PROTEINS: Structure, Function, and Bioinformatics 54:247–259 (2004)

N—H…O, O—H…O, and C—H…O Hydrogen Bonds in Protein–Ligand Complexes: Strong and Weak Interactions in Molecular Recognition

Sanjay Sarkhel and Gautam R. Desiraju*
School of Chemistry, University of Hyderabad, Hyderabad, India

ABSTRACT **The characteristics of N—H…O, O—H…O, and C—H…O hydrogen bonds are examined in a group of 28 high-resolution crystal structures of protein–ligand complexes from the Protein Data Bank and compared with interactions found in small-molecule crystal structures from the Cambridge Structural Database. It is found that both strong and weak hydrogen bonds are involved in ligand binding. Because of the prevalence of multifurcation, the restrictive geometrical criteria set up for hydrogen bonds in small-molecule crystal structures may need to be relaxed in macromolecular structures. For example, there are definite deviations from linearity for the hydrogen bonds in protein–ligand complexes. The formation of C—H…O hydrogen bonds is influenced 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 are common in the macromolecular structures studied here, and there is a gradual lengthening as the extent of furcation increases. C—H…O bonds formed by Gly, Phe, and Tyr residues are noteworthy. 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 are also seen to be relevant in ligand binding. Ligand C—H…O_w interactions are abundant when compared to N—H…O_w and O—H…O_w. This suggests that ligands prefer to use their stronger hydrogen bond capabilities for use with the protein residues, leaving the weaker interactions to bind with water. In summary, the interplay between strong and weak interactions in ligand binding possibly leads to a satisfactory enthalpy–entropy balance. The implications of these results to crystallographic refinement and molecular dynamics software are discussed. Proteins 2004;54:247–259.** © 2003 Wiley-Liss, Inc.

Key words: intermolecular interaction; enthalpy–entropy; Protein Data Bank; Cambridge Structural Database; multifurcation; water

INTRODUCTION

The binding properties of proteins are the essence of functional genomics. It is necessary to know where a protein is localized and when it is expressed, but to find out what it does, one needs to find out to what it binds—and more importantly, how. The specificity of biological processes suggests that the intermolecular interactions involved in the underlying recognition events are also specific, with conserved orientation.[1–3] Hydrogen bonding is directional, and this is what makes it so important in the whole domain of biomolecular recognition.[4,5] Hydrogen bonds are instrumental not only in mediating drug-receptor binding, but they also affect physicochemical properties of a molecule, such as solubility, partitioning, distribution, and permeability, that are crucial to drug development.[6] Another compelling aspect of the hydrogen bond is its composite character. The hydrogen bond, or rather the "hydrogen bridge," is viewed as an interaction that has covalent, electrostatic, and van der Waals character and spans an energy range from 40 kcal/mol to ~0.25 kcal/mol.[7] The composite nature of the hydrogen bond means that the relative proportions of covalency, electrostatics, and van der Waals character in the X—H…A interaction vary smoothly depending on the nature of X and A. This in turn renders the interactions chemically "tunable" with the corresponding implications for function.

No less important in biological processes than specificity is reversibility. Weaker interactions can be made and broken more easily than stronger interactions. In this context, it is of interest to assess the relative significance of strong and weak interactions in the macromolecular recognition process. Is protein–ligand binding governed by conventional, that is, electrostatic, N—H…O and O—H…O hydrogen bonds, or do weaker interactions with a greater dispersive component such as C—H…O also play a role? If so, to what extent are they significant? While several recent studies have concentrated on identifying or validating the presence of C—H…O and other weak hydrogen bonds in macromolecular structures,[8–16] the purpose of this article is twofold: (1) to describe strong and weak hydrogen bonds in a particular category of biological structures that is of importance in drug design, namely,

Grant sponsor: Department of Science and Technology, Government of India; Project number: VI-D&P/2/99/TT.

*Correspondence to: Gautam R. Desiraju, School of Chemistry, University of Hyderabad, Hyderabad 500 046, India. E-mail: desiraju@uohyd.ernet.in

Received 13 May 2003; Accepted 16 June 2003

TABLE I. PDB ID, Resolution, and Description of the Protein–Ligand Complexes in This Study

PDB ID	Resolution (Å)	Protein / substrate(s)
1BXO	0.95	Acid proteinase (Penicillopepsin) / Methyl cyclo[(2S)-2-[[(1R)-1-(*N*-(L-*N*-(3-methylbutanoyl)valyl-L-aspartyl)amino)-3-methylbutyl]hydroxyphosphinyloxy]-3-(3-aminomethyl)phenylpropanoate
8A3H	0.97	Endoglucanase Cel5A / 5-Hydroxymethyl-5,6,7,8-tetrahydro-imidazo[1,2-a]pyridin-6yl-7,8-diol-glucopyranoside
1G2Y	1.00	Hnf-1 dimerization domain / Selenomethionine
1A6G	1.15	Myoglobin / Protoporphyrin
1HDO	1.15	Human biliverdin Ix reductase / NADP
1HET	1.15	Alcohol dehydrogenase E chain / NADH
1RGE	1.15	Guanyloribonuclease / Guanosine-2′-monophosphate
1MRO	1.16	Methyl-coenzyme M reductase / 1-Thioethanesulphonic acid / 7-Thioheptanoyl-threoninephosphate
1C0P	1.20	D-Amino acid oxidase / Flavin-adenine dinucleotide
1I3H	1.20	Concanavalin A / α-D-Mannose
1I76	1.20	Matrix Metallo Proteinase-8 / 2-(Biphenyl-4-sulfonyl)-1,2,3,4-tetrahydro- isoquinoline-3-carboxylic acid
2NLR	1.20	Endoglucanase / 2-Deoxy-2-fluoro-glucose
1D4O	1.21	Transhydrogenase domain III / NADP
1C1D	1.25	L-Phenylalanine dehydrogenase / NADH / L-Phenylalanine
1EQO	1.25	Pyrophosphokinase / Diphosphomethylphosphonic acid adenosyl ester / 6-Hydroxymethyl-7,8-dihydropterin
1OAA	1.25	Sepiapterin reductase / NADP
2TPS	1.25	Thiamine phosphate synthase / Thiamine phosphate
2WEA	1.25	Penicillopepsin / Methyl[cyclo-7[(2R)-((*N*-Valyl)amino)-2-(hydroxyl-(1S)-1-methyoxycarbonyl-2-phenylethoxy)phosphinyloxy-ethyl]-1-naphthaleneacetamide], sodium salt
3CHB	1.25	Cholera toxin B / D-Galactose / Glucose / *n*-(Ethylsulfite)morpholine / *n*-Acetyl-D-galactosamine / o-Sialic acid
1CTQ	1.26	P21Ras / Phosphoaminophosphonic acid-guanylate ester
1QKS	1.28	Cytochrome Cd1 nitrite reductase / Heme D / Heme C
1BK0	1.30	Isopenicillin *N* synthase / L-D-(A-aminoadipoyl)-L-cysteinyl-D-valine
1DJR	1.30	Enterotoxin B-pentamer / M-Carboxyphenyl-α-D-galactopyranoside
1FCY	1.30	Retinoic acid receptor / 6-(5,5,8,8-tetramethyl-5,6,7,8-tetrahydro-naphthalene-2-carbonyl)-naphthalene-2-carboxylic acid / Dodecyl-α-D-maltoside
1FK5	1.30	Lipid-transfer protein / Oleic acid
1HYO	1.30	Fumarylacetoacetate hydrolase / 4-(Hydroxymethylphosphinoyl)-3-oxo-butanoic acid
1I12	1.30	Glucosamine-phosphate *N*-acetyltransferase / Acetyl coenzyme*A
1RUV	1.30	Ribonuclease A / Uridine-2′,3′-vanadate

protein–ligand complexes; and (2) to examine the C—H…O interaction in the context of the stronger N—H…O and O—H…O hydrogen bonds in the chosen structures. Given that biological processes are subtle and that small structural changes can sometimes lead to large functional differences, we believe that the questions posed above are worthy of further study. Not much accurate crystallographic data are available for protein–ligand complexes, and this work makes a beginning with respect to the analysis of hydrogen bonds in these compounds.

MATERIALS AND METHODS

We have examined geometrical characteristics of strong and weak hydrogen bonds in high-resolution crystal structures of protein–ligand complexes as obtained from the Protein Data Bank (PDB).[17] We have compared these interactions with information on similar hydrogen bonds extracted from the Cambridge Structural Database (CSD).[18] To avoid ambiguities, the PDB search was performed only for crystal structures of protein–ligand complexes having a resolution better than 1.3 Å. Of the 69 such structures found, 28 with organic ligands were selected (Table I). Structures not chosen are those in which the ligands are ions, solvent molecules, or disordered species. For complexes containing two or more of the same ligand, only one ligand was considered; for complexes with two or more different ligands, all of them were included in the study. Very small ligands (less than five non-H atoms) and intramolecular hydrogen bonds were not considered.

A few comments on the selection of the structures for this study are in order. The number of protein–ligand complexes in the PDB is quite small. However, for the problem in hand, namely, analysis of different types of hydrogen bonds and in particular, the weaker ones, it is important to have structures of the highest possible resolution. This further reduces the number of available structures. Despite this, it was felt that the resolution criterion should not be relaxed, because the information available from the poorly resolved structures is not really helpful in the assessment of weak hydrogen bonds. A list of the protein–ligand complexes examined in this study is given in Table I, with the respective resolutions. Table I shows that the data set also has a high degree of homology. However, we decided to select all these 28 structures, so that any conserved features of these weak interactions would be revealed in the analysis, and also in order not to truncate an already small set of structures. The results reported in this article should be viewed in this light.

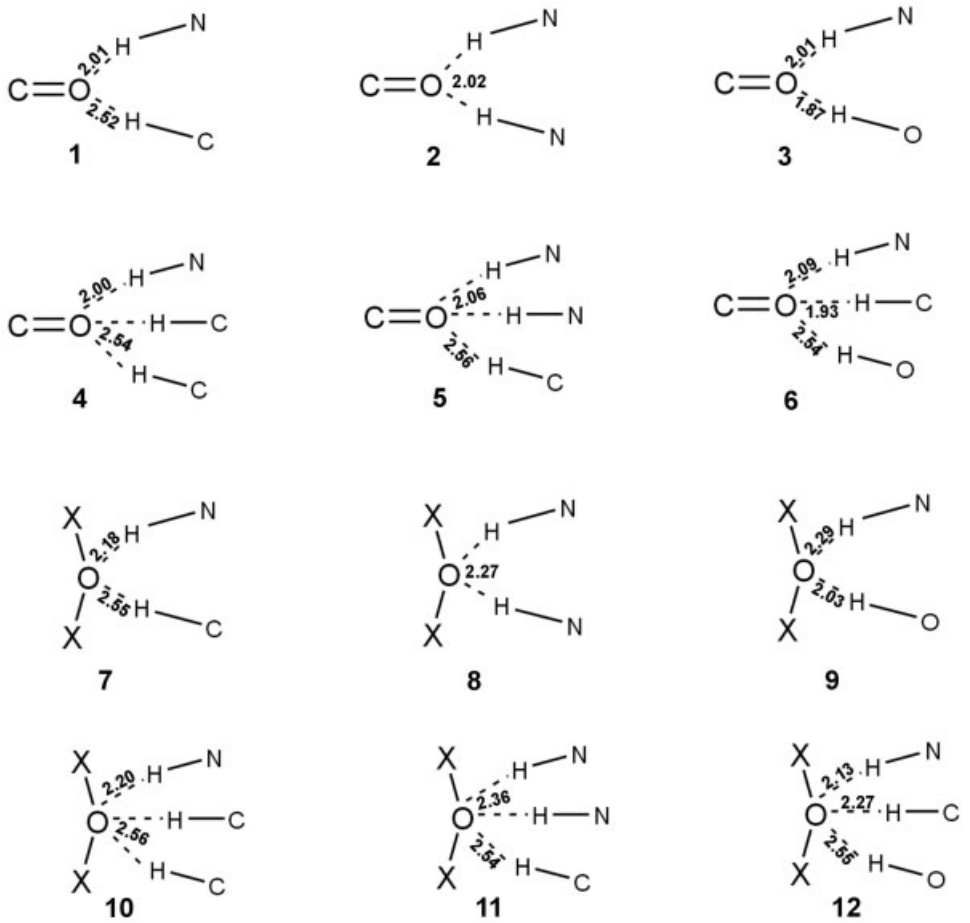

Fig. 1. Synthons 1–12 examined in the CSD analysis. X is any non-hydrogen atom. The numbers indicate corresponding distances.

H-atoms were added to the protein, water, and the ligand with the program Cerius^2.[19] The H-atoms for protein and water molecules were added from the template. For the ligands, the H-atom addition was performed after extracting the ligand from the binding site. The H-atoms were introduced with the "H ADJUST" option after specifying the correct valence of the atoms in the ligand molecule. The H-atom positions were then refined (energy minimization), keeping the position of the non-H atoms fixed. The active site for ligand binding was defined by selecting amino acid residues within a 5 Å radius of the ligand molecule. The H-bonding criteria were set as d (H…A) $\leq$ 2.7 Å and θ (D—H…A) $\geq$ 90°. Interactions between the ligand and water were considered, but not between protein and water, for the sake of simplicity in data handling; however, we do note that such interactions may be just as important in the receptor–ligand interaction.

The Cambridge Structural Database (CSD) serves as a useful benchmark in this study. The interplay of noncovalent interactions in the approximately 250,000 experimentally determined small-molecule crystal structures may now be analyzed reliably. Rigorous statistical analysis of all available data can provide convincing evidence about the "true" nature of nonbonded interactions. One of the primary objectives of this study is to compare the trends in geometrical parameters of hydrogen bonds in protein–ligand complexes with those seen in small-molecule structures. Any similarities and differences noted may then provide information about the inherent nature of these interactions in two different environments. The searches for synthons **1–12** (Fig. 1) in the CSD (Version 5.22, October 2001) were carried out with the following criteria: (1) Three-dimensional coordinates determined (2) no errors, (3) no ions, (4) not disordered, (5) not polymeric, (6) only organics, (7) H-normalized, and (8) R-factor $\leq$ 0.05. For motifs **1–6**, hits with carboxyl and carboxylate groups were not considered. For motifs **7–12**, where X is any atom, contacts with water molecules or with O-atoms attached to P- or S-atoms were not considered. Four-centered interactions were not considered.

RESULTS AND DISCUSSION

There are two general considerations pertinent to the study of the metrics of hydrogen bonds in protein–ligand complexes: (1) Strong interactions such as N—H…O or O—H…O hydrogen bonds may be studied from individual crystal structures, but weak bonds such as C—H…O need to be examined with statistical methods, in other words, by retrieving information from many structures from crystal-

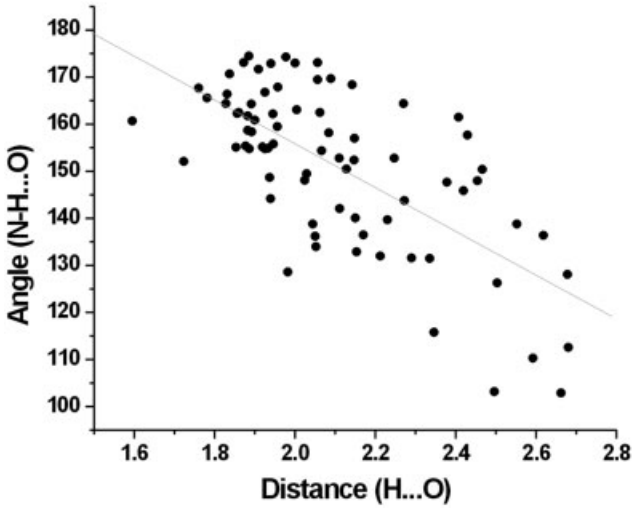

Fig. 2. Distance-angle scatterplot for main-chain (donor)-ligand N—H. . .O interactions.

TABLE II. PDB-Derived geometrical parameters (Mean *d* and θ Values With esu's) for Hydrogen Bonds

(A) Main-Chain–ligand interactions

Interaction	Donor (**X**—**H**…O)	Acceptor (X—H…**O**)
N—H…O	2.10(25), 151(17)	2.14(31), 145(25)
O—H…O		2.12(31), 138(27)
C—H…O	2.50(13), 141(13)	2.40(18), 139(19)

(B) Side-chain–ligand interactions

Interaction	Donor (**X**—**H**…O)	Acceptor (X—H…**O**)
N—H…O	2.09(27), 149(19)	2.08(27), 151(24)
O—H…O	2.03(37), 142(24)	2.20(34), 132(23)
C—H…O	2.51(15), 138(17)	2.41(22), 142(21)

X—**H**…O denotes H-bonds donated by the protein residues. X—H…**O** denotes H-bonds accepted by the residues.

lographic databases.[20] The properties of C—H…O hydrogen bonds are not readily derived from geometrical parameters in one or a small number of crystal structures, because of significant structural interference from many other interactions. Statistical methods are required so that the "chemical signal" from the interaction of choice is distinguished from the "crystallographic noise" from the rest of the structure. There is much precedent, therefore, for the use of the CSD to extract meaningful information about the C—H…O and other weak hydrogen bonds.[21] Accordingly, we have used the PDB here to study hydrogen bonding in a macromolecular environment. (2) The accuracy and precision of the results in a macromolecular X-ray study are typically much less than in a small-molecule crystal structure determination.[4] The following factors contribute to this: (a) macromolecular crystals are not time-stable, and structural fluctuations that occur at the time scale of the X-ray experiment place a limit on the resolution; (b) the flexibility of the side-chains limits the quality of the results; (c) the confidence levels of the metrics are severely affected by the bias due to the restraints imposed in the refinement. The positional error of atoms is roughly a sixth of the resolution.[22] In this work, we restricted ourselves to structures from the PDB with a resolution of less than 1.3 Å. We hoped that by doing so, the above-mentioned problems peculiar to macromolecular structures would be, to some extent, mitigated.

Geometrical Characteristics

The accepted geometrical requirement for a N—H…O or an O—H…O hydrogen bond requires the H…O distance, *d*, to be approximately 2.0 Å with a more or less linear approach.[23] One of the aims of this study was to evaluate this criterion for protein–ligand complexes. This was done by comparing statistical data from biomolecular structures with those obtained from small-molecule crystal structures. The PDB-derived scatter plots of *d* versus θ for the main-chain (Fig. 2) and side-chain (deposited) N—H…O donors show an inverse correlation, a feature characteristic of the hydrogen bond. Many interactions are quite long (2.2– 2.7 Å), but they are well behaved; most lie close to the least squares line and, in this respect, they may be termed weak rather than bad. The fluctuations are clearly more for the side-chain donor interactions in this range, a consequence of their greater flexibility.

One of the features observed in the cone-corrected PDB angular distributions for the N—H…O and O—H…O interactions is their small deviations from linearity (Table II, Fig. 3). This is in contrast to small-molecule crystal structures, where there is a pronounced tendency to linearity. Baker and Hubbard[24] noted this feature long ago in their seminal study of hydrogen bonding in native globular proteins. In our study, the main-chain donor N—H…O distribution resembles the small-molecule cone-corrected distribution except for the fact that the maximum is closer to 175° than to 180° [Fig. 3(a)]. The side-chain donor N—H…O distribution is not uniform and is marked by fluctuations [Fig. 3(f)]. The shift in the distribution of the hydrogen bond angle is unambiguous, and the maximum now occurs near 170°. The O—H…O bonds show a similar trend, and although the side-chain statistics are too "thin," the favored angle is around 170° [Fig. 3(b and g)]. Essentially, the hydrogen bond angles θ, for both O—H…O and N—H…O bonds cluster in similar ranges. Deviations from linearity of the main-chain donor N—H…O bonds may be due to the lesser freedom of the N—H groups to orient appropriately toward the acceptor. However, this would not explain why the more flexible side-chains also form bent O—H…O bonds. Perhaps this latter observation is of entropic origin, in that rotation of the hydrogen bonds can offset the entropy loss that is associated with protein–ligand binding. Despite the limited data set, one may suggest that any particular hydrogen bond (O—H…O, N—H…O, or C—H…O) may have different preferences in chemical and biological contexts. In particular, entropic effects should be considered in biological structures.

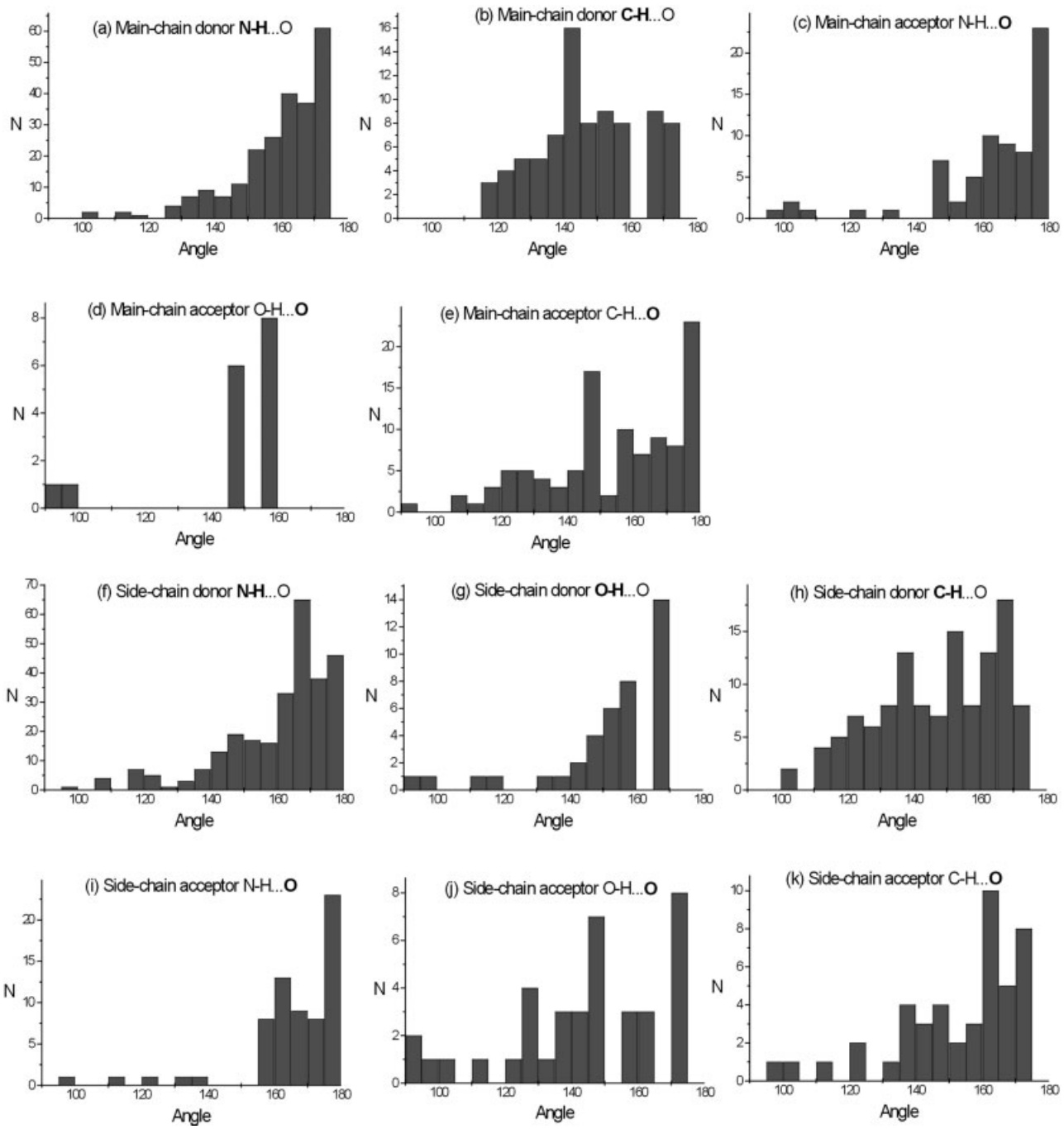

Fig. 3. Histograms showing cone-corrected angular distribution for protein-ligand hydrogen bonds.

For the PDB structures, the d–θ scatterplots (deposited) of the C—H…O interactions formed by the main-chain and the side-chain clearly show that the inverse correlation is better for the more acidic and more rigid main-chain C_α—H donors as compared to the less activated and more flexible side-chain C—H donors. In general, the distance distribution of the C—H…O interactions in the protein–ligand complexes is different from that seen in small molecules and hints that the range of the weak C—H…O interaction is effectively extended in the macromolecular environment.

Protein–ligand binding is different from the simpler situations that prevail in supramolecular chemistry, in that it is generally mediated by multicentered interactions and multipoint recognition. An analysis of the hydrogen bonds in the complexes in Table I shows that acceptor-furcation rather than donor-furcation is preferred. In small-molecule structures, one observes donor-furcation,

TABLE III. Statistics of Donor-Furcation (*n*-Centering) for Hydrogen Bonds in Protein–Ligand Complexes with Water Included as an Acceptor

n	N—H…O	C—H…O	O—H…O	HO_w—H…O
2	160	160	54	95
3	21	13	9	9
4	1	—	1	1
5	1	—	—	—

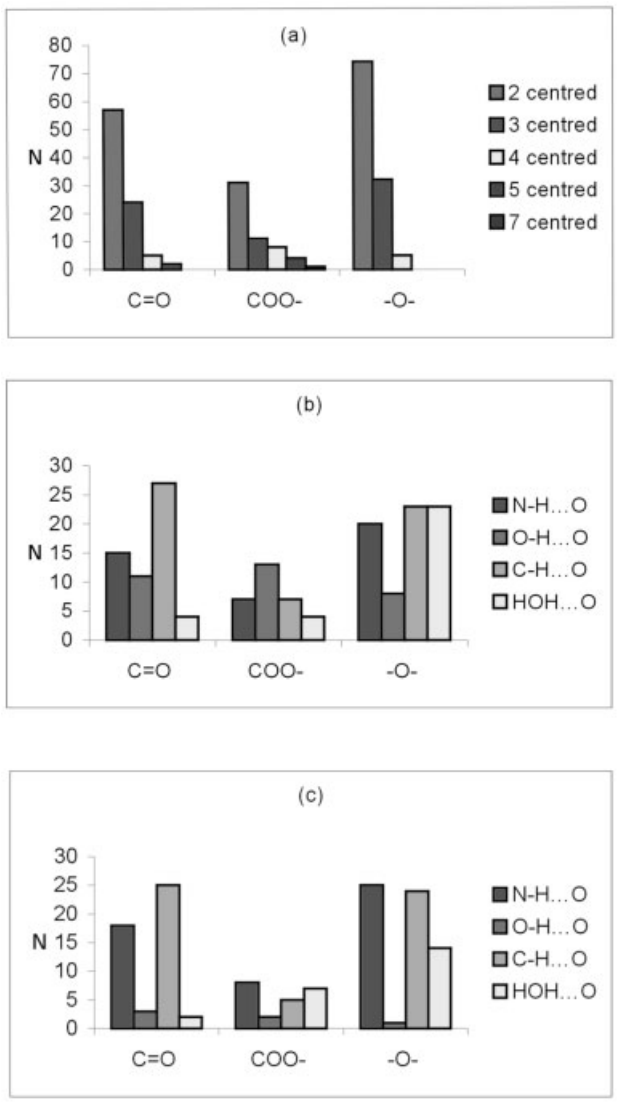

Fig. 4. Histograms of hydrogen bond acceptor functionality (**a**) patterns of "*n*-centered" acceptor furcation, and (**b**) relative distribution of different types of hydrogen bonds for two-centered and (**c**) three-centered interactions. Water has been included as a donor and acceptor for interactions with ligands only.

or the approach of many acceptors toward a donor, if any furcation is present at all. This is characteristic of acceptor-rich molecules, and most small molecules are in this category, because the only simple donor-rich functional groups are —OH and —NH_2, and most other functional groups are acceptor-rich. In protein–ligand complexes,

TABLE 4. Geometrical Parameters for Hydrogen Bond Distances and Angles Within the Acceptor-Furcated Synthon 1

PDB ID	Interaction type	Atom identity	*d*, θ
1BXO	N—H…O	Gly, m.c.	2.29, 131.6
	C—H…O	Tyr, C_α	2.61, 135.7
1COP	N—H…O	Asn, m.c.	1.93, 154.9
	C—H…O	Ala, C_α	2.42, 142.0
1G2Y	N—H…O	Leu, m.c.	1.93, 154.8
	C—H…O	Leu, C_γ	2.46, 132.5
1HET	N—H…O	Ligand	1.96, 170.8
	C—H…O	Ligand	2.28, 138.4
	N—H…O	Phe, m.c.	1.83, 166.4
	C—H…O	Ile, C_α	2.43, 141.3
1I12	N—H…O	Val, m.c.	1.88, 161.2
	C—H…O	Ala, C_α	2.65, 142.1
1QKS	N—H…O	Trp, s.c. ($N_{\varepsilon 1}$)	2.22, 147.6
	C—H…O	Met, C_γ	2.55, 143.6

however, donor-furcation is uncommon (Table III). In contrast, it is acceptor-furcation, or the approach of many donors toward an acceptor, that is observed, with C—H groups also behaving like donors, in addition to N—H and O—H groups. Put another way, there are more donors from the protein residues lined up in the binding site as compared to the number of acceptors in the ligand. The usual acceptors are C═O, COO^-, and —O—. Figure 4(a) is a histogram showing the propensity for "*n*-centered" acceptor-furcated hydrogen bonding in the 28 protein–ligand complexes in this study. Figure 4(b and c) show the breakdown of the two- and three-centered hydrogen bonds in this list with respect to variations in the donors and acceptors. The dominance of C—H…O hydrogen bonds in these distributions is noteworthy.

Another way in which the multicentered hydrogen bonds in protein–ligand complexes may be contrasted to the two-centered bonds in small-molecule structures is that the former correspond to anticooperative geometries.[5] This suggests that the restrictive criteria set for these interactions to qualify as a hydrogen bond probably need to be relaxed in multifurcated systems. To examine the effects of acceptor-furcation on hydrogen bond distances, a CSD search was performed on structures with the synthons[25] shown in Figure 1. The trends (data deposited) seen for synthons **1–6** and **7–12** show that as the extent of furcation increases, there is a gradual lengthening of the

Fig. 5. Examples of anticooperative synthon **1** in protein–ligand complexes: (**a**) a moderate C—H…O from Met (2.55 Å) and a rather long N—H…O from Trp (2.22 Å) (PDB ID: 1QKS); (**b**) a strong N—H…O from Val and a weak C—H…O from Ala (PDB ID: 1I12).

Fig. 7. Examples of C—H…O interactions between protein and ligand: (**a**) a relatively short Gly (donor) C_α—H…O bond (PDB ID: 3CHB); (**b**) the acidic aromatic proton of Tyr involved in a bifurcated C—H…O interaction (PDB ID: 1EQO); (**c**) the aromatic ring of Phe acting as a C—H…O donor (PDB ID: 1I12).

Fig. 10. Examples showing the two-fold acceptor functionality of the water molecule: (**a**) two C—H…O interactions from the ligand (PDB ID: 1HDO);**b**) one weak C—H…O and one strong O—H…O bond (PDB ID: 3CHB).

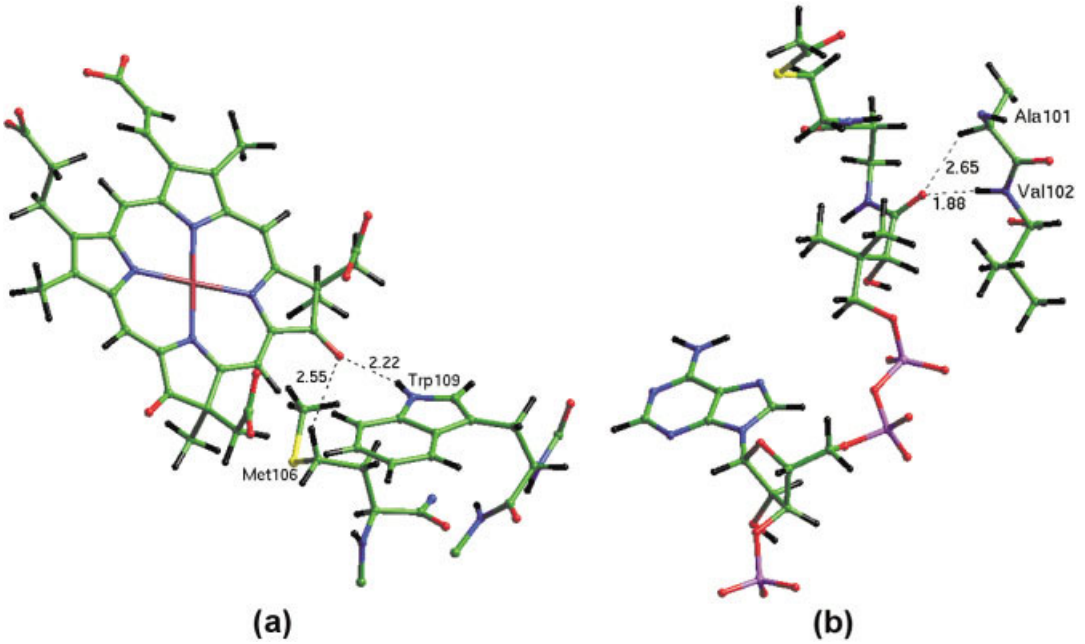

Figure 5

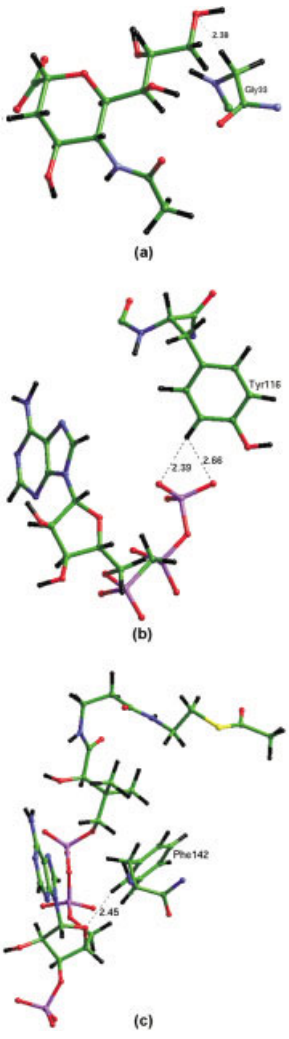

Figure 7

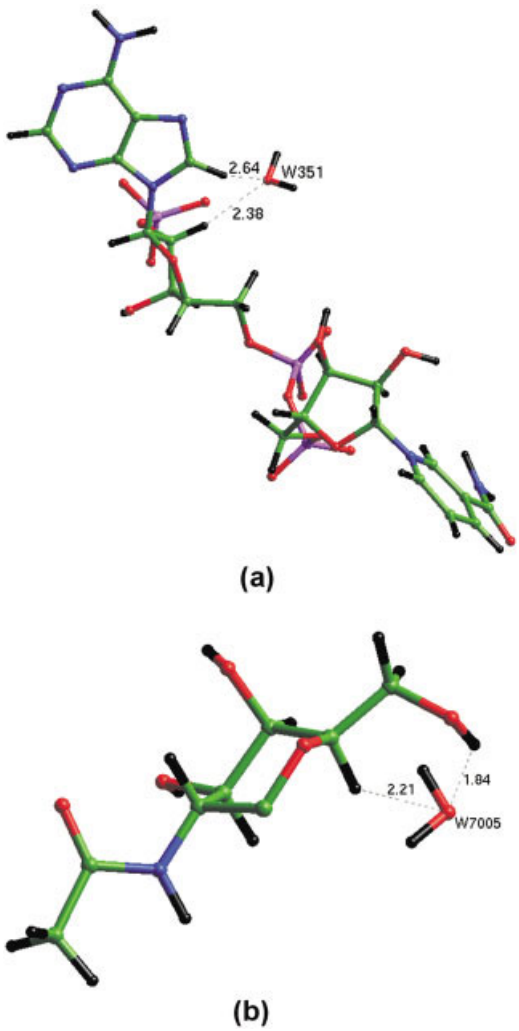

Figure 10

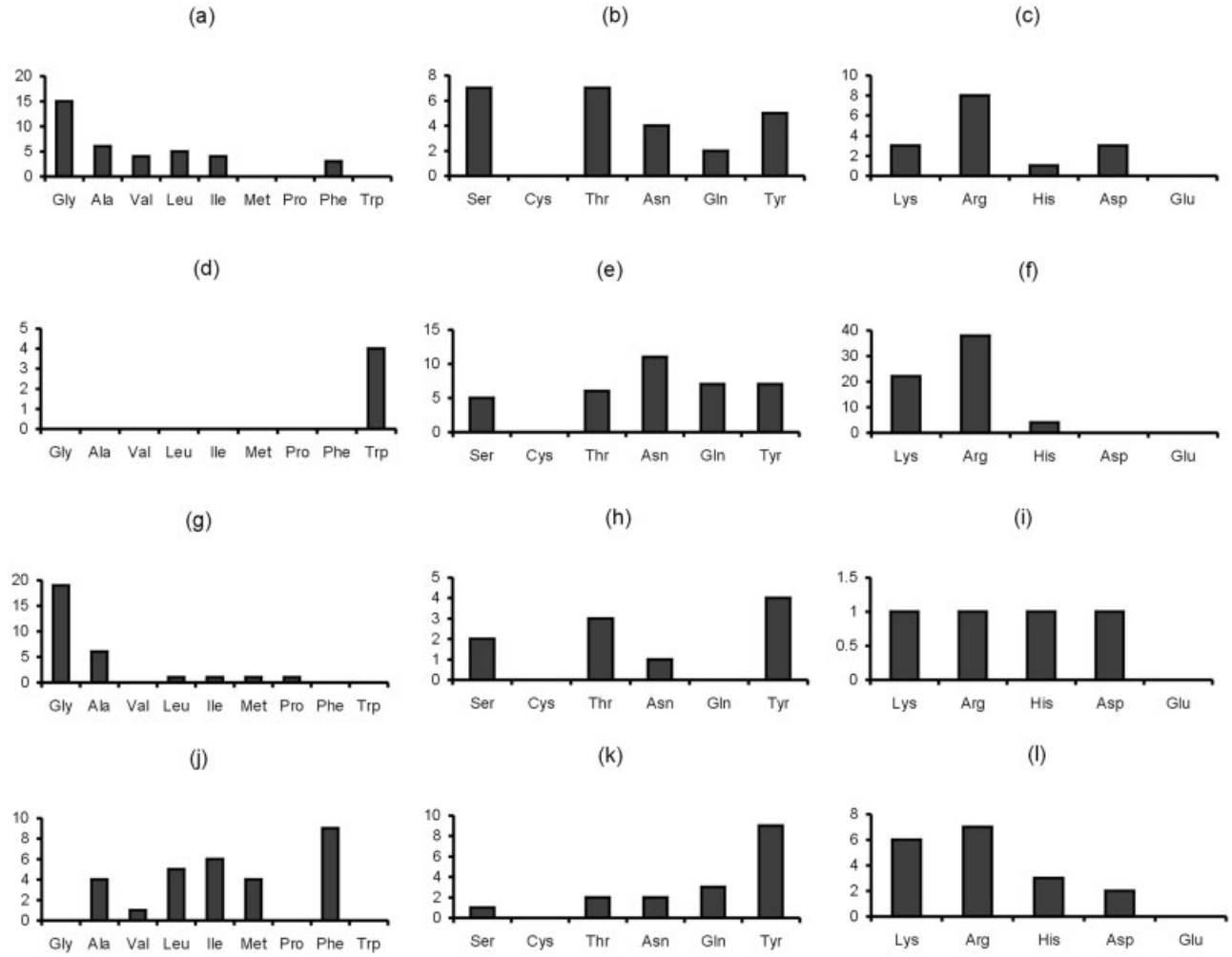

Fig. 6. Frequency distributions of hydrogen bonds for amino acid residues: (**a–c**) main-chain donor **N**—**H**…O for nonpolar, polar, and charged side-chain residues; (**d–f**) side-chain donor **N**—**H**…O and **O**—**H**…O for nonpolar, polar, and charged side-chain residues; (**g–i**) main-chain donor **C**—**H**…O for nonpolar, polar, and charged side-chain residues; (**j–l**) side-chain donor **C**—**H**…O for nonpolar, polar, and charged side-chain residues.

mean interaction distance. These observations pertain to small-molecule structures, where recognition (interaction) sites are few in number; one can appreciate that this effect would be compounded in a macromolecular environment where the recognition patterns are more intricate. Table IV gives the geometrical features associated with the anticooperative synthon *1* formed with N—H…O and C—H…O interactions (Fig. 1). The anticooperativity is demonstrated in Figure 5(a) (PDB ID: 1QKS),[26] wherein the N—H…O length is rather long and the C—H…O length is moderate. A potentially stronger interaction has in effect been weakened by a supposedly weaker interaction that has gained in the process. In Figure 5(b) (PDB ID: 1I12),[27] there is a short N—H…O bond and a long C—H…O interaction, but the geometry is still anticooperative. These anticooperative geometries are hard to understand unless the N—H…O and C—H…O hydrogen bonds are comparable in nature. In small-molecule crystal structures, H…H repulsion would normally disfavor such geometries.

It is of interest to discuss why acceptor-furcation is, in general, prevalent in the crystal structures of protein–ligand complexes, while if any kind of furcation is even present in small molecule structures, it is likely to be donor-furcation. After all, both groups of crystal structures are populated by organic molecules, and the kinds of interactions formed are the same, namely, heteroatom interactions involving C, H, N, O, and a few other atoms. The literature on C—H…O hydrogen bonds in small-molecule crystal structures shows that these interactions generally play only a supportive or passive role.[5] The stronger hydrogen bonds like N—H…O and O—H…O dominate the packing, and their geometrical requirements are the first to be met. In biological structures, however, multipoint recognition is ubiquitous. Whether this has a functional significance or not is beyond the scope of this article. But what is suggested, at least from the 28 structures in this study, is that the marked hierarchy of interactions that is a characteristic of small-molecule crystal structures could be bypassed in biological struc-

tures. One of the consequences of this is acceptor-furcation. Put another way, nature seems to want to use the weaker C—H…O hydrogen bonds in order to optimize the efficiency of protein–ligand interaction. By involving more donor groups, whatever their chemical character, a larger number of interactions come into play even if individual interactions are neither as short nor as linear as the ones found in small-molecule structures. By using more interactions, specificity of recognition is increased. Because individual interactions are weaker, reversibility is possible.

Frequency Distributions of Hydrogen Bonds for Different Amino Acid Residues

Figure 6(a–f) shows the histograms for the different types of amino acid residues involved in N—H…O and O—H…O hydrogen bonds with the ligand. The N—H…O donor capabilities of the nonpolar Gly in the main-chain [Fig. 6(a)] and charged side-chain residues, namely, Arg and Lys [Fig. 6(f)] are noteworthy. The significant numbers of hydrogen bonds formed by Arg and Lys highlight the importance of electrostatics in protein–ligand interactions. Figure 6(g–l) shows the corresponding distributions for C—H…O interactions. The C_α—H…O bond is seen very prominently for the Gly residue [Fig. 6(g)]. This is interesting in light of the earlier observation[10] wherein the nonparticipation of the Gly C_α—H group in C—H…O interactions was ascribed to the highly flexible nature of the residue. While Gly is undoubtedly flexible, such flexibility also allows for induced fit recognition leading in turn to the exceptionally high frequency of C_α—H…O bonds. In this scheme of things, the high frequency of Gly hydrogen bonds may be a feature of protein–ligand complexes and not necessarily of native protein structures. The presence of two C_α—H atoms as compared to only the one present in all other residues is an additional factor that affects this distribution. Both these features are inherent to Gly, and this makes it an important residue in ligand binding. Figure 7(a) shows an example of a Gly C_α—H…O interaction with the ligand (PDB ID: 3CHB).[28] A consequence of enhanced donor capability of the C_α—H of Gly may be profitably exploited in the drug design—one may design acceptor functionalities placed complementarily to the C_α—H groups of the Gly residue in the binding site. Very recently, Scheiner and coworkers[29] have studied the C_α—H donor capability of several amino acids by ab initio calculations on the C_α—H…O/water system. For nonpolar residues like Gly, Ala, and Val, the interaction energies (MP2/6–31 + G**) are −2.5, −2.1, and −2.0 kcal/mol, respectively. These results correlate well with our own analysis.

The side-chain donor capabilities of Tyr present an interesting comparison with regard to C—H…O and O—H…O bonds [Fig. 6(e and k)]. The numbers of C—H…O and O—H…O interactions are comparable. Although it is difficult to draw firm conclusions with such meager data, this observation suggests that one should not underestimate the donor capability of the aromatic C—H protons of the Tyr side-chain even in the presence of the stronger

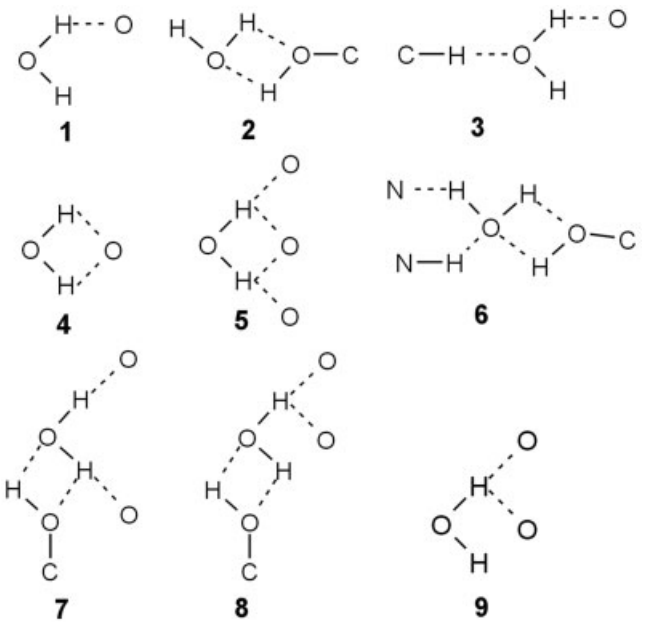

Fig. 8. Selected synthons with ligand–water interactions.

O—H donor. This capability could be inherent or statistical in that there are four C—H groups to every O—H group in the Tyr ring. Figure 7(b) shows an example (PDB ID: 1EQO)[30] where the Tyr acts as a C—H…O donor. The acidic C—H group of the aromatic ring makes one short and one long bond with the phosphate group of the ligand, hinting that the C—H group is activated by the adjacent O—H group.

The hydrogen bonding capabilities of the side-chain atoms of the nonpolar residues are particularly noteworthy [Fig. 6(j)]. In the absence of other, stronger hydrogen bond functionalities, the C—H donors satisfy the hydrogen bonding potential. The Phe residue with an aromatic ring in the side-chain shows a marked preference for such an interaction. Recent ab initio studies[31] predict the C—H group in Trp to be a more effective donor than that in Phe. While C—H (Trp) is undoubtedly more activated than C—H (Phe), the Trp residue appears to utilize the N—H group rather than C—H for hydrogen bond formation [Fig. 6(d)] accounting for our observation. These observations highlight the importance of C—H…O hydrogen bonds in the hydrophobic environment of proteins. Figure 7(c) shows a Phe residue that forms a C—H bond with the ligand (PDB ID: 1I12).[27] This aspect has also been pointed out by Moras and Klaholz, who identify a C—H…O hydrogen bond as the driving force for ligand selectivity of the retinoid SR11254 in the hydrophobic pocket of the retinoic acid receptor (RARγ).[15] The C—H…O interaction assumes lesser importance, as expected, for amino acids with charged side-chain residues. Arg in particular shows an enhanced tendency to form N—H…O hydrogen bonds [Fig. 6(f)].

Lipinski's "Rule of Five" Extended?

The total number of strong hydrogen bonds (N—H…O, O—H…O) made by the various ligands with the protein

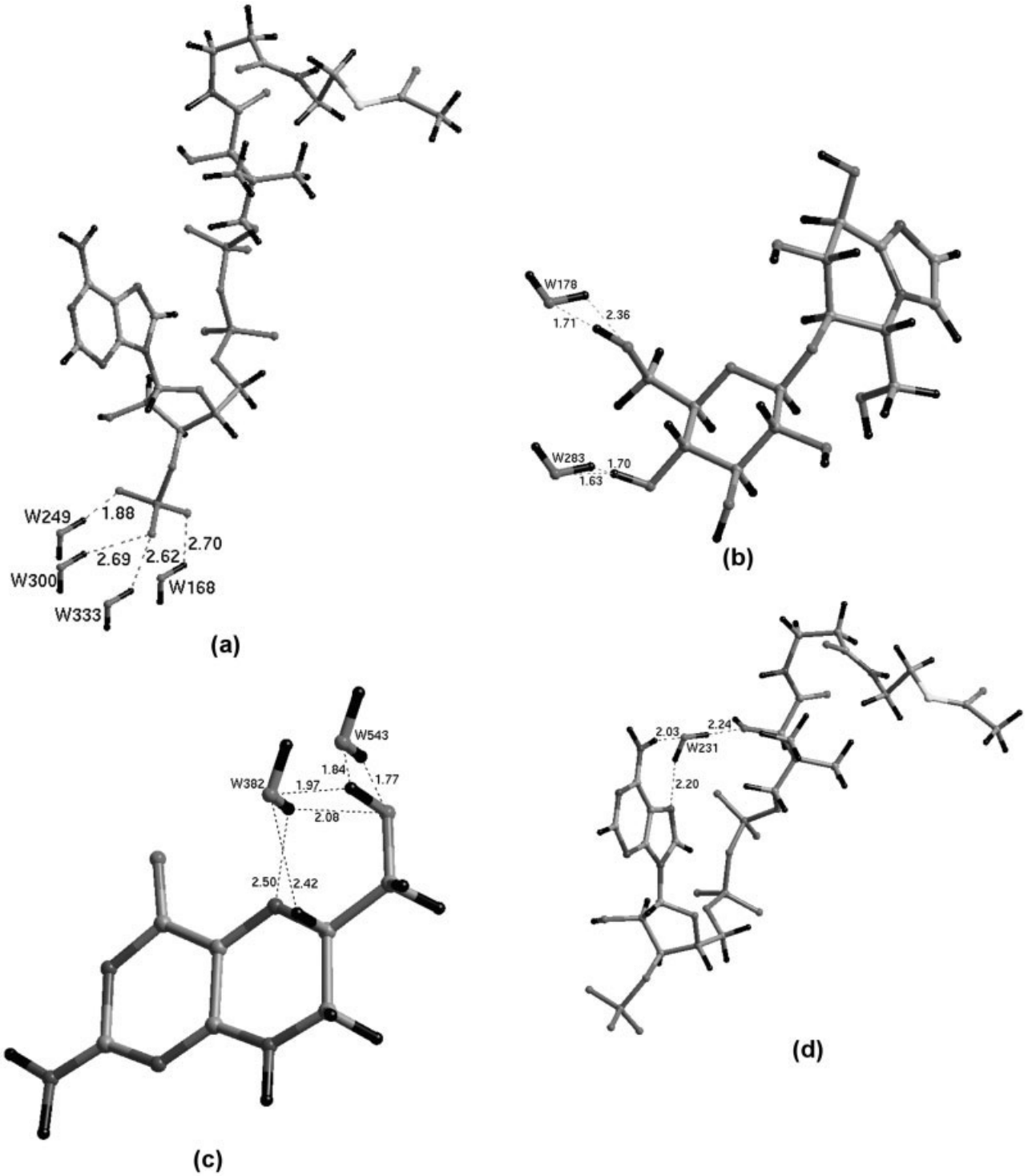

Fig. 9. (**a**) Synthon **1** with four water molecules that satisfy the coordination of the ligand phosphate group (PDB ID: 1I12). (**b**) Tandem hydrogen bonds formed by water with the free ligand O—H groups (PDB ID: 8A3H). (**c**) A unique example where a single free ligand O—H group is involved in tandem hydrogen bonds with two different water molecules (PDB ID: 1EQO). (**d**) Cooperative (assisted by a ligand-donated N—H…O_w bond) donation of two hydrogen bonds from water to ligand (PDB ID: 1I12).

and water molecules stand at 111 (donor) and 297 (acceptor) in the 28 structures studied here. This averages out to be, approximately, 4 (donor) and 10 (acceptor) hydrogen bonds per ligand molecule. These numbers may be important in understanding the thermodynamics of protein–ligand binding and its extension to rational drug design, because they match with Lipinski's "rule of five,"[32] which predicts the potentiality of a

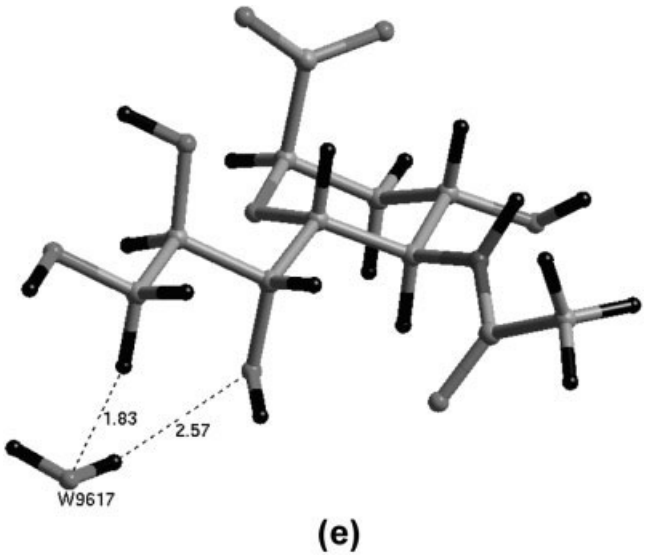

Fig. 9. (**e**) Tandem hydrogen bond formed by an unusually short C—H…O_w and a long O_w—H…O (PDB ID: 3CHB).

molecule to be a drug. The rule states that a maximum of 5 hydrogen bond donor and 10 acceptor functionalities should be present in a molecule (with a molecular weight of ~500) for it to be a useful drug. In general, Lipinski's rule has to do with the hydrophobic–hydrophilic balance in the drug molecule in question.[5] However, the numbers of hydrogen bond donors and acceptors are known to affect the physicochemical properties (solubility, adsorption, distribution) of a molecule and, hence, the efficacy of a drug. Therefore, the coincidence in the numbers of hydrogen bond donors and acceptors at the protein–ligand interface and Lipinski's rule may not be just accidental. The binding of a ligand to a protein requires desolvation of both the protein and the ligand molecules, since both are solvated in their unbound states. For favorable protein–ligand interaction, the enthalpic gain by hydrogen bonding between protein and ligand should exceed the enthalpy loss consequent upon desolvation. Hydrogen bonding functionalities in the ligand, which do not bind to protein residues (due to lack of steric and geometrical complementarity), cannot contribute to this payoff for desolvation. Thus, it is suggested that Lipinski's rule may also have implications in the context of intermolecular interactions.

Ligand–Water Interaction

All protein–ligand interactions must compete with interactions with water; both the protein and ligand are solvated before complexation and lose their solvation shell on complex formation. Conversely, the entropic cost of trapping highly mobile water molecules in the binding site is large. However, in favorable cases, these losses are suitably compensated by the enthalpic gain resulting from water-mediated hydrogen bonds. In effect, the balance between enthalpic and entropic contributions is a fine one, and for a water molecule to be able to contribute to binding affinity, it has to be in a binding site that provides the maximum number of hydrogen bond partners at the optimum distance and orientation.

The distribution of hydrogen bond metrics for the ligand–water interaction shows considerable diversity. The angular distribution of donor water is clearly bimodal (supplementary information). The distance distribution of HO_w—H…O interactions shows a considerable population (65%) in the 2.2–2.7 Å range. This is significantly different from the 14% population found in small-molecule structures. The scatterplot of C—H…O_w interactions from the ligand shows considerable scatter, indicating that these interactions are largely isotropic. However, the uncertainties involved in determining the correct water positions in the crystallographic refinement model and the inadequate treatment of C…O contacts in the crystallographic refinement force fields may well contribute to the above observations.

We have analyzed the different synthons involved between the ligand and water in the binding site (Fig. 8). Interactions between the protein and water were not examined, since they are too numerous. Synthon **1** is the most commonly observed motif (~65%). A good example is shown Figure 9(a) (PDB ID: 1I12).[27] The propensity for the formation of the "tandem" motifs (**2**, **6**, and **8**) is ~13%. Figure 9(b) (PDB ID: 8A3H)[33] shows one such example. The formation of "tandem" or *flip-flop* hydrogen bonds as low-energy intermediates could have significance in maintaining a continuum in the dynamics of bond formation and breaking. Figure 9(c) (PDB ID: 1EQO)[30] shows a good example in which a single O—H group is involved in the formation of two tandem motifs with two water molecules. The effect is enhanced by additional donors, such as N—H (synthon **6**) or O—H. Synthon **3**, found in ~8% abundance, is basically an extension of synthon **1** with cooperative assistance from a C—H…O_w bond. Figure 9(e) (PDB ID: 3CHB)[28] shows an unusually short C—H…O_w contact and a rather long O_w—H…O hydrogen bond. The relative frequencies of the functionalities depend on the density of the acceptor atoms.[34] The high frequency of occurrence of synthon **1** is in accord with this idea; the ligand in the binding site is in a donor-rich environment, and there is little need for water to expand its donor functionality toward the ligand.

In general, water prefers to accept hydrogen bonds from O—H and N—H donors in macromolecular structures.[4] The C—H…O_w interaction is only a "filler" and is resorted to in the absence of stronger donors. In ligand–water interactions, however, we find that the C—H…O_w interactions constitute as much as 54% of the total number of ligand(donor)–water(acceptor) interactions, while the O—H…O_w and N—H…O_w account for only 27% and 19%, respectively. The fact that 85% of the interactions are "two-centered" means that the two-fold acceptor functionality of water is mostly not satisfied. The high number of C—H…O_w interactions could follow from the fact that ligands bind to the protein relatively strongly and, hence, are likely to have used up their conventional hydrogen bonding capacity in their "main" interaction with the

binding site. Another reason for the large number of C—H…O_w bonds could be that they can compensate for the desolvation cost of the strong hydrogen bond functionalities of the ligand; yet they reduce the entropy penalty somewhat by a binding that is "loose" when compared with the strong hydrogen bonds formed by (immobilized) water at the binding surface. In this way, the enthalpy–entropy balance is sustained. Figure 10(a) (PDB ID: 1HDO)[35] and (b) (PDB ID: 3CHB)[28] depict two examples from this study where (a) two C—H…O_w interactions from the ligand satisfy the two-fold acceptor functionality of water, and (b) one C—H…O_w (weak) and one O—H…O_w (strong) hydrogen bond perform the role. In the drug-design context, one may utilize the stronger acidic protons of the ligand for C—H…O_w interactions to lend specificity to the interaction with conserved water molecule(s) that act as a bridge between the ligand and protein molecules. Recent studies show that, while water is observed in protein–ligand complexes, it is not found in the active sites of the uncomplexed protein.[36] This suggests the role that such water molecules play in ligand binding.

CONCLUSIONS

This study provides insights into the relative importance of strong (N—H…O, O—H…O) and weak (C—H…O) hydrogen bonds in protein–ligand complexes. The stronger N—H…O and O—H…O interactions show slight but definite deviations from linearity. Multifurcation of strong with weak hydrogen bonds is common in the structures in this study. The propensity of occurrence of acceptor-furcated bonds, or anticooperative interactions, justifies the need to consider a more liberal distance cutoff criterion for these interactions. For the side-chain O—H…O interaction, an even more flexible cutoff may be allowed. The acidic C_α—H groups preferably form C—H…O hydrogen bonds to the ligands. In particular, the C_α—H donor capability of the nonpolar Gly residue is noteworthy. C—H…O hydrogen bonds are of utility in the hydrophobic binding pocket of proteins. It is possible that they compete favorably with the N—H…O and O—H…O bonds because of the smaller penalty of desolvation associated with their formation.[16] They also satisfy the hydrogen bonding potential of the macromolecule in the absence of stronger hydrogen bond donors and thus contribute enthalpically. The C—H…O bond occupies a middle ground between the highly directional N—H…O and O—H…O bonds, and the directionless van der Waals interactions. All of this calls for a revision in the force fields, so that short C…O contacts are correctly identified as being stabilizing in crystallographic least squares refinement programs and in molecular dynamics simulations. The inadequacies in the treatment of the nonbonded part of the potential energy function could be improved by taking into account these rather dispersive weak interactions.

In the functional genomics era,[37] when much effort is being targeted toward developing high-throughput crystallization and structure determination in drug discovery,[38] systematic studies of protein–ligand complexation provides information on the nature and role of noncovalent interactions and, in particular, the weaker ones. Typically, within a family of protein–ligand complexes, one might attempt to find patterns of weak interactions that are conserved. Alternatively or additionally, one may add or remove specific interactions by site-directed mutagenesis experiments and substitute desired amino acid residue(s) in the binding site in order to more quantitatively assess the changes in binding affinity and/or ligand specificity in complex formation. However, the intrinsic nature of macromolecular systems means that there is a compromise between several interactions of varying strengths. The stronger interactions contribute enthalpically, but the weaker interactions, due to their inherent flexibility, make only modest entropic demands on the system. In this way, the enthalpy–entropy balance is maintained. This study is limited by the paucity of sufficiently high-resolution crystal structures of protein–ligand complexes. This has always been a problem,[39] but by its very nature, this difficulty diminishes with time. If the earlier experiences with the CSD are any indication,[4,20,21] the trends observed in this study will only be reinforced as the size of the PDB increases.

ACKNOWLEDGMENTS

We are grateful to Accelrys, Inc. for their continued cooperation.

REFERENCES

1. Behr J-P, editor. The lock and key principle: perspectives in supramolecular chemistry. Chichester, UK: Wiley; 1994. Vol. 1.
2. Desiraju GR. Crystal engineering: the design of organic solids. Amsterdam: Elsevier; 1989.
3. Glusker JP. Directional aspects of intermolecular interactions. In: Weber E, editor. Topics in current chemistry: design of organic solids. Berlin and Heidelberg: Springer-Verlag; 1998. p 1–56.
4. Jeffrey GA, Saenger W. Hydrogen bonding in biological structures. Berlin: Springer-Verlag; 1991.
5. Desiraju GR, Steiner T. The weak hydrogen bond in structural chemistry and biology. New York: Oxford University Press; 1999.
6. Abraham MH, Ibrahim A, Zissimos AM, Zhao YH, Comer J, Reynolds DP. Application of hydrogen bonding calculations in property based drug design. Drug Discov Today 2002;7:1056–1063.
7. Desiraju GR. Hydrogen bridges in crystal engineering: interactions without borders. Acc Chem Res 2002;35:565–573.
8. Derewenda ZS, Lee L, Derewenda U. The occurence of C—H…O hydrogen bonds in proteins. J Mol Biol 1995;252:248–262.
9. Wahl MC, Sundaralingam M. C—H…O hydrogen bonding in biology. Trends Biochem Sci 1997;22:97–102.
10. Fabiola GF, Krishnaswamy S, Nagarajan V, Pattabhi V. C—H…O hydrogen bonds in β sheets. Acta Crystallogr D 1997;53:316–320.
11. Bella J, Berman HM. Crystallograhic evidence for C_α—H…O hydrogen bonds in a collagen triple helix. J Mol Biol 1996;264:734–742.
12. Chakrabarti P, Chakrabarti S. C—H…O hydrogen bond involving proline residues in alpha-helices. J Mol Biol 1998;284:867–873.
13. Babu MM, 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.
14. Jiang L, Lai L. C—H…O hydrogen bonds at the protein–protein interfaces. J Biol Chem 2002;277:37732–37740.
15. Moras D, Klaholz BP. C—H…O hydrogen bonds in the nuclear receptor RARγ—a potential tool for drug selectivity. Structure 2002;10:1197–1204.
16. Pierce AC, Sandretto KL, Bemis GW. Kinase inhibitors and the case for C—H…O hydrogen bonds in protein–ligand binding. Proteins 2002;49:567–576.

17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN. Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nuc Acid Res 2000;28:235–242. Available online at http://www.rcsb.org/pdb.
18. Allen FH, Kennard O. 3D search and research using the Cambridge Structural Database. Chem Des Autom News 1993;8:31–37. Available online at http://www.ccdc.cam.ac.uk/prods/csd/csd.html.
19. Cerius^2 Program. Molecular Simulations, 9685 Scranton Road, San Diego, CA 92121-3752 (USA) & 240/250 The Quorum, Barnwell Road, Cambridge CB5 8RE (UK).
20. Desiraju GR. The C—H…O hydrogen bond: structural implications and supramolecular design. Acc Chem Res 1996;29:441–449.
21. Wheeler K. Statistically significant conclusions from the Cambridge Structural Database. In: Whitesell JK, editor. Organised molecular assemblies in the solid state. Chichester, UK: Wiley; 1999. p 141–183.
22. Böhm H-J, Klebe G. What can we learn from molecular recognition in protein–ligand complexes for the design of new drugs? Angew Chem Int Ed Engl 1996;35:2588–2614.
23. Taylor R, Kennard O, Versichel W. The geometry of the N—H…O═C hydrogen bond: 3. Hydrogen-bond distances and angles. Acta Crystallogr 1984;B40:280–288, and the two previous articles in the series.
24. Baker EN, Hubbard RE. Hydrogen bonding in globular proteins. Progr Biophys Mol Biol 1984;44:97–179.
25. Desiraju GR. Supramolecular synthons in crystal engineering: a new organic synthesis. Angew Chem Int Ed Engl 1995;34:2311–2327.
26. Fulop V, Moir, JWB, Ferguson SJ. Hajdu J. The anatomy of a bifunctional enzyme: structural basis for reduction of oxygen to water and synthesys of nitric oxide by cytochrome Cd1. Cell 1995;81:369–377.
27. Peneff C, Mengin-Lecreulx D, Bourne Y. The crystal structures of apo and complexed *Saccharomyces cerevisiae* Gna1 shed light on the catalytic mechanism of an amino-sugar *N*-acetyltransferase. J Biol Chem 2001;276:16328–16334.
28. Merritt EA, Kuhn P, Sarfaty S, Erbe JL, Holmes RK, Hol WG. The 1.25 Å resolution refinement of the cholera toxin B-pentamer: evidence of peptide backbone strain at the receptor-binding site. J Mol Biol 1998;282:1043–1059.
29. Scheiner S, Kar T, Gu Y. Strength of the C—H…O hydrogen bond of the amino acid residues. J Biol Chem 2001;276:9832–9837.
30. Blaszczyk J, Shi G, Yan H, Ji X. Catalytic center assembly of Hppk as revealed by the crystal structure of a ternary complex at 1.25 Å resolution. Structure 2000;8:1049–1058.
31. Scheiner S, Kar T, Pattanayak J. Comparison of various types of hydrogen bonds involving aromatic amino acids. J Am Chem Soc 2002;124:13257–13264.
32. 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 Del Rev 1997;23:3–25.
33. Varrot A, Sculein M, Pipelier M, Vasella A, Davies GJ. (PDB citation. To be published). Lateral protonation of a glycoside inhibitor: structure of the *Bacillus agaradhaerens* Cel5A in complex with a cellobiose-derived imidazole at 0.97 Å resolution.
34. Steiner T, Saenger W. Role of C—H…O hydrogen bonds in the coordination of water molecules: analysis of neutron diffraction data. J Am Chem Soc 1993;115:4540–4547.
35. Pereira PJB, Macedo-Ribeiro S, Parraga A, Perez-Luque R, Cunningham O, Darcy K, Mantle TJ, Coll M. Structure of human biliverdin Ix beta reductase, an early fetal bilirubin Ix producing enzyme. Nat Struct Biol 2001;8:215–220.
36. Ni H, Sotriffer CA, McCammon JA. Ordered water and ligand mobility in the HIV-1 integrase-5CITEP complex: a molecular dynamics study. J Med Chem 2001;44:3043–3047.
37. Nagano N, Orengo CA, Thornton JM. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol 2002;321:741–765.
38. Stewart L, Clark R, Behnke, C. High-throughput crystallization and structure determination in drug discovery. Drug Discov Today 2002;7:187–196.
39. Klebe G. The use of composite crystal-field environments in molecular recognition and the *de novo* design of protein–ligands. J Mol Biol 1994;237:212–235.