J Mol Model (2006) 12: 436–444
DOI 10.1007/s00894-005-0067-x

ORIGINAL PAPER

Simona Cotesta · Martin Stahl

The environment of amide groups in protein–ligand complexes: H-bonds and beyond

Received: 12 May 2005 / Accepted: 18 October 2005 / Published online: 13 December 2005

Abstract A comprehensive structural analysis of interactions involving amide NH and C=O groups in protein–ligand complexes has been performed based on 3,275 published crystal structures (resolution≤2.5 Å). Most of the amide C=O and NH groups at the protein–ligand interface are highly buried within the binding site and involved in H-bonds with corresponding counter-groups. Small percentages of C=O and NH groups are solvated or embedded in hydrophobic environments. In particular, C=O groups show a higher propensity to be solvated or embedded in a hydrophobic environment than NH groups do. A small percentage of carbonyl groups is involved in weak hydrogen bonds with CH. Cases of dipolar interactions, involving carbonyl oxygen and electrophilic carbon atoms, such as amide, amidinium, guanidium groups, are also identified. A higher percentage of NH are in contact with aromatic carbons, interacting either through hydrogen bonds (preferably with the NH group pointing towards a ring carbon atom) or through stacking between amide plane and ring plane. Comprehensive studies such as the present one are thought to be important for future improvements in the molecular design area, in particular for the development of new scoring functions.

Keywords Protein–ligand interactions · Weak polar interactions · Structural database · Molecular modeling

S. Cotesta · M. Stahl
Molecular Design, Pharmaceutical Division,
F. Hoffmann-La Roche AG, Discovery technologies,
Bldg. 092/2.10D,
CH-4070, Basel, Switzerland

S. Cotesta (✉)
Computer Assisted Drug Design, Novartis Pharma AG,
Novartis Institute for Biomedical Research (NIBR),
Bldg. WKL 136.P15,
CH-4002, Basel, Switzerland
e-mail: simona.cotesta@novartis.com

Introduction

Over the past decade, the amount of crystal-structure information available both of small organic molecules and protein–ligand complexes has increased so significantly that our ability to use this wealth of information for structure-based design is still lagging behind. The analysis of this structural data can yield new insights into both conformational preferences of small molecules and intermolecular interactions, but only few comprehensive and systematic studies of this kind have been undertaken. For example, several types of non-covalent interactions are yet to be exploited fully in the context of drug design, e.g. weak H-bonds involving CH groups as donors [1–7], interactions between donors and π-systems as acceptors [7–11] and so-called dipolar interactions [12]. We believe that a lot can be learned from crystal structures, even about those non-covalent interactions that are seemingly well understood, such as classical hydrogen bonds or hydrophobic interactions. Our thinking about these interactions is often reduced to simplified positive statements about the most commonly seen interaction types. For example, in drug design one aims at satisfying hydrogen-bond donors with acceptors, because it is known that hydrogen bonds can be important for recognition and orientation of a ligand within the binding site. With the exception of buried, charged hydrogen bonds, they are not as important for affinity prediction because the newly formed hydrogen bonds between ligand and receptor compete with hydrogen bonds formed with water. Other types of environments are in most cases completely neglected, whereas they could be also favorable for both pose prediction and hopefully for affinity prediction. Little is known about the relative propensity of a particular functional group to form one interaction or another, and we hardly know anything about the borderline between what is tolerated in a protein–ligand complex and which geometric arrangements are so unlikely (energetically unfavorable) that they will never be observed. To a great extent it is this lack of information that causes the still poor performance of empirical scoring functions in docking and *de novo* design tools.

A full understanding of the interactions of a particular functional group requires the analysis of the crystal environment of all its occurrences in a structural database. This is not possible with currently available query interfaces for the CSD (Quest) [13] and the PDB (Relibase+, [14]), since these allow the user to single out specific non-bonded interactions, but offer no assistance in generating a global view of the interactions a functional group may form.

We decided to start a research program aimed at a systematic increase in our knowledge of non-bonded interactions through data mining in crystal-structure databases. We implemented new search techniques to address simple questions regarding recurring functional groups of drug-like molecules. Here we present first results on amide carbonyl groups and amide NH groups. To what extent are they involved in forming hydrogen bonds? In what environment are they found if they do not form hydrogen bonds?

Material and methods

Data set selection

Protein–ligand complexes were extracted from the PDB [15] and filtered using Reliscript [16]. Complexes containing only non-covalently bound ligands with a resolution below 2.5 Å were selected. Structures classified as DNA or RNA were eliminated, as were protein structures binding only pure peptides, sugars, nucleic acids or common cofactors as ATP, ADP, GTP, and hemes. These selection criteria led to a set of 3,275 PDB entries. Ligands were extracted and saved in mol2 format. Protein binding sites were extracted in PDB format. Residues with at least one atom at a distance below 7 Å from any ligand atom were defined as part the binding site. Structural water molecules positioned in the binding sites were stored as part of the binding site.

Interaction analysis tool

A computational tool for analyzing protein–ligand interactions and ligand conformations was developed as a Python (v2.3.2) script using several OEChem [17] modules. The program calculates distances, angles, and torsion angles between atoms matched by user-defined SMARTS [18] sub-structures. Ring centers can be defined and used as dummy atoms, planes can be defined, and angles between planes or between vectors and planes can be calculated. SMARTS searches can be made very generic through the use of various types of wildcards. Query results can be aligned and visualized in a multi-mol2 file. In addition to these classical ingredients of 3D database search tools, further descriptors are calculated. The accessible solvent surface area (SASA) and the hydrophobic buried surface (HBS) are determined for each atom matching the query by placing a spherical distribution of points [19] around the atom at a radius of 1.4 Å plus the van der Waals radius of the atom. Points lying within the van der Waals radius of another atom are considered inaccessible. We do not calculate absolute surface areas but employ a normalized %SASA value with the accessible surface of the query atom (considering its covalently bound neighbors) as a reference state, i.e. for any given atom type a %SASA value of 0 means that the atom is completely buried in the binding site, whereas a %SASA value of 100 corresponds to a completely accessible atom. The hydrophobic buried surface (HBS) is defined as the fraction the buried surface of an atom occluded by hydrophobic atoms. In this context, hydrophobic atoms are defined as any carbon or divalent sulfur atoms.

Environment of acceptors and donors

The environment of amide C=O and NH groups was examined in both proteins and ligands. All atoms within a sphere of 4.0 Å around the amide oxygen/nitrogen were retrieved. For the initial classification, the closest atom in the environment (the "contact atom") was selected. Depending on the atom type and distance of the contact atom to the query atom, it was determined whether the query atom forms a classical H-bond. In this fashion, 98% of the H-bonds are identified; only in 2% of the cases is a hydrogen bond formed with a more distant atom. The cut-off contact distance for H-bond assignment was 3.5 Å. For ligands, the Tripos mol2 atom types N.3, N.2, N.1, O.3, O.2, N.ar, O.co2 were treated as acceptors, and N.3, N.2, O.3, N.pl3, N.am, N.4 were treated as donors. In the case of protein atoms, backbone and carbonyl oxygen were considered as acceptors, nitrogen atoms as donors. The hydroxyl groups of Thr, Tyr and Ser, histidine nitrogen atoms and carboxylic acid oxygen atoms were considered as either donors or acceptors. For amide groups within ligands and protein binding sites, potential hydrogen bonds to water were searched separately. Intramolecular H-bonds within the protein were detected separately.

Results and discussion

Classical H-bond formation and solvation

Not surprisingly, the majority of the C=O and NH groups in amides, both in ligands and proteins, are involved in H-bonds. 80% of protein C=O groups (9,466) and 85% of the protein NH groups (9,421) form a classical H-bond with ligand or protein donors or water. Eighty three percent of the ligand C=O groups (2,161) and 72% of the ligand NH groups (1,909) forms an H-bond. Protein NH groups more often form an intermolecular H-bond (56%) than protein C=O groups do (31%). The latter tend to interact more often with water or protein donors. At first sight, this may seem surprising, since in ligands the percentage of carbonyl and NH groups in contact with proteins is similar: 60% for carbonyls and 65% for NH groups. This discrepancy arises from the fact that ligands on average possess more

acceptors than donors. If we restrict ourselves to a subset of cases with exactly equal donor and acceptor distribution, proteins and ligands yield the same statistics: considering interactions between secondary amides only in both proteins and ligands, one finds that 89% of the protein NH groups form an intermolecular H-bond with the ligand amide carbonyl group and that 97% of the protein C=O groups form an H-bond with a ligand amide NH.

Hydrogen bonds to solvent were considered where water molecules were explicitly present in the binding site. In the absence of explicit water molecules, the degree to which a functional group may be solvated was assessed by determining of the fraction of its free surface area (%SASA). With a threshold value as low as 10%, 84% of protein amide carbonyls and 91% of the protein amide NH groups are buried within the binding site and hence not solvent accessible. A similar situation is found when considering ligand amides: 78% of the ligand carbonyls and 88% of the ligand NH groups are buried within the binding site. These findings are summarized in Fig. 1. As expected, the percentage of donor/acceptor groups with high accessible solvent surface is larger for the groups not involved in H-bonds. If we assume that functional groups without explicit H-bond partner but a high %SASA value are forming H-bonds to water, we arrive at a total percentage of H-bonds of 86 and 91% for protein and ligand carbonyl groups, and 88 and 78% for protein and ligand NH groups.

Taken together, these observations indicate that most of the donor/acceptor groups are forming hydrogen bonds. For molecular design, the crystal structure data seem to confirm the well-known rule that H-bond donors as well as acceptors should be satisfied with appropriate partners. We were interested to learn more about the small percentage of NH and C=O groups not forming H-bonds. In which kind of environment are these donors and acceptors groups enclosed? In what types of interaction are they involved?

Focus on groups not forming H-bonds

We define the hydrophobic buried surface (HBS) of an atom as the contribution of hydrophobic atoms to the solvent-inaccessible part of an atom's surface. In general, most of the C=O and NH groups are in a mixed environment. HBS distributions for our dataset resemble Gaussian curves with peaks at 60%; for functional groups not involved in H-bonds the distribution is slightly shifted towards more hydrophobic environments with peaks at 70%. Fig. 2 shows the distribution of %HBS for protein and ligand C=O (Fig. 2a,c) and NH (Fig. 2b,d) groups not involved in H-bonds. Histogram bars are subdivided into sections denoting the type of atom closest to the selected functional group. Protein groups (a,b) and ligand groups (c,d) yield very similar histograms. Distributions for C=O groups (a) and NH groups (b) have the same overall shape, but differ in the type of contact atom. Carbonyl groups are typically in contact with a hydrophobic atom (mostly carbon, sulfur in only five cases). The %HBS distribution for NH groups shows a significantly higher number of acceptor contact atoms (red bars in Fig. 2). These are interactions that were not classified as H-bonds because the donor-acceptor distance is larger than 3.5 Å. They could be considered as "weak classical" H-bonds. These findings suggest that in structure-based design embedding carbonyl groups in a hydrophobic environment may be tolerated, whereas NH groups generally require a more polar environment. Under what circumstances do we find either functional group in contact with hydrophobic atoms?

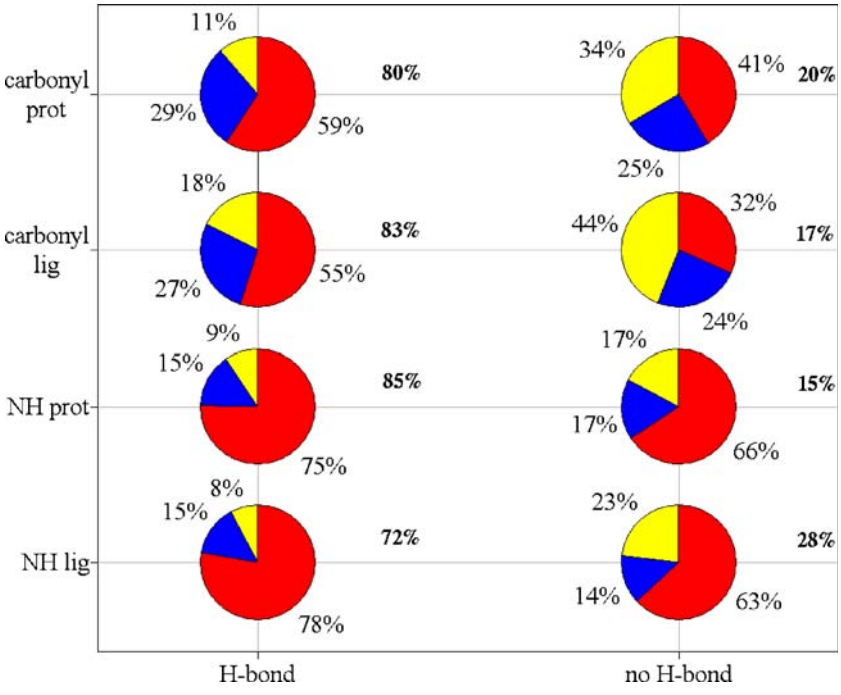
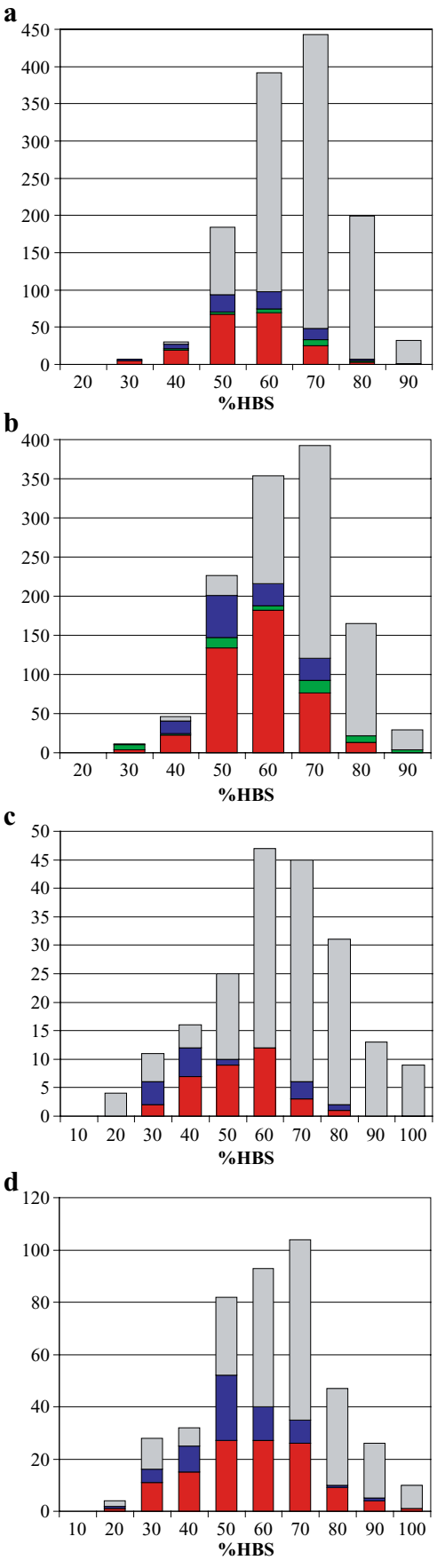


Fig. 1 Percentage of amide protein/ligand carbonyls and nitrogens involved in H-bonds and their solvent accessible surface area (%SASA). Colors: *red*: %SASA=0, *blue*: %SASA≤10, *yellow*: %SASA>10. On the *left*, the represented functional groups are forming H-bonds, on the *right* the functional groups are not involved in H-bonds. H-bond definition is based on donor–acceptor heavy atom distance with cutoff at 3.5 Å. On the *Y* axis the four different analyzed situations are reported. In *bold*, nearby each pie chart, the percentage of records, taken into account in the analysis is reported



◀ **Fig. 2** Distribution of the percentage of hydrophobic surface for functional groups not involved in H-bonds and buried in the binding site (%SASA<=10). **a** protein C=O, **b** protein NH, **c** ligand C=O, **d** ligand NH. Color codes represent the atom type of the contact atoms: *red*: acceptors, *blue*: donors, *green*: halogen, *gray*: hydrophobic (carbon or sulfur)

A small percentage of C=O and NH groups not forming H-bonds and with %SASA≤10 have a carbon atom as the closest non-bonded contact atom: 10% and 7% of protein and ligand carbonyls, 6 and 12% of protein and ligand NH groups. A number of interactions between formally unpolar and polar atoms have been extensively studied, e.g. weak hydrogen bonds between donors and aromatic ring π acceptors [8, 9, 11, 20–25] or weak H-bonds involving CH donors [1–6]. Dipolar interactions between electrophilic carbon atoms and electron-rich heteroatoms have recently been described in detail [12]. Such interactions have been observed in X-ray structures of both small organic molecules and macromolecules. In the following paragraphs, we attempt to classify C=O and NH groups with carbon atoms as closest nonbonded neighbors in terms of such known interaction types.

Carbonyl groups in contact with carbon

Amide carbonyl groups not forming H-bonds, inaccessible to solvent (%SASA=10%) and in contact with a carbon atom were classified into three subgroups, (1) C=O groups forming dipolar interactions with electrophilic carbon atoms contained in polar groups such as amide, amidinium or guanidinium groups, (2) C=O groups interacting with aromatic carbon atoms and (3) and C=O groups forming weak H-bonds with CH groups.

The number of protein carbonyl groups forming so-called dipolar interactions was found to be relatively low in the subset of non-hydrogen bonding C=O groups under investigation here. In particular five (0.5% of the sub-group of carbonyls interacting with carbons) dipolar interactions involving protein carbonyl groups interacting with amide carbons and one (0.6%) interaction involving ligand oxygen and amide protein carbons were identified. A much larger number of such interactions can be found considering all C=O groups regardless of their environment and of the types of additional interactions they may form. Obviously, dipolar interactions typically occur within an interaction network where the C=O groups are involved in other strong interactions. In Fig. 3, two examples of dipolar interactions are depicted: dipolar interaction occurring between a ligand amide oxygen, already interacting through an H-bond with the protein molecule, and a protein carbon amide (Fig. 3a) and the much rarer case of a protein amide carbonyl group not interacting through an H-bond with the ligand but forming an H-bond to a ligand amide carbon (Fig. 3b).

Among those dipolar interactions that were identified, the majority occurs in serine protease structures binding ligands with amidine and guanidine groups within the S1 pocket (53 interactions, 5%). This interaction is observed

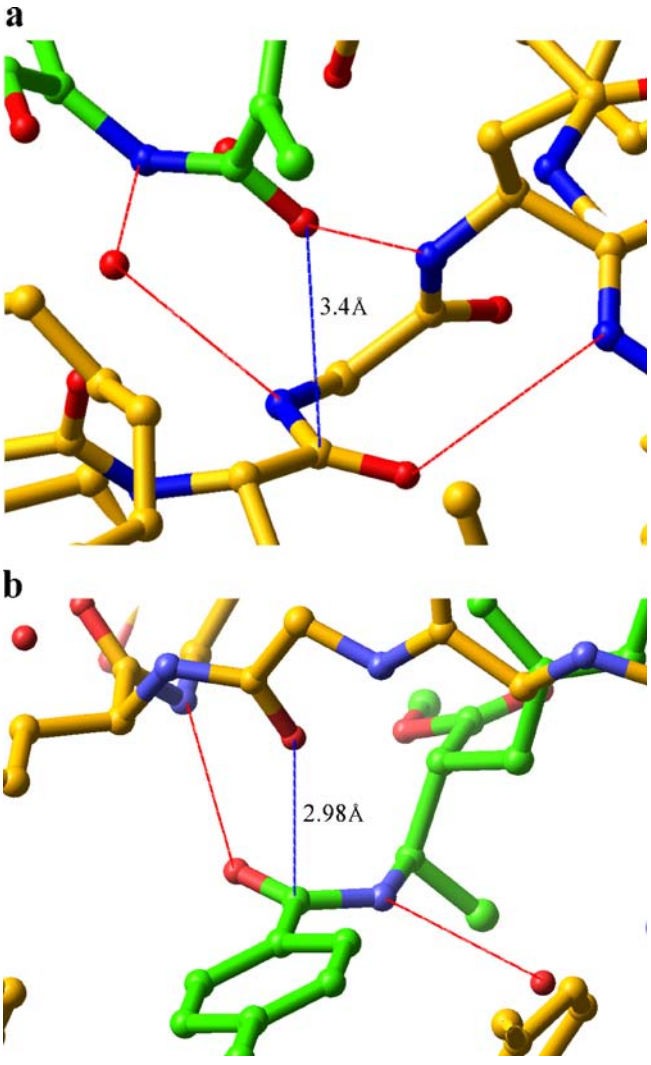


Fig. 3 Example of dipolar interactions between carbonyl oxygen and amide carbons. **a** Dipolar (O carbonyl and C amide) interaction within an interaction network (*red line* H-bonds, *blue line* dipolar interaction). Protein: PDB 1a4h, [28] N terminal domain of the yeast HSP90 chaperone in complex with geldanamycin; **b** Dipolar interaction not in an interaction network (*red line* H-bonds, *blue line* dipolar interaction). Protein: PDB 1ezq, [29] Factor Xa

often simply because of the large number of such structures in the PDB. In all cases, the carbonyl oxygen points towards the central amidine carbon with the angle C=O...C greater than 100°. An example of this kind of interaction is shown in Fig. 4. Only one ligand C=O group was found interacting in this way with the guanidinium carbon CZ of an arginine.

Secondly, amide C=O groups in contact with aromatic carbon atoms were investigated. A total of 360 (36%) protein C=O groups and 41 (27%) ligand C=O groups fall into this class. Angle distributions between the amide plane and the ring plane and between the normal to the ring plane and the vector connecting the carbonyl oxygen to the ring center were analyzed. The distributions obtained in this way reveal that there exists no preferred geometry (plots not shown). The only avoided geometric arrangement is the

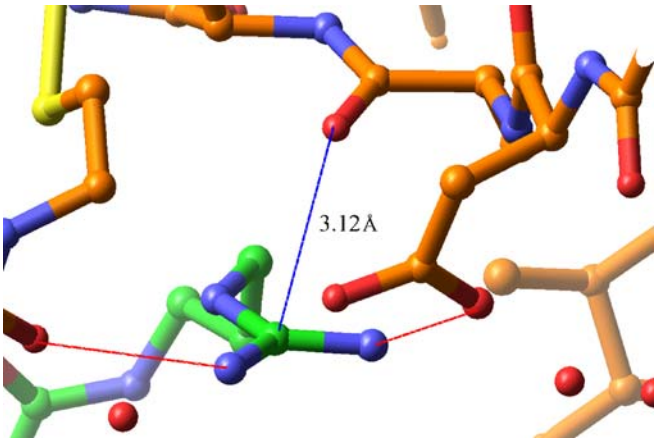


Fig. 4 Example of a dipolar interaction between backbone carbonyl oxygen and gaunidinium carbon. The distance between oxygen and carbon is 3.12 Å and the angle between the carbonyl O formed with the protein carbon and ligand carbon is 113°. The PDB protein entry is 1k2l, [30] a human thrombin (Ser-protease family)

one in which the carbonyl oxygen would point towards the ring center, forming an angle with the normal to the plane smaller than 30°. This is in accordance with classical electrostatics considerations. The electron-rich C=O oxygen and aryl ring π system should repel each other. Alternatively, CH groups in aromatic rings could interact with the carbonyl oxygen forming a weak H-bond. Only few interactions of this kind (45, 4%) could be identified, without showing a clear geometrical preference.

Finally, carbonyls interacting with aliphatic carbons were analyzed. A total of 317 (32%) interactions with C ···· O distance <3.8 Å were found. Weak H-bonds between CH groups and acceptors are well documented [1, 6], and the strength of the interaction is clearly correlated with the acidity of the donor hydrogen. A thorough assessment of the acidity of individual CH groups would be beyond the scope of this work. Instead, we considered interactions to CH groups in α-position to heteroatoms. The frequency distribution of the C···O distances shows a regular Gaussian distribution with a peak at 3.3 Å, coinciding with the sum of the van der Waals radii of carbon and oxygen. However, a significant number of interactions also occurs at distances of about 3.1 Å, indicating an attractive interaction. Only few ligand carbonyl groups were found to interact in this way with protein CH groups. In nine cases (6%), distances below 3.5 Å were measured to C_α carbon atoms. The remaining interactions between amide C=O groups and carbon atoms are at larger distances (> 3.8 Å) and were not further analyzed.

Carbonyl groups in contact with acceptors

Judging from the histogram in Fig. 2a,c a small percentage of amide carbonyls seems to be in contact with acceptor groups (red). In most cases, the closest contact atom is further away than 3.5 Å, alleviating electrostatic repulsion. In a number of cases, the electrostatic interaction is a secondary interaction, the primary and attractive one being

a hydrogen bond to the carbonyl group (Fig. 5a). In a number of cases, it is quite likely that the actual orientation of an amide group is flipped and no real repulsion is seen (Fig. 5b). It can be concluded that interactions between C=O and acceptors are (as of course assumed) strongly avoided unless compensated by an attractive interaction.

NH groups in contact with carbon

Amide NH groups not forming H-bonds and inaccessible to solvent (%SASA≤10%) in contact with carbons were further classified into those interacting with aromatic carbon atoms and others interacting with aliphatic carbon atoms. Approximately half of the amide NH groups were found to be in contact with aromatic carbon atoms: 350 interactions for protein amide NH groups (57%), 118 (51%) for ligand amide NH groups. NH donors can form weak hydrogen bonds to π-acceptors. These interactions have been observed in proteins [9, 20], small molecules [26] and protein–ligand complexes [7, 27]. Alternatively, an amide may stack onto the plane of a phenyl ring. In geometric terms, these interactions can be characterized by the angle ω formed by the N-ring center vector and the normal to the ring plane. When this angle is less than about 25°, the amide N is approximately above the ring center. Stacking and H-bonding can be further distinguished by the angle formed by the amide plane and the ring plane. If the two planes are parallel, the interaction is stacked; when they are nearly perpendicular, the interaction is an H-bond. An additional H-bonding geometry also described in the literature is the one in which the N is positioned above one of the ring carbon atoms [26]. The latter geometry has also been defined as H-bond. We found all three geometric arrangements. In Fig. 6a, the distribution of the angle ω is shown. The histogram was generated from both ligand and protein amide NH groups and shows a first peak at 20° and a second larger peak at 35°. The first peak represents either a stacked or an H-bond interaction with the NH pointing towards the ring center, whereas the second peak represent H-bonds with the amide nitrogen pointing directly towards the carbon ring atoms. The prevalence of this geometry was already observed in interactions between NH and phenyl in small molecule X-ray structures [26]. In Fig. 6b, the angle between amide plane and ring planes is plotted as a function of the angle ω. In this way it is possible to distinguish between stacked interactions and H-bonds. The stacked geometry clearly occurs more often than the H-bond. In

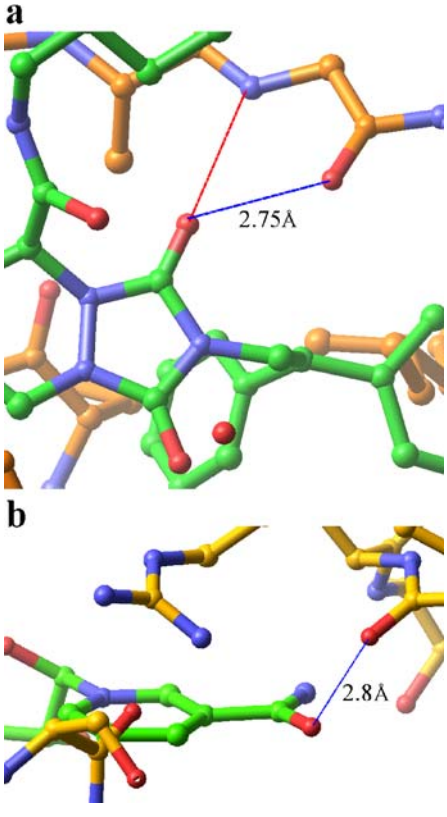


Fig. 5 Carbonyl groups in contact with other carbonyl groups: **a** one of the carbonyl oxygen is interacting through H-bonds with the protein backbone (PDB 1c4v, [31] thrombin); **b** none of the carbonyls interacts favorably (PDB 1hwx, [32] bovin liver glutamate dehydrogenase)

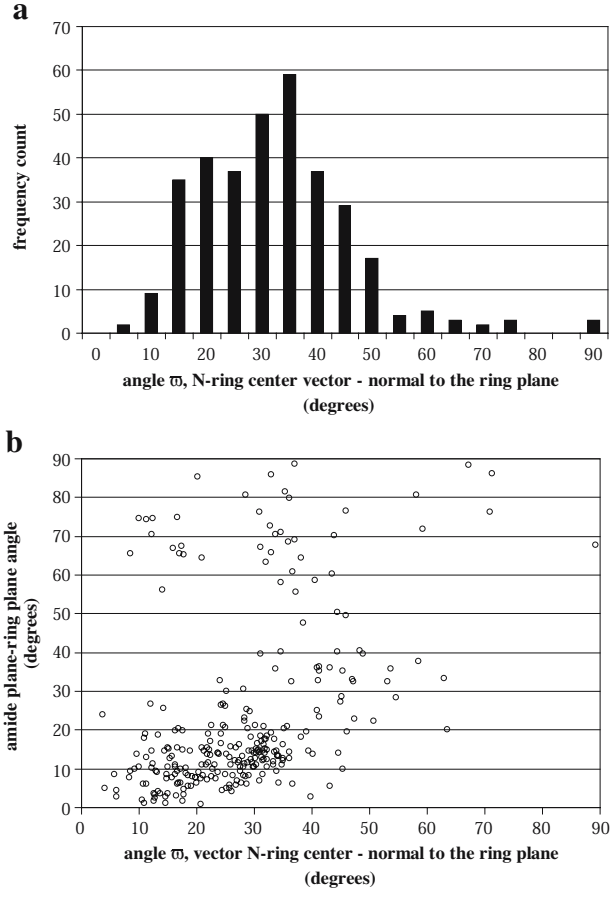


Fig. 6 NH aromatic π acceptors: **a** distribution of the angle ω, formed by the vector *N* to the ring center and the normal to the ring plane; **b** *scatter plot* showing the angle among the amide plane and the ring plane as a function of the ω angle

Fig. 7, three examples showing the three different geometries are shown: a) H-bonds with the NH pointing towards the ring center, b) stacked and c) H-bond with NH pointing directly towards an aromatic ring carbon. These interactions take place in every type of environment and for a diverse set of protein structures. The widespread occurrence of the interaction between amide NH and aromatic

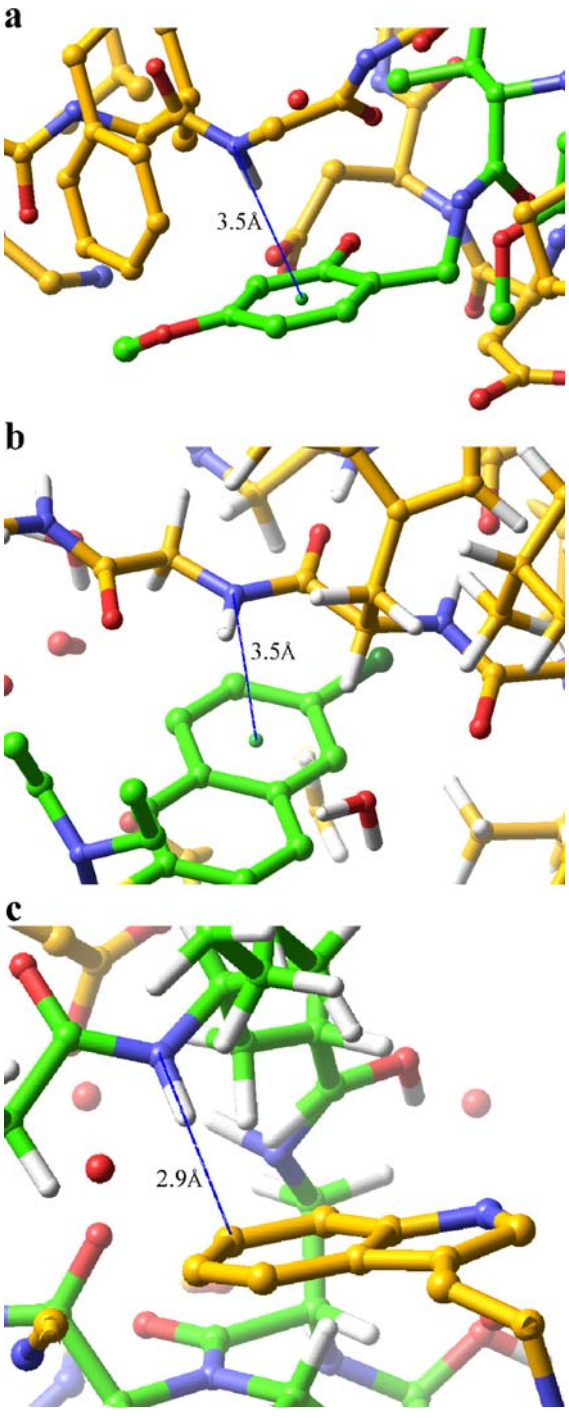


Fig. 7 NH aromatic π acceptors: **a** H-bonds (PDB, 1a8g, [33] HIV-1 protease); **b** stacked (PDB, 1ql7, factor Xa; **c** NH pointing directly towards an aromatic ring carbon (PDB, 1tmb, [34] alpha thrombin)

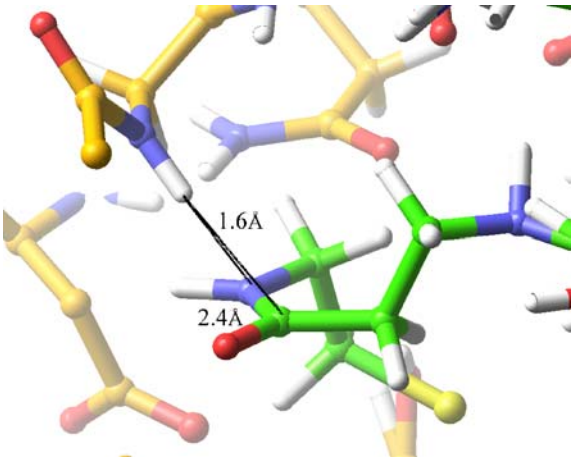


Fig. 8 Protein amide N interacting with π-electrons of a ligand amide (PDB 6csc, chicken citrate synthase)

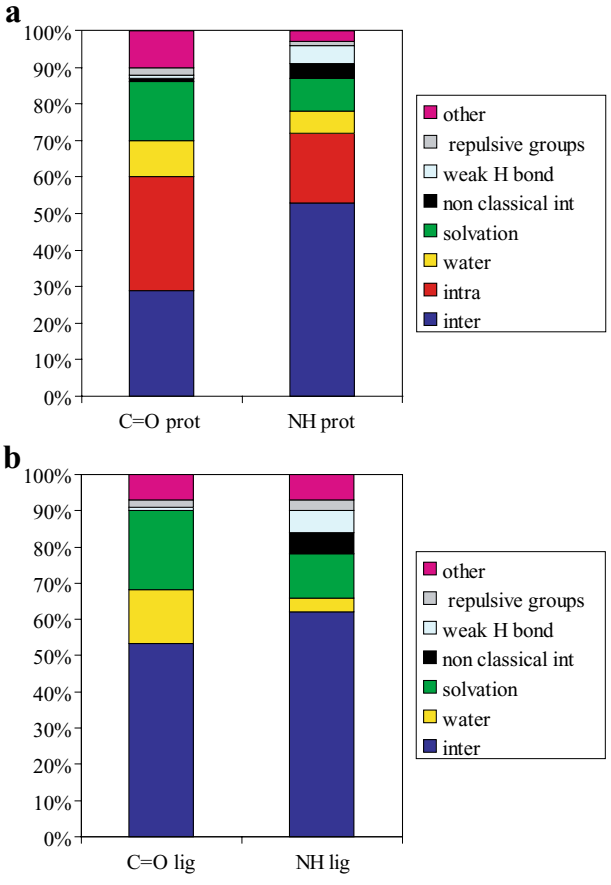


Fig. 9 Percentage of functional groups participating in classical and non classical interactions: **a** protein carbonyl and NH groups, **b** ligand carbonyl and NH groups. Colors: *blue*, inter-molecular H-bonds, *red*, intra-molecular H-bonds, *yellow*, H-bonds with water molecules, *green*, functional groups exposed to the solvent, *black*, H-bonds interacting through "non classical" interactions, i.e. CH... C=O and NH...aromatic ring, *light cyan*, weak classical H-bonds (donor–acceptor distance greater than 3.5 Å), *gray*, repulsive interactions (donor–donor or acceptor–acceptor), *magenta*, functional groups not involved in any specific interactions and in contact with a hydrophobic atom

ring systems indicates that it should be added to the standard repertoire of interactions considered by molecular design experts and automated tools.

NH groups in contact with donors

A small percentage of NH groups have H-bond donors as closest contact atoms. As in the case of repulsive interactions to carbonyl groups, the majority of donor–donor interactions are at a distance greater than 3.5 Å. Secondary electrostatic interactions occurring in conjunction with hydrogen bonds again play a large role. Several cases should not actually be classified as electrostatically repulsive interactions, much rather, two amides stack onto each other with the N atoms coming particularly close. This type of arrangement occurs in particular between ligand amide NH and Asn and Gln protein amides. A perpendicular arrangement between two amides, such as the one depicted in Fig. 8, might be classified as a donor–π-acceptor interaction. Direct NH-donor contacts are clearly avoided.

Conclusions

Our results show that most of the amide C=O and NH groups at the protein–ligand interface are highly buried (% SASA<10) within the binding site and involved in H-bonds with corresponding counter-groups. In Fig. 9a,b the interaction types formed by protein and ligand amide groups, as found in PDB protein–ligand complexes, are summarized. All in all, the percentage of protein carbonyl groups not performing specific interactions and in contact with hydrophobic atoms (magenta) is greater than the corresponding percentage of protein NH groups, whereas similar percentages have been found in the case of ligand carbonyl and NH groups. Carbonyl groups, both in ligand and in proteins, show a higher propensity for being in contact with water (yellow, green) than amide NH groups. Truly repulsive interactions are tolerated in very few cases for both functional groups. The presence of these unfavorable ligand–protein contacts can be explained as a consequence of packing effects. A ligand could be allowed to have unfavorable contacts only if other favorable contacts compensate. The propensity of NH groups to form NH-π H-bonds is higher than the propensity of C=O groups to form H-bonds to CH groups. NH-π H-bonds are typically formed with the NH group off the center of the aromatic ring, as found before in an analysis of small molecule X-ray structures from the Cambridge Structural Database [26]. If all H-bonds—classical, nonclassical and long-distance H-bonds—are taken together, it seems that amide NH groups in proteins have a stronger tendency to form hydrogen bonds than carbonyl groups do. The small number of dipolar interactions identified in our searches indicates that these interactions typically take place in interaction networks.

We note that the present study does not allow us to rank order the various types of interactions identified in an energetic fashion, even though rarely observed geometric arrangements are not likely to be strongly attractive in nature. In this context, it is important to remember that, in spite of the large number of entries in the PDB, there is still a significant bias towards protein–ligand complexes of particular classes of enzymes and ligands. This makes it necessary to validate any statistical trends carefully and to correct them if necessary. Also, we have here used a rigid interaction classification scheme, thereby neglecting the fact that intermolecular interactions always occur in networks and that a single functional group may simultaneously participate in various types of interactions.

Nevertheless, our findings suggest that rule-based systems, such as scoring functions, which to date distinguish between van-der-Waals or lipophilic contacts and classical hydrogen bonds only, might be augmented by directional and angular terms describing weaker hydrogen bond type interactions. Further rules penalizing geometric arrangements not found in protein–ligand complexes should further enhance the success rate of structure-based design. NH groups pointing into lipophilic pockets, and without the chance to form NH-π H-bonds, should clearly be penalized, whereas less stringent penalties should be applied to C=O groups in a lipophilic environment.

We believe that the analysis and systematic exploration of the full environment of a functional group in all database entries is a key to successful rule building and knowledge generation. For the work presented here, we have restricted ourselves to a very limited set of functional groups. Clearly, many further studies are required to liberate the knowledge hidden in structural databases, and flexible database search tools are a key prerequisite to conduct them.

Acknowledgements The authors thank Prof. Klaus Mueller for stimulating discussions about non-bonded interactions, Dr. Harald Mauser and Dr. Olivier Roche for their suggestions and all colleagues at Roche Basel for a supportive research atmosphere.

References

1. Derewenda ZS, Lee L, Derewenda U (1995) J Mol Biol 252:248–262
2. Fabiola GF (1997) Acta Crystallogr D Biol Crystallogr 53:316–320
3. Klaholz B, Moras D (2002) Structure (Camb) 10:1197–1204
4. Pierce AC, Sandretto KL, Bemis GW (2002) Proteins 49:567–576
5. Manikandan K, Ramakumar S (2004) Proteins 56:768–781
6. Sarkhel S, Desiraju GR (2004) Proteins 54:247–259
7. Desiraju GR, Steiner T (1999) The weak hydrogen bond in structural chemistry and biology. Oxford University Press, Oxford
8. Steiner T (2002) Angew Chem Int Ed Engl 41:49–76
9. Steiner T, Koellner G (2001) J Mol Biol 305:535–557
10. Steiner T (2002) Biophys Chem 95:195–201
11. Steiner T, Schreurs AM, Kanters JA, Kroon J (1998) Acta Crystallogr D Biol Crystallogr 54 (Pt 1):25–31
12. Paulini R, Muller K, Diederich F (2005) Angew Chem Int Ed Engl (in press)
13. Allen FH (2002) Acta Crystallogr B 58:380–388
14. Hendlich M, Bergner A, Gunther J, Klebe G (2003) J Mol Biol 326:607–620

15. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235–242
16. Bergner A, Gunther J, Hendlich M, Klebe G, Verdonk M (2001) Biopolymers 61:99–110
17. OEChem Toolkit (2004) 1.3.2 OpenEye Scientific Software, Inc. (http://www.eyesopen.com)
18. James CA, Weininger D, Delany J (2001) Daylight Theory Manual. Daylight Chemical Information System, Inc, Los Altos
19. Stahl M, Bur D, Schneider G (1999) J Comput Chem 20:336–347
20. Mitchell JB, Nandi CL, McDonald IK, Thornton JM, Price SL (1994) J Mol Biol 239:315–331
21. Scheiner S, Kar T, Pattanayak J (2002) J Am Chem Soc 124:13257–13264
22. Thomas A, Meurisse R, Charloteaux B, Brasseur R (2002) Proteins 48:628–634
23. Toth G, Watts CR, Murphy RF, Lovas S (2001) Proteins 43:373–381
24. Weiss MS, Brandl M, Suhnel J, Pal D, Hilgenfeld R (2001) Trends Biochem Sci 26:521–523
25. Worth GA, Wade RC (1995) J Phys Chem-Us 99:17473–17482
26. Malone JF, Murray CM, Charlton MH, Docherty R, Lavery AJ (1997) J Chem Soc Faraday T 93:3429–3436
27. Sampson NS, Vrielink A (2003) Acc Chem Res 36:713–722
28. Prodromou C, Roe SM, O'Brien R, Ladbury JE, Piper PW, Pearl LH (1997) Cell 90:65–75
29. Maignan S, Guilloteau JP, Pouzieux S, Choi-Sledeski YM, Becker MR, Klein SI, Ewing WR, Pauls HW, Spada AP, Mikol V (2000) J Med Chem 43:3226–3232
30. Dullweber F, Stubbs MT, Musil D, Sturzebecher J, Klebe G (2001) J Mol Biol 313:593–614
31. Krishnan R, Mochalkin I, Arni R, Tulinsky A (2000) Acta Crystallogr D Biol Crystallogr 56 (Pt 3):294–303
32. Peterson PE, Smith TJ (1999) Structure Fold Des 7:769–782
33. Ringhofer S, Kallen J, Dutzler R, Billich A, Visser AJ, Scholz D, Steinhauser O, Schreiber H, Auer M, Kungl AJ (1999) J Mol Biol 286:1147–1159
34. Maryanoff BE, Qiu X, Padmanabhan KP, Tulinsky A, Almond Jr HR, Andrade-Gordon P, Greco MN, Kauffman JA, Nicolaou KC, Liu A, Bungs PH, Fusetani N (1993) Proc Natl Acad Sci USA 90:8048–8052