CHAPTER 3

# STRONG AND WEAK HYDROGEN BONDS IN THE PROTEIN–LIGAND INTERFACE

## 3.1 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 [3.1, 3.2]. Because of their weakness, however, they are also made and broken rapidly during complexation, conformational change and folding [3.3]. 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 facilitate 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 [3.4, 3.5].

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 [3.6, 3.7]. The subject was reviewed in depth by Glusker in 1995 [3.8]. Work by Sundaralingam on nucleic acids and Derewenda on globular proteins in the mid to late 1990's widened the scope of this field [3.9–3.11]. Since 2000, there have been a number of papers that have attempted to analyze the systematics of hydrogen bonds in biological structures [3.12–3.18].

Previously 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 [3.19] have been examined from the Protein Data Bank (PDB) [3.20]. These interactions were compared with the interactions found in small molecule crystal structures from the Cambridge Structural Database (CSD) [3.21]. Some of the important conclusions derived in this study are: (a) both strong and weak hydrogen bonds are involved in ligand binding, (b) the restrictive geometrical criteria set up for hydrogen bonds in small molecule crystal structures may need to be relaxed in macromolecular structures due to extensive

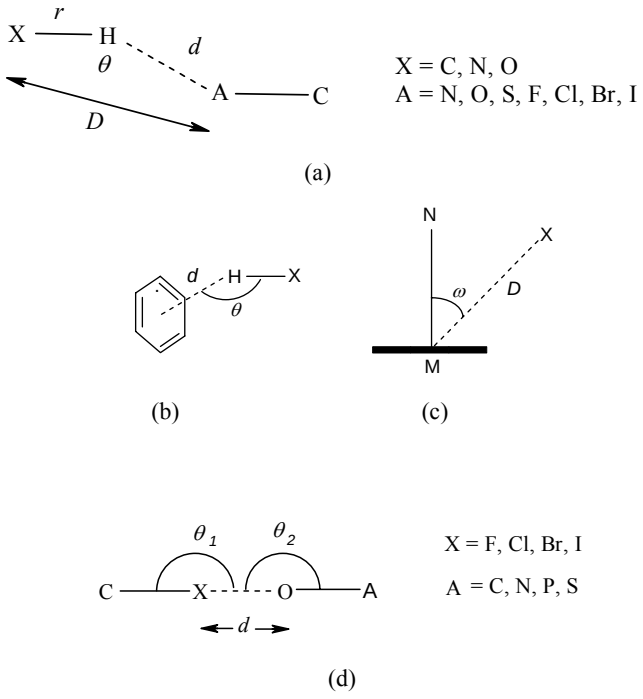
multifurcation, (c) the formation of C−H···O hydrogen bonds is enhanced by the activation of the $C_\alpha$−H atoms and by the flexibility of the side chain atoms, (d) in contrast to small-molecule structures, anti-cooperative geometries were found to be common in the 28 macromolecular structures, (e) there is a gradual lengthening of hydrogen bond as the extent of furcation increases, (f) the C−H···O bonds formed by Gly, Phe, and Tyr residues are noteworthy, (g) number of hydrogen bond donors and acceptors agree with Lipinski's rule-of-five that predicts drug-like properties, (h) 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 the earlier study [3.19] 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. Therefore it was felt that the conclusions of the 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), which was discussed in the previous chapter. 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 the previous study [3.19], other new interactions involving halogen atoms (both as electrophiles and nucleophiles), π-acceptors and sulfur-atom acceptors are examined. Eventually, it is hoped 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 chapter is to document these interactions reliably in a representative group of protein–ligand crystal structures.

## 3.2 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 the earlier study by Sarkhel and Desiraju [3.19] and 224 from Nissink *et al.* [3.22]. Because of the wrong assignment of ligand, PDB ID 1G2Y.pdb from the previous study [3.19] is not included in the present

study. The external kinase test set of 233 protein–ligand complexes was taken from the PDB (Appendix I, Table 8). 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 [3.23]. 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. The standard H-bonding criteria were set as $d$ (H···A) ≤ 3.0 Å and $\theta$ (X−H···A) ≥ 90°. For other weak interactions, the criteria are mentioned in the respective sections. A schematic description of the various interactions is given in Scheme 3.1.




**Scheme 3.1**: (a) Representative hydrogen bond. A–C is a single or double bond, (b), (c) parameters for X−H···π interactions (d) Parameters for halogen···O interactions. O–A is a double bond.

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 [3.24]. 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 viz. CHARMM22, AMBER96, OPLSAA, and MMFF94x were used to derive standard hydrogen bond geometries ($d$ and $\theta$) in the earlier studied 28 crystal structures by Sarkhel and Desiraju [3.19, 3.24, 3.25]. The MMFF94x force field 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 [3.26, 3.27].

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

The large number of structures and more personalized requirements led the design of HBAT software 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 [3.27–3.31]. These programs did not fit my 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. To evaluate the accuracy of the program, HBAT was used to reproduce the

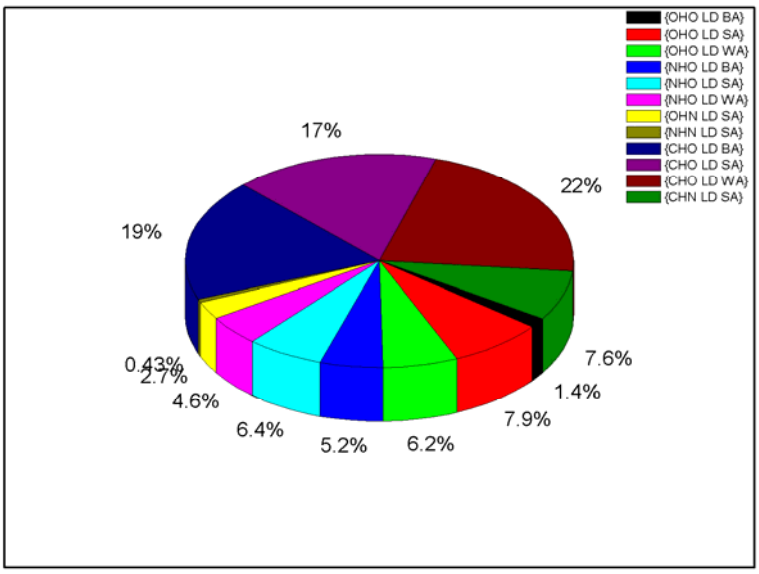
geometries in the earlier manual study of 28 complexes by Sarkhel and Desiraju [3.19] and also in other recent papers [3.32–3.35]. Results obtained are in excellent agreement with the earlier studies [3.19, 3.32–3.35]. This exercise has given me enough confidence to carry out the present study of 484 protein–ligand complexes.

## 3.3 Results and discussion

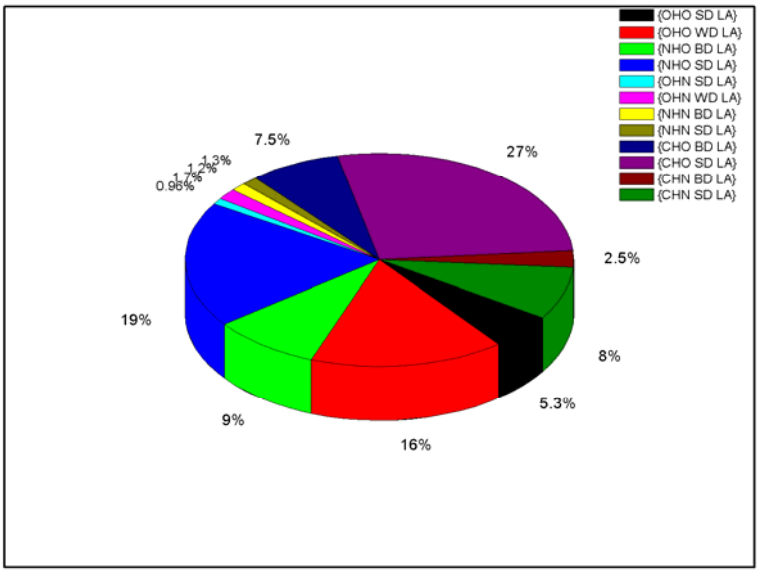
### 3.3.1 *Hydrogen bond geometry. Lengths and angles.*

In any database analysis, evaluation of large data sets provides a more unbiased identification of a chemical signal in the presence of crystallographic noise [3.36, 3.37]. A data set consisting of more than 100 representatives is ideal for the study of interaction geometry in crystal structures [3.38, 3.39]. 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 [3.40]. 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 3.1*a* and *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.




(a)




(b)

**Figure 3.1**: Distribution of all possible of hydrogen bond types in the active sites of protein–ligand complexes (a) ligand as donor (b) ligand as acceptor.

When the ligand is a donor, the percentage of strong hydrogen bonds is 34% while weak hydrogen bonds account for 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 shows 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 less than 2.0 Å and the hydrogen bonds are linear (Figure 3.2). (See Appendix I, Figure 9).

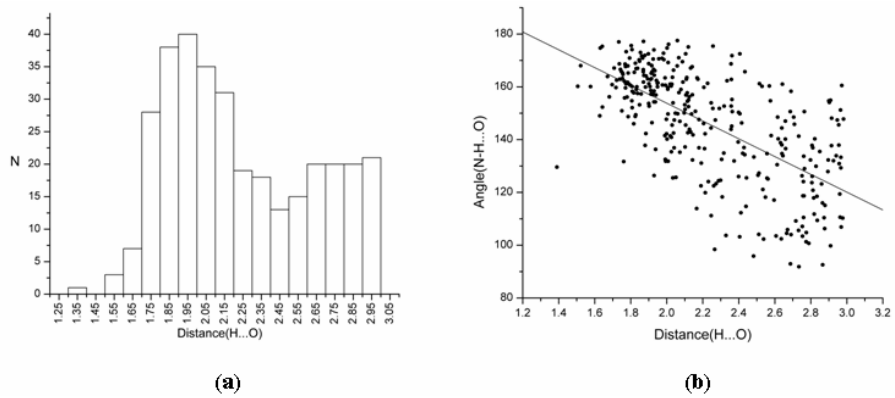



**Figure 3.2**: Distance H···O histogram and *d*-$\theta$ scatterplot ($R^2$ = –0.65) for {NHO BD LA} hydrogen bonds in the active sites of the 251 protein–ligand complexes considered in this study.

It was suggested earlier [3.19] based on PDB cone-corrected angular distributions, that there are small deviations from linearity for both N−H···O and O−H···O 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 180º. The cone-corrected angular distributions for {OHO LD BA} and {OHO SD LA} are similar with maxima in the range 175-180$^o$. The inverse length-angle correlations are also well behaved in all these cases (Figure 3.3). 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 protein–ligand interfaces are comparable to what is observed in small molecules. Baker and Hubbard have discussed hydrogen bond nonlinearity in their 1984 review [3.6]. However, based on the present observations it can be asserted 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 170-180º. There is another maximum at 135-150º which corresponds generally to multifurcated geometries. Also similar is {CHO LD BA} with a maximum of 175-180º. For {CHO SD LA} the maximum is still around 170-175º. For {CHO LD SA} the favored angle is around 140-145º with a narrow range of linearity. Unlike strong hydrogen bonds, the cone corrected weak C−H···O geometries show two distinct maxima at 130-150º and 170-180º. 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 (Figure 3.3). Also seen are C−H···$O_w$ bonds formed by main chain and side chain C−H groups to water as acceptor (See Appendix I, Figure 10). These are discussed later.

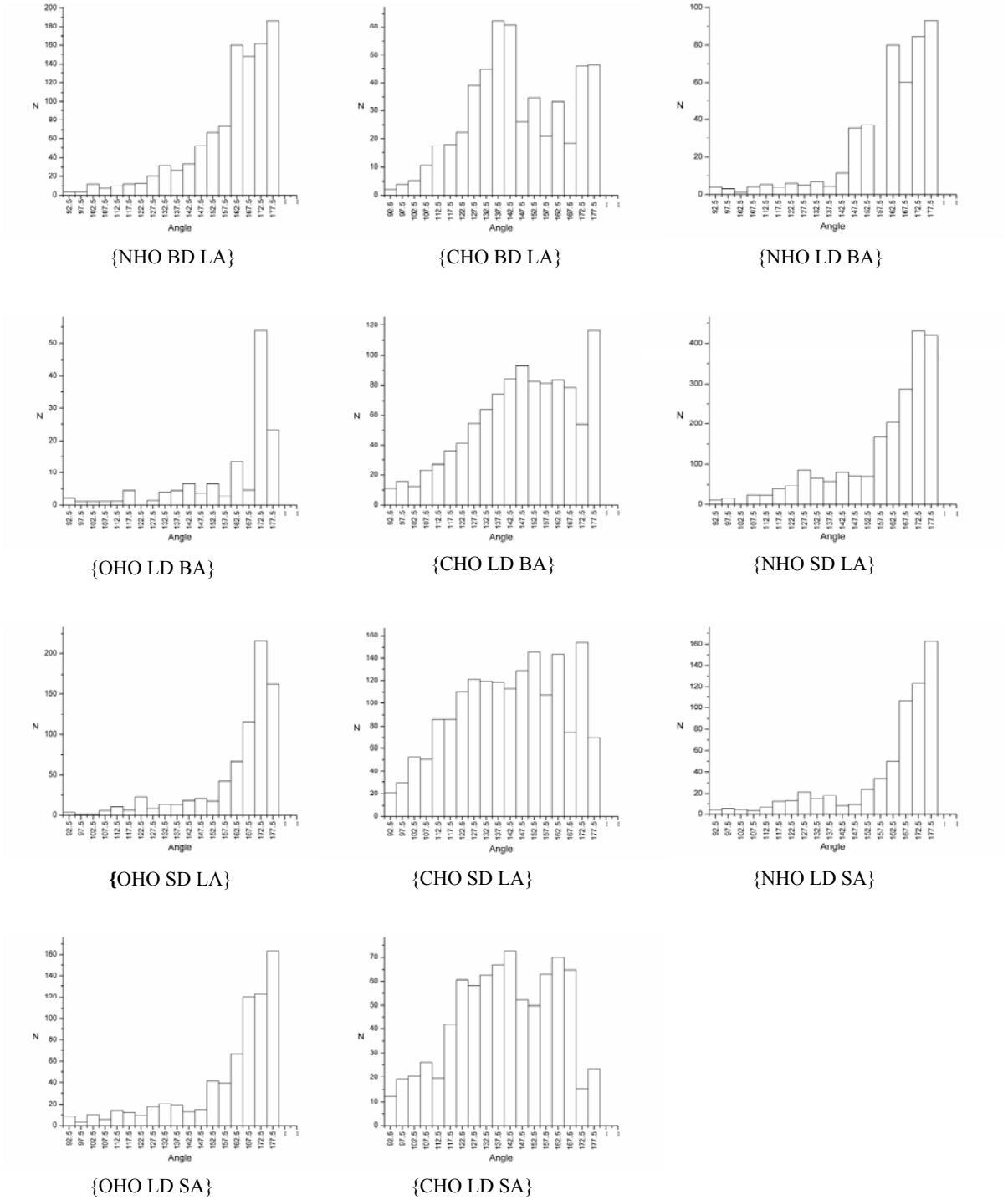
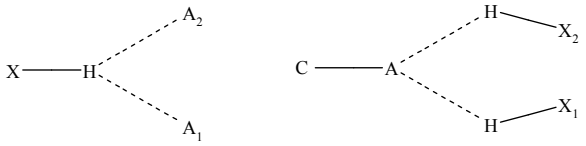



**Figure 3.3**: Histograms showing cone-corrected angular distribution for strong and weak hydrogen bonds in active sites of 251 protein–ligand complexes.

### 3.3.2 *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 3.2). 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 [3.41].




(a) Bifurcated donor (b) Bifurcated acceptor

**Scheme 3.2**: Notations for bifurcated donor/acceptor.

In the active sites of protein–ligand complexes, the level of furcation ranges from bifurcated to hexafurcated. Table 3.1 shows that donor and acceptor furcation occur roughly to the same extent (33299 furcated donors and 33038 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.

**Table 3.1**: Acceptor and donor furcations in the active sites of 251 protein–ligand complexes.

| | Furcated acceptors in active site | | | Furcated donors in active site | | |
|---|---|---|---|---|---|---|
| | In protein and water | In ligand | Total | In protein and water | In ligand | Total |
| Furcation Level | | | | | | |
| 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 3.2**: Strong and weak hydrogen bonds for ligands at various levels of donor and acceptor furcation.

| | Furcated acceptor | | | Furcated donor | | |
|---|---|---|---|---|---|---|
| Furcation level | 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* |

It is emphasized here 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 3.2). These ideas have been noted by the earlier study of 28 protein–ligand complexes by Sarkhel and Desiraju [3.19]. 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 3.2 shows that the total number of hydrogen bonds (O−H···O, N−H···O, C−H···O) to the 1104 furcated acceptors in Table 3.1 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 (Figure 3.4*a* and *b*).

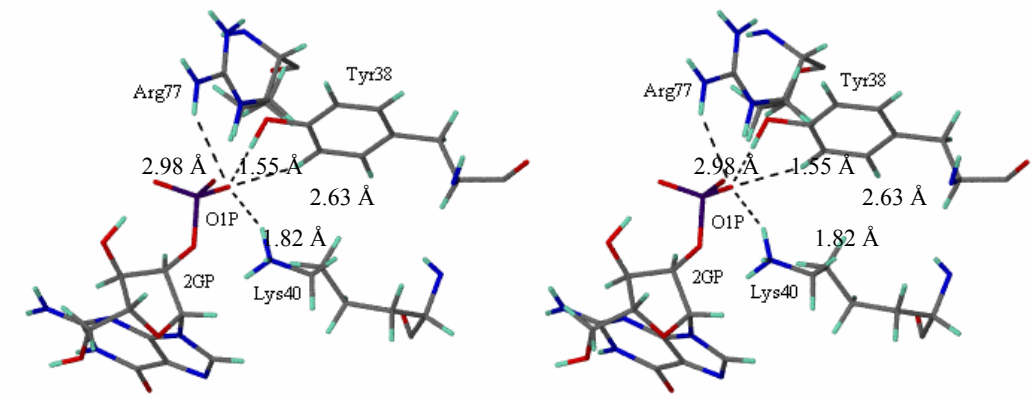



**Figure 3.4**: (a) Stereo view of tetrafurcated interaction in 2AAD.pdb. The acceptor centre is the O1P atom of the ligand (2GP). O1P interacts with residues Tyr38, Lys40 and Arg77 through C−H···O and O−H···O and N−H···O bonds. The respective H···O distances are also shown.

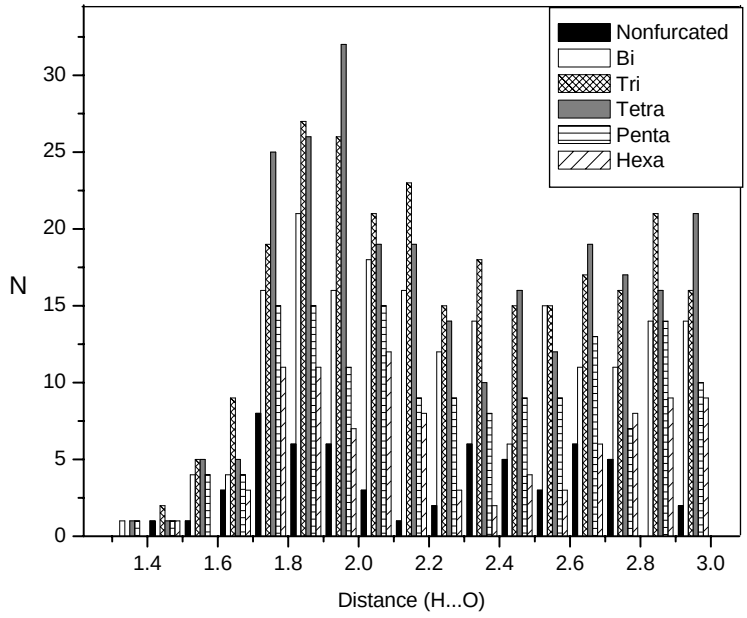



**Figure 3.4**: (b) Histogram of the distance *d* of N−H···O interactions to furcated ligand acceptors.

### 3.3.3 *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 by Sarkhel and Desiraju [3.19], which considered only 28 crystal structures, and all of very good resolution. One of the aims of the present study is to identify the resolutions limits for macromolecular crystal structures where the hydrogen bond geometries are as reliable as that obtained in small molecule structures. Therefore dataset of 251 structures classified 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). Table 3.3 contains a list of structures and Figure 3.5 gives this information pictorially.

**Table 3.3**: 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 | | |

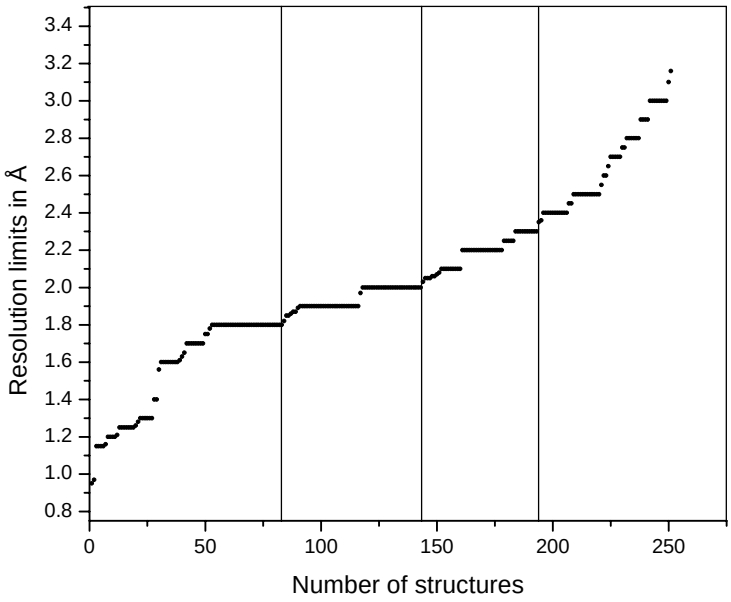
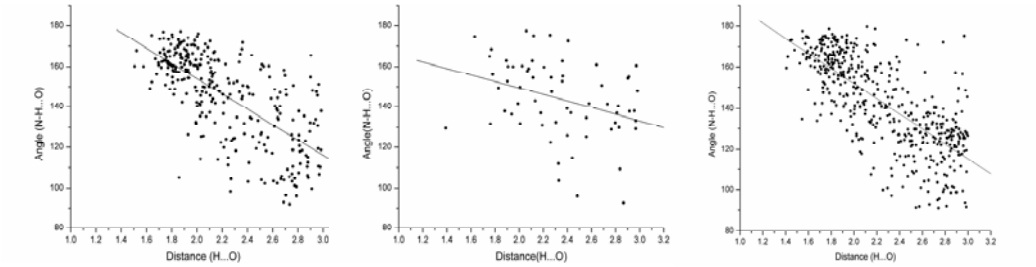



**Figure 3.5**: Distribution of crystal structures as a function of resolution.

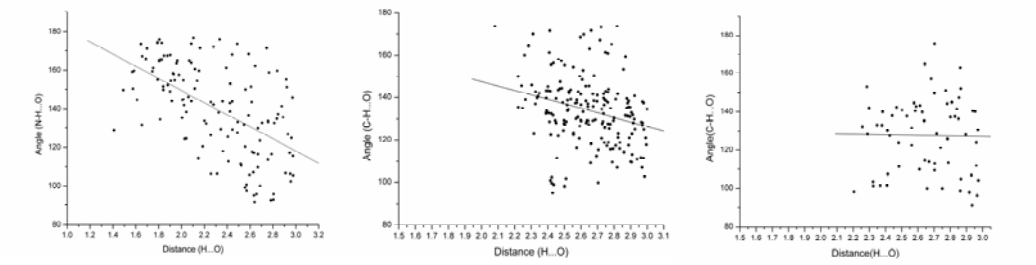
As prototypes of strong and weak hydrogen bonds, {NHO BD LA}, {NHO SD LA} and {CHO BD LA}, {CHO SD LA} are described here for comparisons of hydrogen bond geometry as a function of resolution. The $d$-$\theta$ scatterplots (cone-corrected) were analyzed for these prototype interactions (Appendix I, Figure 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 (Figure 3.6), 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. (Appendix I, Figure 11).




{NHO BD LA} ($R^2 = -0.69$)
Resolution < 2.3 Å
(a)

{NHO BD LA} ($R^2 = -0.35$)
Resolution > 2.3 Å
(b)

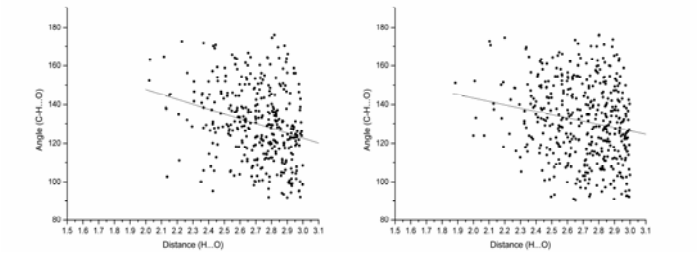
{NHO SD LA} ($R^2 = -0.74$)
Resolution < 2.3 Å
(c)




{NHO SD LA} ($R^2 = -0.55$)
Resolution > 2.3 Å
(d)

{CHO BD LA} ($R^2 = -0.26$)
Resolution < 2.0 Å
(e)

{CHO BD LA} ($R^2 = -0.02$)
Resolution > 2.0 Å
(f)




{CHO SD LA} ($R^2 = -0.26$)
Resolution < 2.0 Å
(g)

{CHO SD LA} ($R^2 = -0.18$)
Resolution > 2.0 Å
(h)

**Figure 3.6**: *d*-$\theta$ 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.

### 3.3.4 *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 (Appendix I, Table 6). Residues like Gly, Ile, Phe and 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 (Figure 3.7) and this is in keeping with the hydrophobic nature of the protein interior. This result is similar to the earlier observed trend in the smaller set of 28 structures [3.19] 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 $C_\alpha$−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. This fact is ascribed 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 earlier observations [3.19] and suggests that the highly flexible nature of Gly is very well exploited in biological recognition [3.42, 3.43].

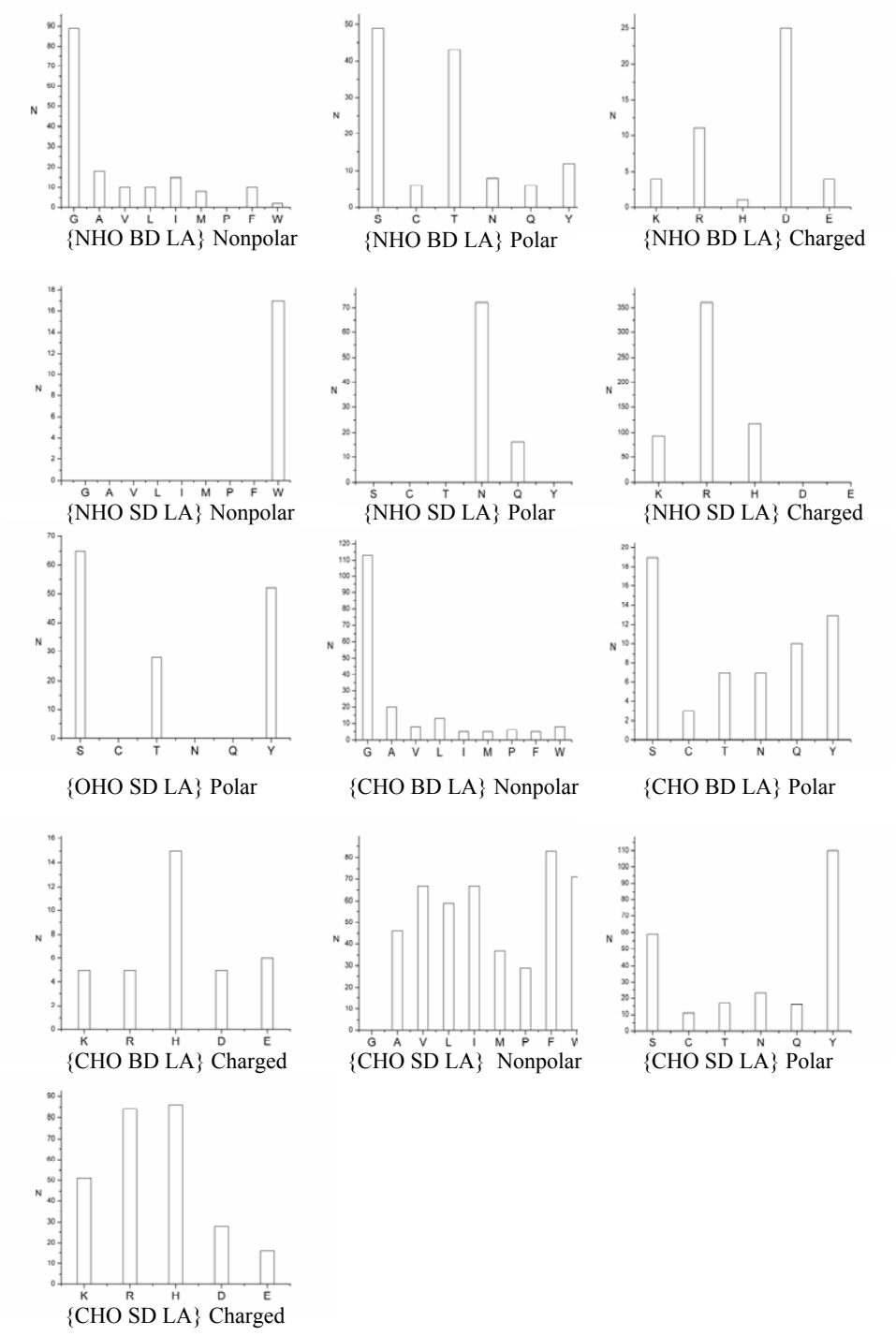



**Figure 3.7**: Frequency distribution of nonpolar, polar, and charged amino acids for strong and weak hydrogen bonds in the protein–ligand interface.

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 (Figure 3.8). The energies for the O−H···O and C−H···O interactions in Complex I are −6.68 and −1.83 kcal/mol. For Complex II, the values are −9.43 and −0.56 kcal/mol respectively.

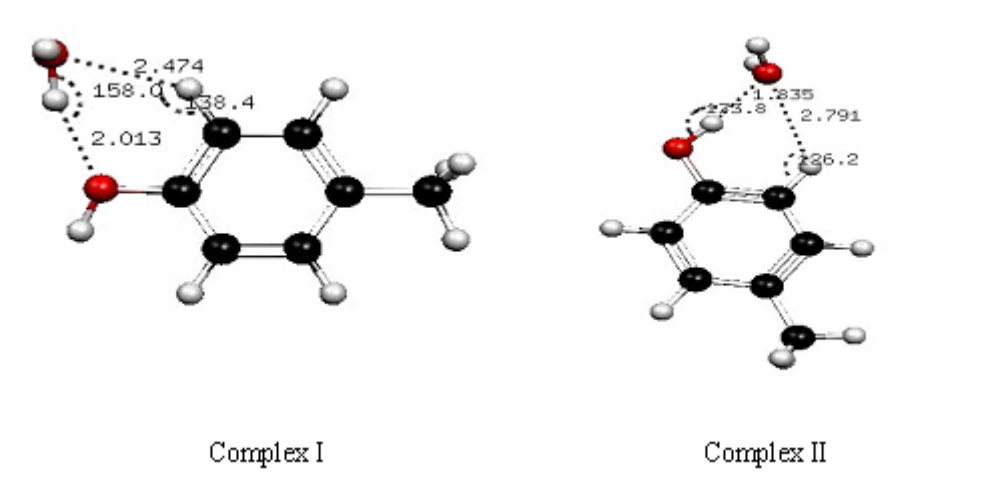



**Figure 3.8**: Two possible conformations of the interaction of the Tyr side chain with water.

### 3.3.5 *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 29718, 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 3.4 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···$O_w$.

**Table 3.4**: Percent distribution of various bond types among the 29718 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 |

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 the previous study [3.19]. For {OHO WD LA} the number of interactions observed is 592 of which 53% are below a H···O distance of 2.2 Å (Appendix I, Table 7). This is in contrast to the earlier study [3.19], 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 [3.44]; it is also chemically reasonable.

### 3.3.6 *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 acceptor in a ligand should be less than 5 and 10 respectively [3.45]. 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.

### 3.3.7 *Protein–ligand interactions in kinases*

An external test set of 233 protein–ligand complexes of various kinases was compiled from the PDB (Appendix I, Table 8) 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 (Appendix I, Figure 12, Table 8–10). These results suggest a similar trend for hydrogen bond geometries and constitute a useful validation of the principles enunciated in this analysis.

### 3.3.8 *Other weak interactions*

(a) *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 [3.13, 3.32]. In this 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, the convention adopted is shown in Scheme 3.1*b* and *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, $\theta \geq 100^0$ and $\omega \leq 40^0$ appear to be satisfactory and this geometric criterion is generally accepted [3.13]. In the present study 4 N−H···π (two each to Trp and Tyr), 3 O−H···π (two to His and one to Trp) and 159 C−H···π interactions are observed (Appendix I, Figure 13, Table 11). 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.

(b) *Halogen bonds*

Short oxygen···halogen interactions have been known since the 1950s [3.46]. A recent survey of protein and nucleic acid structures reveals similar halogen bonds as potentially stabilizing inter- and intramolecular interactions that can affect ligand binding [3.34]. A typical halogen bond is represented by Scheme 3.1*d*. Protein–ligand complexes were analyzed for possible halogen bonds following the van der Waals radii criterion. Seven halogen bonds are found in the following structures: 1CKP.pdb, 1CLE.pdb, 1BMA.pdb, 4EST.pdb, 2NLR.pdb, 1ETA.pdb listed in Table 3.5. These interactions were observed between C−F, C−Cl and C−I and carbonyl O-atoms in the main chain.

**Table 3.5**: Halogen bonds observed in the protein–ligand interfaces is tabulated along with respective PDB ID.

| | Type | Ligand | Residue ID | Acceptor | Residue ID | *d* | $\theta_1$ | $\theta_2$ | PDB ID |
|---|---|---|---|---|---|---|---|---|---|
| 1 | C−Cl···O=C | PVB | 1 | Asp | 86 | 3.0 | 134.0 | 78.4 | 1CKP |
| 2 | C−Cl···O=C | ENH | 703 | Thr | 96 | 2.9 | 147.7 | 74.2 | 1CLE |
| 3 | C–Cl···O=C | ENH | 703 | Arg | 100 | 2.8 | 152.9 | 81.0 | 1CLE |
| 4 | C−F···O=C | TFA | 256 | Cys | 199 | 2.9 | 133.9 | 108.2 | 1BMA |
| 5 | C−F···O=C | FPA | 5 | Pro | 3 | 2.8 | 102.1 | 90.0 | 4EST |
| 6 | C−F···O=C | G2F | 603 | Glu | 120 | 1.9 | 107.5 | 121.9 | 2NLR |
| 7 | C−I···O=C | T44 | 128 | Ala | 109 | 3.1 | 149.3 | 106.0 | 1ETA |

(c) *Halogen as nucleophile*

The acceptor capability of organic halogen, X (X = F, Cl, Br, I), has not been studied in detail in macromolecules [3.47]. 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 (Appendix I, Table 12). 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.

(d) *Hydrogen bonds involving sulfur atoms*

Sulfur atoms are larger and have a more diffuse electron cloud than oxygen and nitrogen, but are nevertheless capable of participating in hydrogen bonds [3.48]. The acceptor functionality of sulfur atoms have been studied here. Sulfur 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 (Appendix I, Table 13). 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.

## 3.4 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 to simple non-furcated hydrogen bonds. Resolution limits are important with respect to the hydrogen bond geometry. Strong hydrogen bonds retain good geometries up to 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 sulfur 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. I conclude that the results of the previous study of 28 structures are largely applicable to a set of structures that is nearly twenty 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.