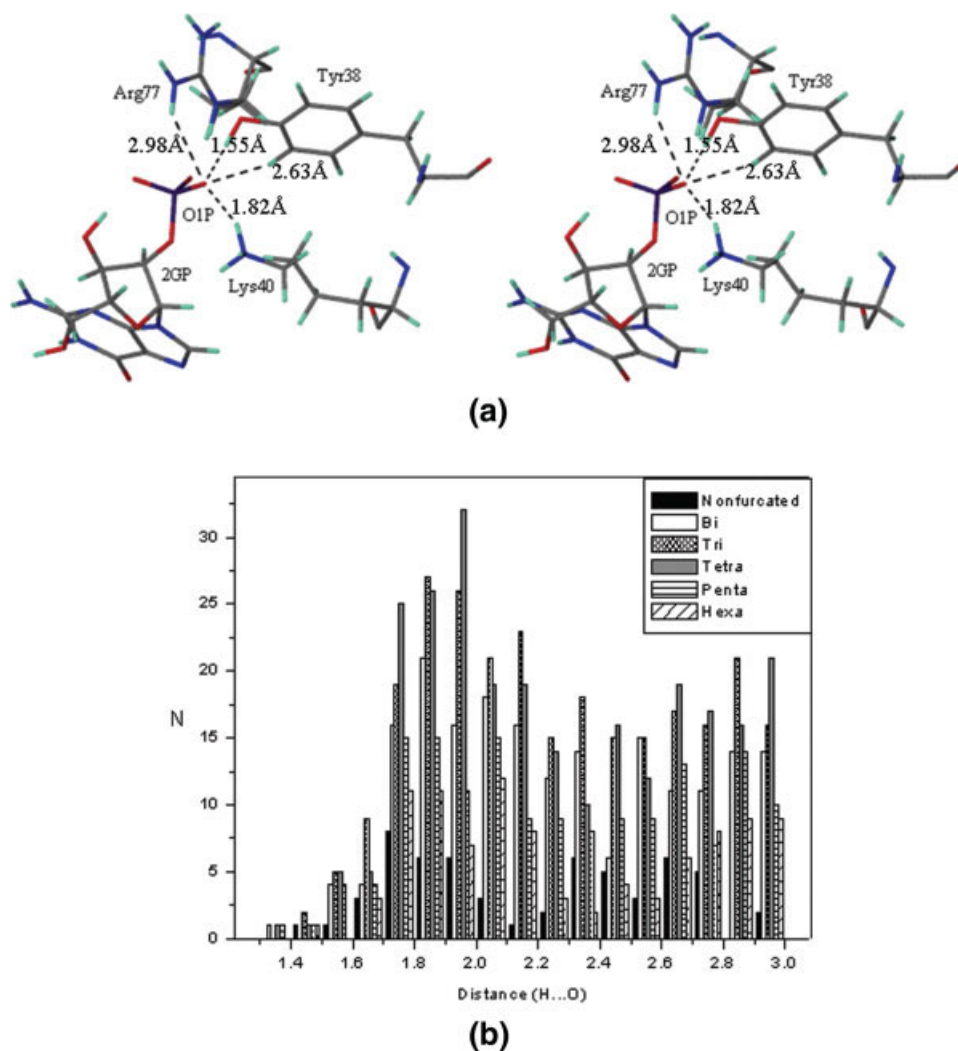
Fig. 4. (**a**) Tetrafurcated interaction in 2AAD.pdb. The acceptor centre is the O1P atom of the ligand (2GP). O1P interacts with residues Tyr38, Lys40 and Arg77 through C—H···O and O—H···O and N—H···O bonds. The respective H···O distances are also shown. (**b**) Histogram of the distance d of N—H···O interactions to furcated ligand acceptors. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

ideas have been noted by us in our earlier study of 28 protein–ligand complexes.[19] It is less likely (for electrostatic and statistical reasons) that a strong interaction like O—H···O occurs repeatedly in a furcated interaction. Instead, strong interactions tend towards non-furcated geometries while weak interactions occur in furcated situations. The overall message is that both interaction strength and close packing are important. A furcated geometry typically has one or a small number of strong interactions and many weaker interactions. This optimizes both interaction geometry and efficiency of space-filling. Table II shows that the total number of hydrogen bonds (O—H···O, N—H···O, C—H···O) to the 1104 furcated acceptors in Table I is 2863. This corresponds to an average level of furcation of 2.6 interactions to each furcated acceptor in the active site.

In summary, furcation occurs for both donor and acceptor sites on ligands. Acceptor furcation is more common than donor furcation and this could be due to steric reasons. The majority of furcated interactions exhibit longer d (H···O) distances than the simple non-furcated hydrogen bonds and this is as might have been expected [Fig. 4(a,b)].

Hydrogen Bond Geometry: The Resolution Problem

The resolution in a macromolecular crystal structure determination has a direct effect on the geometry of both strong and weak interactions. The issue of reliability of hydrogen bond metrics as a function of crystallographic resolution is interesting, and could not be addressed in our earlier study, which considered only 28 crystal structures,
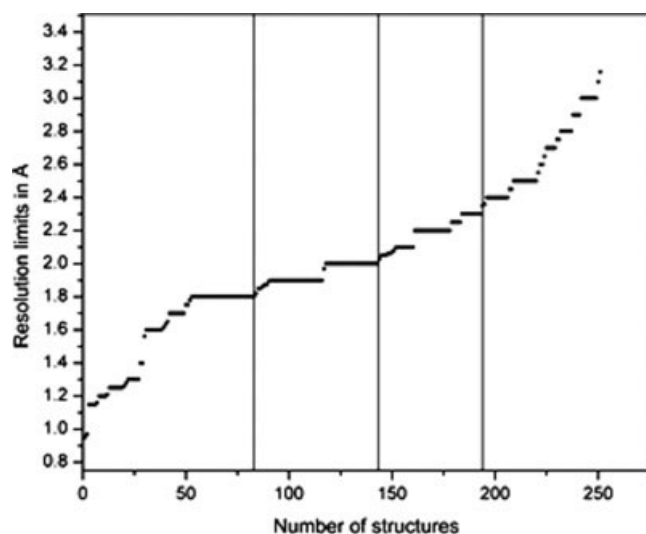
Fig. 5. Distribution of crystal structures as a function of resolution.

and all of very good resolution. One of the aims of the present study is to identify the resolutions limits for macromolecule crystal structures where the hydrogen bond geometries are as reliable as that obtained in small molecule structures. Therefore we classified our 251 structures on the basis of resolution. The categories chosen were as follows: resolution <1.8 Å (83 structures); resolution 1.8–2.0 Å (60 structures); resolution 2.0–2.3 Å (50 structures); resolution >2.3 Å (58 structures). Figure 5 gives this information pictorially and Table III contains a list of structures.

As prototypes of strong and weak hydrogen bonds, {NHO BD LA}, {NHO SD LA} and {CHO BD LA}, {CHO SD LA} are described here for comparisons of hydrogen bond geometry as a function of resolution. The d–θ scatterplots (cone-corrected) were analyzed for these prototype interactions (SM, Fig. 11). From a visual inspection of these plots, it was concluded that strong hydrogen bonds {NHO BD LA} and {NHO SD LA} retain acceptable geometries till a resolution of 2.3 Å, whereas for {CHO BD LA} and {CHO SD LA} the threshold is 2.0 Å (see Fig. 6), with bond geometries beyond these limits being poor. Accordingly, crystal structures within a resolution of 2.3 Å may be safely considered for strong hydrogen bonds like N—H···O. For C—H···O the corresponding limit is 2.0 Å and signs of non-linearity are observed above this. (SM, Fig. 11)

Residue Frequency

The percentage distributions of various residues in the total protein and in the active site are comparable for nonpolar and polar amino acids. (SM, Table VI) Residues like Gly, Ile, Phe, Tyr are among the most common residues. For {NHO BD LA} Gly is the major nonpolar donor, while Ser and Thr are the major polar donors. The percentage of charged residues in the active site (20%) is slightly less than in the overall protein (24%) with Asp and Arg being the major charged donors (see Fig. 7) and this is in keeping with the hydrophobic nature of the pro-

tein interior. This result is similar to what we observed earlier in the smaller set of 28 structures[19] except that among the charged residues, Asp is now the major donor instead of Lys. For {CHO BD LA}, Gly, Ser and Tyr, and His are dominant as donors in the three respective classes of amino acids. These observations provide a hint that residues which are smaller in size and have greater flexibility participate well in both strong and weak hydrogen bonds. Sometimes a so-called weak donor like C_α—H can behave like a strong donor in the presence of a charged side chain; this is observed for His.

Residues like Trp, Asn and Arg are the major donors in the {NHO SD LA} category. However, Lys and His also interact equally well with the ligand through their side chains. The major participation of charged residues here suggests that ligand binding is dominated by electrostatic interactions. At this point, it is interesting to discuss the donor capability of Gly. For Gly the total number of {NHO BD LA} and {CHO BD LA} interactions are 89 and 113, and this is the highest among all amino acids. We ascribe this fact to Gly being the smallest residue in terms of volume (63.8 Å^3) and also occurring most frequently in the active site (9.63%). This result reaffirms our previous observations and suggests that the highly flexible nature of Gly is very well exploited in biological recognition.[42,43]

For N—H···O bonds with N—H side chain donors, the major donor residues are Trp (nonpolar) Asn (polar) and Arg (charged). The number of interactions for {OHO SD LA} polar residues follows the order Ser > Tyr > Thr. The alkyl hydroxyl present in Ser has less steric hindrance than in Thr and is therefore more commonly used. For nonpolar residues, the numbers of {CHO SD LA} interactions are greater for Phe and Trp. Similarly, Tyr, Arg and His are the major amino acids in the other categories participating in {CHO SD LA} interactions. An interesting case in this category is the side chain phenyl ring donor capacity of Tyr, wherein a large number of C—H donors are present along with a strong O—H donor. The total number of C—H···O and O—H···O interactions exhibited are 87 and 52, respectively. To explain this phenomenon an ab initio calculation between Tyr side chain and water was carried out at the 6-31G** basis set level. Two minima were obtained corresponding to Complex I and Complex II (see Fig. 8). The energies for the O—H···O and C—H···O interactions in Complex I are −27.96 and −7.65 kJ/mol. For Complex II, the values are −39.45 and −2.38 kJ/mol, respectively.

Interactions Involving Water

The hydrogen bonding capacity of water makes it easy to interact with protein, ligand or neighboring water molecules. The total number of hydrogen bonds (N—H···O, O—H···O, C—H···O) formed by water (as donor or acceptor) in the active sites of the 251 complexes under consideration is 29,718, in other words these are nearly 118 such hydrogen bonds for each structure, on average. The frequency of occurrence of these bonds follows the order, side chain > main chain > water > ligand. Table IV

TABLE III. List of Entries of PDB Structures Based on Four Resolution Limits

Resolution < 1.8 Å (83 entries)

1BXO	8A3H	1A6G	1HDO	1HET	1RGE	1MRO	1C0P	1I3H	1I76
2NLR	1D4O	1C1D	1EQO	1OAA	1Q0N	2TPS	2WEA	3CHB	1CTQ
1QKS	1DJR	1FCY	1FK5	1HYO	1I12	1RUV	1TPP	2CTC	1HFC
1AOE	1CIL	1LIC	1MRK	1PHD	1PHG	1ROB	2TMN	1C5C	2CPP
1SNC	1ABE	1B17	1ETA	1HYT	1IDA	1XID	1XIE	4DFR	1C5X
3CLA	4EST	1A28	1A6W	1APT	1APU	1AQW	1ATL	1B58	1B59
1BMA	1C83	1CBS	1COY	1D3H	1EJN	1GLQ	1HVR	1LST	1MRG
1NCO	1PPC	1QBR	1QBU	1RDS	1SRJ	1TNG	1TNH	1WAP	2FOX
2QWK	5ABP	6RNT							

Resolution 1.8–2.0 Å (60 entries)

1JAP	1FLR	2AK3	1AEC	1F3D	2CMD	1HSL	1A4Q	1ABF	1BYB
1FEN	1GLP	1HPV	1HSB	1KEL	1LNA	1MLD	1MMQ	1MTS	1PPH
1RNT	1SLT	1TMN	1TNI	1TNL	1TYL	1UKZ	1VGC	2GBP	2H4N
3ERT	3TPI	7TIM	2TSC	1AZM	1BBP	1CBX	1CDG	1CLE	1DO1
1DG5	1DMP	1EED	1EIL	1EPO	1FKG	1FRP	1HIV	1MBI	1POC
1PSO	1QCF	1QPE	1TRK	25C8	2AAD	2IFB	3CPA	5ER1	6RSA

Resolution 2.0–2.3 Å (50 entries)

3ERD	1ACO	1CKP	1NIS	1LAH	1LCP	1D4P	5CPP	1DY9	1EBG
1F0R	1F0S	1LDM	1MDR	1OKL	2YHX	4PHV	1A4G	1CL2	1COM
1DR1	1EPB	1ETR	1FKI	1HDC	1HTF	1IMB	1LPM	1OKM	1PDZ
1PPI	1YDR	1YEE	2CHT	2PCP	1A42	1CPS	1DD7	1EOC	2PK4
1BGO	1BLH	1DHF	1ETS	1HOS	1PBD	1PTV	1TLP	1YDT	3GPB

Resolution > 2.3 Å (58 entries)

1B9V	1AI5	1BYG	1CVU	1DOG	1IVB	1MUP	1NGP	1RNE	2ACK
2ADA	4AAH	4LBD	1FL3	1QPQ	1A9U	1AAQ	1ASE	1BMQ	1DID
1EAP	1ETT	1FGI	1LYB	1UVT	2YPI	4FBP	1RT2	1CQP	1IVQ
1TDB	1DBB	1DBJ	1IBG	1MCQ	2PHH	1BKO	1ULB	1ACJ	1ACL
1ACM	1LYL	1UVS	2LGS	1BAF	2DBL	3HVT	4CTS	1CX2	1DWC
1DWD	1FAX	1HAK	1HRI	2RO7	4COX	2MCP	1DWB		

shows that there are many more C—H···O_w interactions than for N—H···O_w and O—H···O_w interactions. A small number (15.29%) of hydrogen bonds are of the type O_w—H···O_w.

Among the 1503 ligand–water interactions, {CHO LD WA} interactions constitute as many as 66%, while {OHO LD WA} and {NHO LD WA} account for only 19 and 13%, respectively. A similar trend was observed in our previous study.[19] For {OHO WD LA} the number of interactions observed is 592 of which 53% are below a H···O distance of 2.2 Å. (SM, Table VII) This is in contrast to our earlier study, where the maximum number (about 66%) of O_w—H···O hydrogen bonds were observed in the range of 2.2–2.7 Å. A possible reason for this discrepancy could be the inaccuracy in the H-atom addition method for water in our previous study. A total of 8624 water molecules are present in the 251 active sites. For these, the average coordination number is 2.1, if only strong hydrogen bonds are considered. If weak interactions are added, then this number rises to 3.4 per water molecule, which is in good agreement with the earlier study of Steiner[44]; it is also chemically reasonable.

Lipinski's Rule Extended

The numbers of hydrogen bond donors and acceptors are known to affect the physico-chemical properties (solubility, adsorption, distribution) of a molecule and hence the efficacy of a drug. Lipinski's rule-of-five states that for better permeation and absorption, the number of donors and acceptors in a ligand should be less than 5 and 10, respectively.[45] The present dataset contains 302 ligands. The total number of strong hydrogen bonds made by the various ligands with protein and water molecules stand at 973 (ligand as donor) and 2001 (ligand as acceptor). Accordingly, each ligand has 3.2 donors and 6.6 acceptors on average. This figure satisfies Lipinski's rule-of-five for hydrogen bond donors and acceptors in that the number of acceptors present per ligand is around twice the number of donors.

Protein–Ligand Interactions in Kinases

An external test set of 233 protein–ligand complexes of various kinases was compiled from the PDB (SM Table VIII) to assess the general applicability of important conclusions derived so far. The nature of strong and weak hydrogen bonds (linearity of hydrogen bonds), prevalence of multifurcated interactions, distinctiveness of interaction patterns as a function of resolution and other attributes were analyzed in this dataset (SM. Fig. 13, Table VIII–X). These results suggest a similar trend for hydrogen bond geometries and constitute a useful validation of the principles enunciated in this analysis.
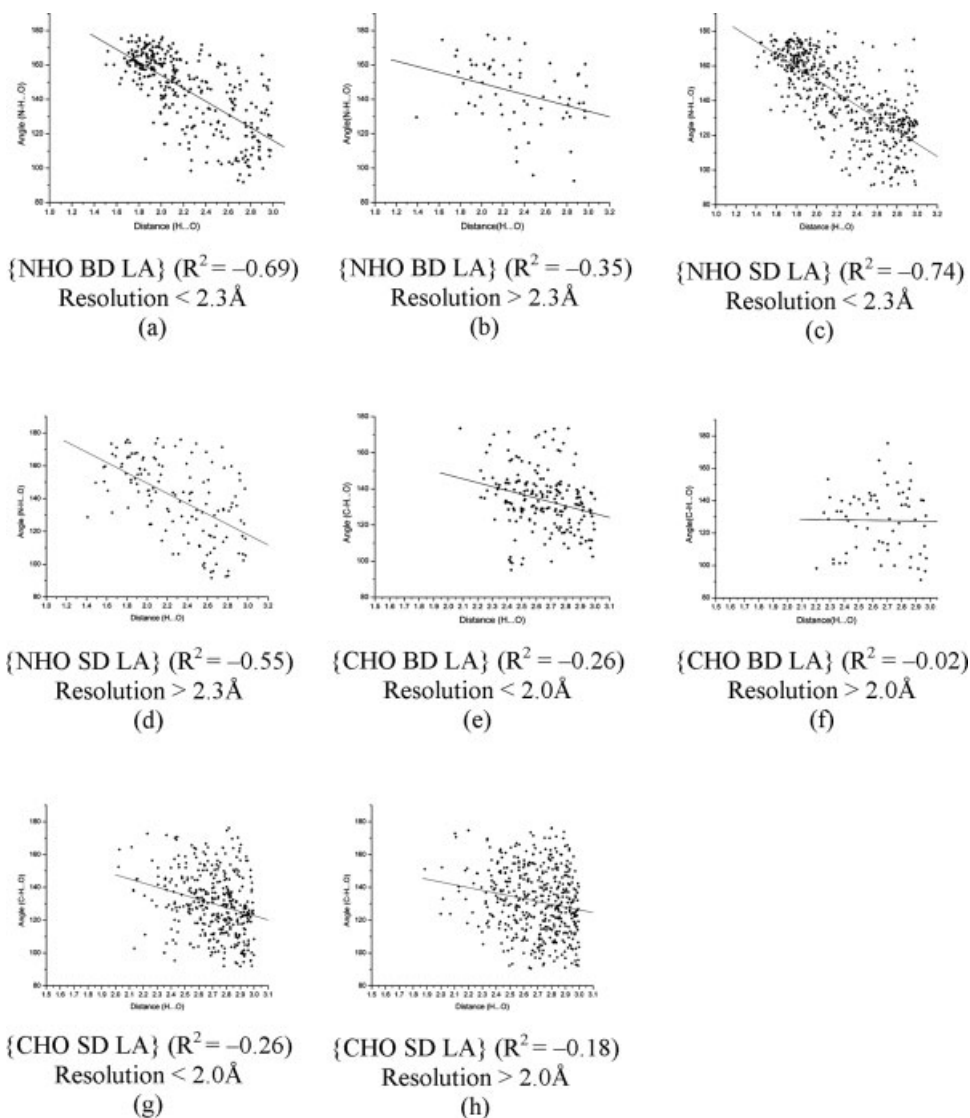
Fig. 6. d–θ Scatterplots for {NHO BD LA}, {NHO SD LA} (**a**)–(**d**), and {CHO BD LA}, {CHO SD LA} (**e**)–(**h**). Notice the different appearance of these two sets of plots below and above the threshold resolution limits.

Other Weak Interactions
X—H···π hydrogen bonds to amino acid residues

The most common π-acceptors in proteins are the side chains of Phe, Tyr, Trp and occasionally His.[13,32] In our study, the ligand π-acceptors have not been taken into account and so the results are restricted to π-acceptors in the side chains of Phe, Tyr, Trp and His. But this is not a serious limitation. For Trp, the five membered and six-membered rings were treated separately. In the present study, we have adopted the convention of Scheme 1(b,c) to locate ligand–π interactions. It is difficult to derive an ideal geometry for these interactions to multi-atom acceptors. However $d \leq 3.5$ Å, $\theta \geq 100°$ and $\omega \leq 40°$ appear to be satisfactory and this geometric criterion is generally accepted.[13] In the present study we have observed 4 N—H···π (2 each to Trp and Tyr), 3 O—H···π (2 to His and 1 to Trp), and 159 C—H···π interactions. For the 159

C—H···π interactions, the acceptor frequency is Trp (41%), Tyr (28%), Phe (14%) and His (17%). The percentage occurrences of these residues in the active sites are Trp (2.39%), Tyr (5.18%), Phe (5.11%), and His (2.83%). The high frequency of C—H···π bonds to Trp is accounted for by the larger accessible areas afforded by the two fused rings. The doubling of the number of interactions to Tyr when compared with Phe, two residues that occur nearly equally in the active sites, is nicely accounted for by the increased acceptor capability of the Tyr aromatic ring.

Halogen bonds

Short oxygen···halogen interactions have been known since the 1950s.[46] A recent survey of protein and nucleic acid structures reveals similar halogen bonds as potentially stabilizing inter- and intra-molecular interactions that can affect ligand binding.[34] A typical halogen bond is
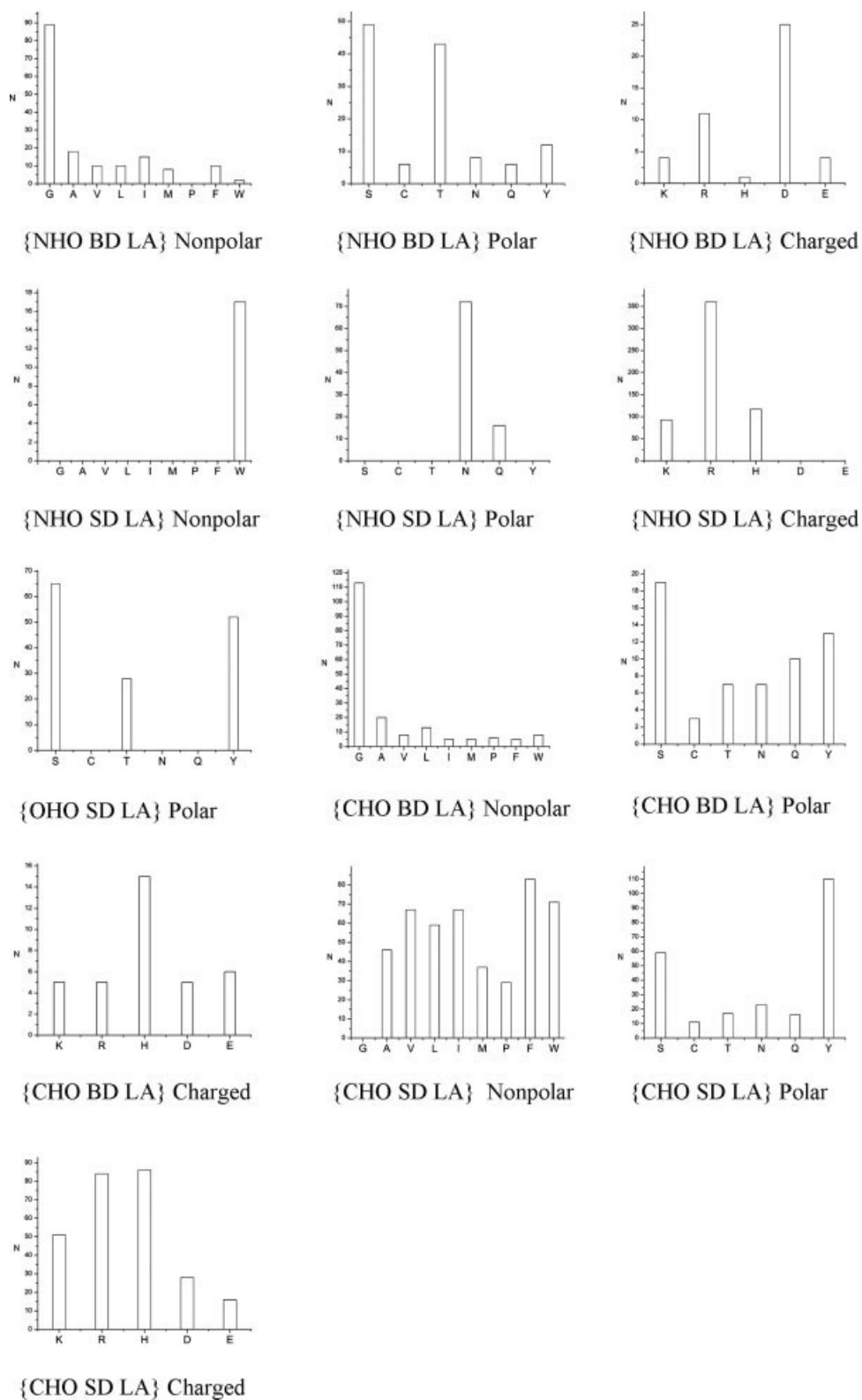
Fig. 7.   Frequency distribution of nonpolar, polar, and charged amino acids for strong and weak hydrogen bonds in the protein–ligand interface.
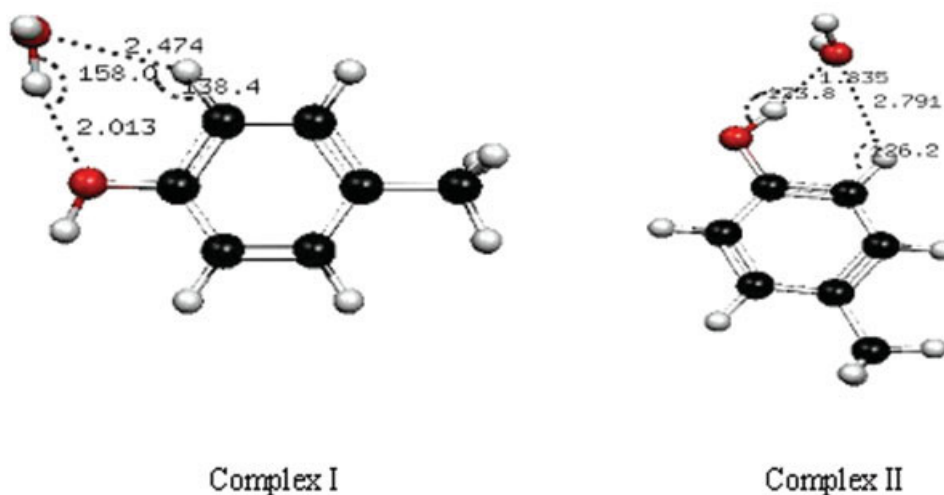
Complex I   Complex II

Fig. 8. Two possible conformations of the interaction of the Tyr side chain with water. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

TABLE IV. Percentage Distribution of Various Bond Types Among the 29,718 Hydrogen Bonds Formed by Water in the Active Site of 251 Protein–Ligand Complexes

	Type	%
1	{NHO LD WA}	0.42
2	{OHO LD WA}	0.58
3	{CHO LD WA}	2.05
4	{OHO WD LA}	1.99
5	{NHO BD WA}	8.43
6	{CHO BD WA}	8.84
7	{OHO SD WA}	3.99
8	{NHO SD WA}	10.49
9	{CHO SD WA}	25.48
10	{OHO WD WA}	15.29
11	{OHO WD BA}	12.80
12	{OHO WD SA}	9.59

represented by Scheme 1(d). Protein–ligand complexes were analyzed for possible halogen bonds following the van der Waals radii criterion. We found seven halogen bonds in the following structures: 1CKP.pdb, 1CLE.pdb, 1BMA.pdb, 4EST.pdb, 2NLR.pdb, 1ETA.pdb listed in Table V. These interactions were observed between C—F, C—Cl and C—I and carbonyl O-atoms in the main chain.

Halogen as nucleophile

The acceptor capability of organic halogen, X (X = F, Cl, Br, I), has not been studied in detail in macromolecules.[47] While these interactions are weak they seem to play a definite role in protein–ligand stabilization when halogenated ligands are present. The number of X—H···halogen (here X = O, N, C) interactions for O—H, N—H, and C—H donors is 5, 12, and 35, respectively. Almost all O—H···Cl interactions are observed between water and ligand. For C—H···Cl and C—H···F interactions side chain C—H groups are frequently used.

Hydrogen bonds involving sulphur atoms

Sulphur atoms are larger and have a more diffuse electron cloud than oxygen and nitrogen, but are nevertheless capable of participating in hydrogen bonds.[48] We have studied the acceptor functionality of sulphur atoms. Sulphur is present in amino acid residues like Met and Cys or it may occur in the ligand. In all these situations, a hydrogen bond is presumed to exist if the distance d (H···S) is ≤ 2.9 Å. The numbers of such cases are 12, 15, and 24 for O—H, N—H and C—H respectively with these donors belonging to either ligand, protein or water. For hydrogen bonds of the type O—H···S and N—H···S the acceptor is found more often in the ligand than in the protein.

CONCLUSIONS

The nature of strong (O—H···O, N—H···O) and weak (C—H···O) hydrogen bonds in the protein–ligand interface has been studied in a dataset of 251 protein–ligand complexes using a new in-house computer program (HBAT). Reasonable accuracy in locating hydrogen atoms positions in these complexes were achieved using the MMFF94x force field in the MOE software. The fundamental property of hydrogen bonds, namely linearity, holds by and large for all strong hydrogen bonds in these structures. Strong hydrogen bonds have more consistent distance and angle attributes, while the weak C—H···O interactions have variable geometry. Main chain hydrogen bonds are, in general, shorter and more linear than those formed by side chain donors and acceptors. Furcated ligand–receptor interactions are manifested by both donors and acceptors. Acceptor furcation is more common than donor furcation. The majority of furcated interactions exhibit longer d (H···O) distances when compared with simple non-furcated hydrogen bonds. Resolution limits are important with respect to the hydrogen bond geometry. Strong hydrogen bonds retain good geometries up to

TABLE V. Halogen Bonds Observed in the Protein–Ligand Interfaces is Tabulated Along With Respective PDB ID

	Type	Ligand, residue ID	Acceptor, residue ID	d	θ_1	θ_2	PDB ID
1	C—Cl···O=C	PVB, 1	Asp, 86	3.05	134	78.47	1CKP
2	C—Cl···O=C	ENH, 703	Thr, 96	2.98	147.7	74.22	1CLE
3	C—Cl···O=C	ENH, 703	Arg, 100	2.85	152.9	81.02	1CLE
4	C—F···O=C	TFA, 256	Cys, 199	2.91	133.9	108.2	1BMA
5	C—F···O=C	FPA, 5	Pro, 3	2.88	102.1	90.09	4EST
6	C—F···O=C	G2F, 603	Glu, 120	1.99	107.5	121.9	2NLR
7	C—I···O=C	T44, 128	Ala, 109	3.12	149.3	106.0	1ETA

a resolution of 2.3 Å, whereas for weak bonds the limit is 2.0 Å. Residues like Gly and Ala, which are smaller in size and have greater flexibility, participate well in both strong and weak hydrogen bonds. In this respect, Gly frequently interacts with the ligand. The side chain donor capacity of Tyr, with respect to both O—H···O and C—H···O interactions, is noteworthy. Other weak interactions involving halogen atom (both as electrophiles and nucleophiles), π-acceptors and sulphur atom acceptors are also important in the protein–ligand interface. Strong and weak hydrogen bonds involving water are ubiquitous in the active sites. Water is found to interact with amino acid residues and ligands forming O—H···O, N—H···O and C—H···O bonds. The hydrogen bond donor–acceptor ratio for the ligands is in accordance with Lipinski's rule-of-five. We conclude that the results of our previous study of 28 structures are largely applicable to a set of structures that is nearly 10 times as large. An encouraging aspect of this study is that macromolecular crystal structures with resolutions up to 2.0 Å may be used to analyze hydrogen bond geometry provided a reliable way is found to fix H-atom positions.

ACKNOWLEDGMENTS

REFERENCES

1. Glusker JP. Directional aspects of intermolecular interactions. In: Weber E, editor. Topics in current chemistry: design of organic solids. Berlin, Heidelberg: Springer-Verlag; 1998. pp 1–56.
2. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. J Mol Biol 2003;326:1239–1259.
3. Allocati N, Masulli M, Pietracupa M, Federici L, Di Ilio C. Evolutionarily conserved structural motifs in bacterial GST (glutathione-*S*-transferase) are involved in protein folding and stability. Biochem J 2006;394:11–17.
4. Shokhen M, Albeck A. Is there a weak H-bond → LBHB transition on tetrahedral complex formation in serine proteases? Proteins 2004;54:468–477.
5. Bledsoe RK, Madauss KP, Holt JA, Apolito CJ, Lambert MH, Pearce KH, Stanley TB, Stewart EL, Trump RP, Willson TM, Williams SP. A ligand-mediated hydrogen bond network required for the activation of the mineralocorticoid receptor. J Biol Chem 2005;280:31283–31293.
6. Baker EN, Hubbard RE. Hydrogen bonding in globular proteins. Prog Biophys Mol Biol 1984;44:97–179.
7. Jeffrey GA, Saenger W. Hydrogen bonding in biological structures. Berlin: Springer-Verlag; 1991.
8. Glusker JP. Intermolecular interactions around functional groups in crystals: data for modeling the binding of drugs to biological macromolecules. Acta Crystallogr Sect D 1995;51:418–427.
9. Wahl MC, Rao ST, Sundaralingam M. The structure of r(UUCGCG) has a 5′-UU-overhang exhibiting Hoogsteen-like trans U.U base pairs. Nat Struct Biol 1996;3:24–31.
10. Derewenda ZS, Derewenda U, Kobos PM. (His)C ε-H···O=C ⟨ hydrogen bond in the active sites of serine hydrolases. J Mol Biol 1994;241:83–93.
11. Fabiola GF, Krishnaswamy S, Nagarajan V, Pattabhi V. C—H···O hydrogen bonds in β sheets. Acta Crystallogr Sect D 1997;53:316–320.
12. Thomas A, Benhabiles N, Meurisse R, Ngwabije R, Brasseur R. Pex, analytical tools for PDB files. II. H-Pex: noncanonical H-bonds in α-helices. Proteins 2001;43:37–44.
13. Steiner T, Koellner G. Hydrogen bonds with π-acceptors in proteins: frequencies and role in stabilizing local 3D structures. J Mol Biol 2001;305:535–557.
14. Scheiner S, Kar T, Gu Y. Strength of the Cα H···O hydrogen bond of amino acid residues. J Biol Chem 2001;276:9832–9837.
15. Jiang L, Lai L. CH···O hydrogen bonds at protein–protein interfaces. J Biol Chem 2002;277:37732–37740.
16. Reddy CK, Das A, Jayaram B. Do water molecules mediate protein-DNA recognition. J Mol Biol 2001;314:619–632.
17. Prasad T, Prathima MN, Chandra N. Detection of hydrogen-bond signature patterns in protein families. Bioinformatics 2003;19:167–168.
18. Babu M, Singh SK, Balaram P. A C—H···O hydrogen bond stabilized polypeptide chain reversal motif at the C terminus of helices in proteins. J Mol Biol 2002;322:871–880.
19. Sarkhel S, Desiraju GR. N—H···O, O—H···O, and C—H···O hydrogen bonds in protein–ligand complexes: strong and weak interactions in molecular recognition. Proteins 2004;54:247–259.
20. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acid Res 2000;28:235–242. http://www.rcsb.org/pdb
21. Allen FH, Kennard O. 3D search and research using the Cambridge Structural Database. Chem Des Autom News 1993;8:31–37. http://www.ccdc.cam.ac.uk/prods/csd/csd.html
22. Nissink JW, Murray C, Hartshorn M, Verdonk ML, Cole JC, Taylor R. A new test set for validating predictions of protein–ligand interaction. Proteins 2002;49:457–471.
23. Kleywegt G. Validation of protein crystal structures. Acta Crystallogr Sect D 2000;56:249–265.
24. Chemical Computing Inc. MOE Program, Chemical Computing Inc., Montreal, Quebec, Canada.
25. Kollman P, Allinger N. Molecular mechanics aspect. In: Schleyer P, editor. Encyclopedia of computational chemistry. New York: Wiley; 1998.
26. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 1999;285:1735–1747.
27. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. Acta Crystallogr Sect D 1994;50:760–763.
28. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1994;238:777–793.
29. Lindauer K, Bendic C, Suhnel J. HBexplore—a new tool for identifying and analysing hydrogen bonding patterns in biological macromolecules. Comput Appl Biosci 1996;12:281–289.

30. Sobolev V, Sorokine A, Prilusky J, Abola E, Edelman M. Automated analysis of interatomic contacts in proteins. Bioinformatics 1999;15:327–332.
31. Babu M. NCI: a server to identify non canonical interactions in protein structures Nucleic Acid Res 2003;31:3345–3348.
32. Brandl M, Weiss MS, Jabs A, Sühnel J, Hilgenfeld R. C−H⋯π-interactions in proteins. J Mol Biol 2001;307:357–377.
33. Koellner G, Kryger G, Millard CB, Silman I, Sussman JL, Steiner T. Active-site gorge and buried water molecules in crystal structures of acetylcholinesterase from Torpedo californica. J Mol Biol 2000;296:713–735.
34. Auffinger P, Hays FA, Westhof E, Ho PS. Halogen bonds in biological molecules. Proc Natl Acad Sci USA 2004;101:16789–16794.
35. Klaholz B, Moras D. C−H⋯O hydrogen bonds in the nuclear receptor RARγ-a potential tool for drug selectivity. Structure 2002; 10:1197–1204.
36. Chakrabarti P, Chakrabarti S. C−H⋯O hydrogen bond involving proline residues in α-helices. J Mol Biol 1998;284:867–873.
37. Kim J, Mao J, Gunner MR. Are acidic and basic groups in buried proteins predicted to be ionized? J Mol Biol 2005;348:1283–1298.
38. Park S, Saven JG. Statistical and molecular dynamics studies of buried waters in globular proteins. Proteins 2005;60:450–463.
39. Ippolito JA, Alexander RS, Christianson DW. Hydrogen bond stereochemistry in protein structure and function. J Mol Biol 1990; 215:457–471.
40. Ragone R. Hydrogen-bonding classes in proteins and their contribution to the unfolding reaction. Protein Sci 2001;10:2075–2082.
41. Desiraju GR, Steiner T. The weak hydrogen bond in structural chemistry and biology. Oxford: Oxford University Press; 1999. pp 1–28.
42. Porter CT, Bartlett GJ, Thornton JM. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acid Res 2004;32:129–133.
43. Yan B, Sun Y. Glycine residues provide flexibility for enzyme active sites. J Biol Chem 1997;272:3190–3194.
44. Steiner T. Water molecules which apparently accept no hydrogen bonds are systematically involved in C−H⋯O interactions. Acta Crystallogr Sect D 1995;51:93–97.
45. Lipinski CA, Lombardo F, Dominy BW, Feeney PL. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 1997;23:3–25.
46. Ouvrard C, Le Questel JY, Berthelot M, Laurence C. Halogen-bond geometry: a crystallographic database investigation of dihalogen complexes. Acta Crystallogr Sect B 2003;59:512–526.
47. Kovács A, Varga Z. Halogen acceptors in hydrogen bonding. Coord Chem Rev 2006;250:710–727.
48. Gregoret LM, Rader SD, Fletterick RJ, Cohen FE. Hydrogen bonds involving sulfur atoms in proteins. Proteins 1991;9:99–107.