Received: 23 April 2013, Revised: 23 August 2013, Accepted: 07 October 2013, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jmr.2332

Structural analysis of protein-ligand interactions: the binding of endogenous compounds and of synthetic drugs

Anna M. Gallina[a], Peer Bork[b] and Domenico Bordo[a]*

The large number of macromolecular structures deposited with the Protein Data Bank (PDB) describing complexes between proteins and either physiological compounds or synthetic drugs made it possible a systematic analysis of the interactions occurring between proteins and their ligands. In this work, the binding pockets of about 4000 PDB protein-ligand complexes were investigated and amino acid and interaction types were analyzed. The residues observed with lowest frequency in protein sequences, Trp, His, Met, Tyr, and Phe, turned out to be the most abundant in binding pockets. Significant differences between drug-like and physiological compounds were found. On average, physiological compounds establish with respect to drugs about twice as many hydrogen bonds with protein atoms, whereas drugs rely more on hydrophobic interactions to establish target selectivity. The large number of PDB structures describing homologous proteins in complex with the same ligand made it possible to analyze the conservation of binding pocket residues among homologous protein structures bound to the same ligand, showing that Gly, Glu, Arg, Asp, His, and Thr are more conserved than other amino acids. Also in the cases in which the same ligand is bound to unrelated proteins, the binding pockets showed significant conservation in the residue types. In this case, the probability of co-occurrence of the same amino acid type in the binding pockets could be up to thirteen times higher than that expected on a random basis. The trends identified in this study may provide an useful guideline in the process of drug design and lead optimization. Copyright © 2014 John Wiley & Sons, Ltd.
Additional supporting information may be found in the online version of this article at the publisher's web site.

Keywords: protein-ligand interaction; lead optimization; secondary target; drug discovery

INTRODUCTION

Many biological processes are based on the selective interaction between proteins and small molecules that occur through the formation of a transient complex between the small molecule, hereinafter referred to as the ligand, and the target protein. In most cases, the binding site is located at the protein surface and displays a concave shape, hence the name binding pocket. Although in some instances, especially those involving enzymes, the docking of the ligand to the target protein leads to the transient formation of inter-molecular covalent bonds (e.g., Fersht, 1993; Singh *et al*., 2011), usually the protein-ligand complex formation is characterized by the presence of intermolecular hydrogen bonds, enhanced sometimes by Coulomb interactions, and of van der Waals interactions. These interactions are responsible for the establishment of the protein-ligand selectivity upon which a large fraction of biochemical reactions is based. Of particular relevance for the biomedical research are the cases in which a drug molecule displays affinity for distinct biological targets. In fact, the unwanted cross-reactivity for secondary targets represents an element of major concern in the development of new drugs, as this is often associated to deleterious side effects, especially in chemotherapy. In some instances, however, spurious cross-reactivity may reveal alternative and new therapeutic possibilities for commercialized drugs (Campillos *et al*., 2008), in particular for orphan diseases (Ekins *et al*., 2011; Sardana *et al*., 2011; Dakshanamurthy *et al*., 2012).

About one-fourth of the entries contained in the current release of the Protein Data Bank (PDB, Dutta *et al*., 2009) describe complexes between proteins and small molecular compounds, not including the light compounds commonly used in buffers or in crystallization solutions, such as salts or heavy atom compounds. Furthermore, the PDB contains numerous instances in which a specific ligand molecule is found in distinct protein complexes. In these cases, the comparison of the residues present in the binding pockets may reveal hidden relationships between ligand specificity and binding pocket composition. If the proteins involved are homologous, structurally equivalent binding pocket amino acids, as well as the respective interaction with the bound ligand, can be compared. An example on this regard is Imatinib, an anticancer drug that binds specific kinases. In the PDB, twelve distinct kinase complexes could be found, with similarity among each other ranging between 24% and 98% identical residues. On the other hand, the comparison of the binding pocket composition can also be carried out when the involved proteins are not homologous. In these cases, even

* *Correspondence to: Domenico Bordo, IRCCS Az. Ospedaliera Universitaria S. Martino – IST – National Cancer Research Institute. Largo R. Benzi, 10, 16132 Genova, Italy.*
E-mail: bordo@fisica.unige.it

[a] *A. M. Gallina, D. Bordo*
IRCCS Az. Ospedaliera Universitaria S. Martino – IST, National Cancer Research Institute, Largo R. Benzi, 10, 16132 Genova, Italy

[b] *P. Bork*
European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany

in the lack of structural equivalence, it is still possible to compare the composition of the respective binding pockets with the aim of identifying the co-occurrence of amino acid or interaction types in distinct target proteins.

In recent years, several studies of protein-ligand interactions found in experimentally determined three-dimensional protein-ligand complexes have been carried out, some focused on the residues directly involved in the catalytic process (Zvelebil and Sternberg, 1988; Bartlett *et al*., 2002), other based upon specific databases (e.g., Wang *et al*., 2005; Benson *et al*., 2007; Reddy *et al*., 2008). In this work, the analysis of the interaction between ligands and target proteins was carried out in a systematic survey of the PDB focused on all the amino acids in interaction with the bound molecule. Ligands were classified as “drugs” or “compounds” according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa *et al*., 2012). Statistical preferences in binding pocket amino acid and interaction types were separately deduced for drugs and compounds, and differences between the two types of ligands were analyzed. Furthermore, conserved features observed in families of homologous proteins bound to the same ligand, or the co-occurrence of the same amino acid or interaction type in the binding pocket of proteins belonging to distinct homology families, but in complex with the same ligand, were deduced and discussed. The results described here, especially those hinting at the differences between natural compounds and artificial drug molecules, might provide an useful guide in lead identification and optimization. Those involving groups of unrelated proteins bound to the same ligand could be especially useful in the analysis of off-target interactions in the context of drug repurposing.

METHODS

The present analysis is based on experimentally determined protein-ligand complexes deposited with the Protein Data Bank (PDB). Covalently-bonded protein-ligand complexes were not included, as the formation and breakage of covalent bonds involve free energies which are usually orders of magnitude greater than those associated to non-covalent interactions. As the analysis involves the pairwise comparison among all the proteins in complex with a selected ligand, to avoid exceedingly long computational times ligands found in more than 25 PDB entries were excluded from the analysis. In this way, small molecules displaying broad but poor specificity for proteins, such as buffer components or compounds used in X-ray crystallography such as additives or heavy atom compounds, usually present in a large number of PDB entries, were not considered. Also, the instances in which the binding pocket was located at the interface between polypeptide chains were excluded. Although the binding at the protein-protein interface is frequent and often functionally relevant (e.g., Fuller *et al*., 2009), the associated functional and evolutionary constraints are conceivably more complex than those occurring on a protein-small molecule complex involving a single polypeptide chain. The analysis of the binding at the protein-protein interface will be the focus of further studies.

In the PDB, ligands are referred to as heterogeneous compounds and are identified with a HET_ID code. The current version of PDB contains about 12,600 distinct heterogeneous compounds. To identify differences between natural and synthetic compounds, we used the KEGG database (Kanehisa *et al*., 2012), that classifies a part of the ligand present in the PDB as either “drug” or “compound”. Some ligands have both drug and compound classifications. We obtained KEGG classification by either using cross references to KEGG present in the PDBsum (www.ebi.ac.uk/pdbsum; Laskowski *et al*., 2005), or by querying the ChemSpider Web server (www.ChemdSpider.com) with the InChi code associated to the ligand. The InChi code could be determined by querying the DrugBank (www.DrugBank.ca; Knox *et al*., 2011). This resulted in the selection of 1118 distinct ligand and 3992 PDB entries on which the analyses described below in the following texts were carried out. The data mining and the analyses described here were performed by several software tools written in PERL. These scripts allow a fully automated periodic scan of the PDB for new entries describing complexes with known ligands, as well as for the inclusion of new ligands. The updated description of the binding pockets and the associated comparisons for ligands with no more than 25 PDB entries are stored in the PLI database (protein-ligand interactions; Gallina *et al*., 2013) and available on the web site http://bioinformatics.istge.it/pli/.

Identification of the amino acids of the binding pocket

For each protein-ligand complex, the amino acids forming the binding pocket, that is, the amino acids in interaction with the bound ligand, were determined with the program LigPlot (Wallace *et al*., 1995) as made available on the PDBsum Web pages. The interactions were classified as either hydrogen bonds or van der Waals interactions.

Identification of the group of homologous proteins bound to the same ligand

For each ligand included in this study, the associated PDB entries were obtained from the PDBsum web site and subsequently sorted in homology families according to the mutual primary sequence similarity. The homology was determined by using the SAS section of the PDBsum that, with the primary sequence of each PDB entry used as query and having the PDB database as a search space, provides the output of the FASTA algorithm (Pearson and Lipman, 1988). Only the PDB entries included in the FASTA output describing complexes with the selected ligand and having an associated expectation value smaller than, or equal to 10^{-3} (conventionally considered a positive indication of homology) were included in the same homology group of the query. In order to maximize the size of each homology group and to minimize the number of distinct groups, the procedure was recursively repeated for each PDB entry containing the selected ligand.

Comparison of the binding pockets of homologous proteins

With the purpose of avoiding overrepresentation of very similar proteins in complex with the same ligand (e.g., point mutated proteins or identical proteins solved in different laboratories), only proteins having at most 95% identical residues were considered. This resulted in the identification of 299 distinct ligands having at least two homologous PDB entries. The binding pockets of each pair of homologous proteins were compared and the equivalent residues, namely, those occupying the same position in the FASTA pair wise alignment, were scrutinized. To analyze the conservation of each residue type, all pair wise comparisons were carried out within each homology family and for each ligand. The degree of conservation of each amino acid type was compared with the overall pair wise sequence similarity. For this purpose, seven similarity bins associated to the pair wise primary sequence similarity were defined, ranging from 0.25 to 0.95 fraction of identical residues, each 0.1 wide. In detail, the similarity intervals were 0.25-0.35, 0.35-0.45, 0.45-0.55, 0.55-0.65, 0.65-0.75, 0.75-0.85, and 0.85-0.95. In each pair wise comparison, the residue types forming the binding pocket and their conservation were separately tallied in the appropriate similarity bin. Subsequently, the fraction of conserved residues was calculated for each similarity bin and each residue type. Finally, these fractions were compared with the expected average amino acid mutation rate for that bin, approximated by the middle similarity value of the bin. The ratio between the two values, referred to as the Conservation Index, represents therefore the propensity of an amino acid type to be conserved when part of a the binding pocket with respect to the rest of the protein. An example of calculation of the Conservation Index, together with the list of the Conservation Indexes for each similarity shell have been provided in Supplementary Materials (see the foot note of Table S1). As large fluctuations were sometime observed among contiguous similarity bins, likely i.e. the reduced number of instances, the Conservation Index was calculated for each amino acid type as the weighted average value of the seven similarity bins.

Comparison of the binding pockets for non-homologous proteins

In the cases in which a selected ligand was found in the PDB in complex with proteins belonging to distinct homology families, that is, having different three-dimensional fold (for this purpose, we used the CATH classification of Sillitoe *et al*., 2013), the type and number the amino acid forming the binding pocket could still be compared. In detail, one member of each family was randomly selected, and pair wise comparisons were carried out between proteins belonging to distinct homology families. In this case, the pair wise comparison was carried out by considering in turn each protein as reference. Due to the lack of structural equivalence, each amino acid of the binding pocket of the reference protein was considered co-occurring in the second protein if a residue of the same type was found in the binding pocket of the second protein. The procedure was recursively repeated on all amino acids of the binding pocket of the reference protein, with the condition that each amino acid of the second binding pocket could be counted (at most) only once. After all pair wise comparisons, for each residue type the total number of co-occurring instances was compared with the total number of occurrences. The ratio between the two numbers, referred to as "probability of co-occurrence" represents the probability to observe a given amino acid type, found in the binding pocket of a selected protein in complex with a specific ligand, also in the biding pocket of an unrelated protein in complex with the same ligand. A similar analysis was also carried out for conserved chemico-physical properties such as aromatic or basic/acidic side chains; in these cases, the presence of an aromatic or of a basic/acidic side chain, respectively, was tallied with the same rule of counting only once occurrences in binding pockets.

RESULTS

In this work the non-covalent intermolecular interactions observed in protein-ligand complexes having known three-dimensional structure were analyzed. The protein-ligand interactions were classified as hydrogen bonds or van der Waals interactions. The analysis did not include the cases in which the small molecule interacts with more protein subunits as this situation, albeit common, is likely subject to more complex structural and evolutionary constraints. The analysis involved 1118 ligands, out of the 12,600 distinct heterogeneous compounds found in the current PDB release, for which a KEGG classification in either "compound" and/or "drug" was provided. Drugs include most prescription and "over the counter" (OTC) drugs, whereas compounds are small molecules of biological origin present as either endogenous or exogenous compounds in living organisms. In detail, of the ligands included in the analysis, 760 were compounds, of which 341 occurring in human pathways (Figure 1), 81 ligands were drugs (including nutriaceuticals), while the remaining 277 ligands were classified as both compounds and drugs, of which 95 were present in human metabolic pathways. Furthermore, of the total number of ligands considered here, 652 are classified as enzyme substrates or cofactors (Figure 1).

Amino acid composition of the binding pockets of drugs and compounds

To identify trends involving the amino acids in contact with the bound ligand, the average composition of the binding pockets was determined for the 3992 protein-ligand complexes considered here (see Methods Section). The observed binding pocket amino acid frequencies were compared to the average amino acid composition observed in proteins (Creighton, 1993; see Figure 2(a)). No significant difference in the binding pocket composition was found between drugs and compounds (Figure 2(b)). Notably, for both drugs and compounds, five of the six rarer amino acids in protein sequences, Trp, His, Met, Tyr, and Phe, display instead higher propensity for being in binding pockets, with frequency for Phe and Tyr of 8.0% and 7.7%, respectively (Figure 2(a), and (b)). The same analysis was also carried out on the complexes of human proteins with endogenous compounds (i.e., present in the human metabolic pathways), of which 113 instances were found, but no significant deviation from the trends represented in Figure 2 was observed (not shown). These results can be compared with those deduced from a study on the protein-protein interaction aimed at identifying hot

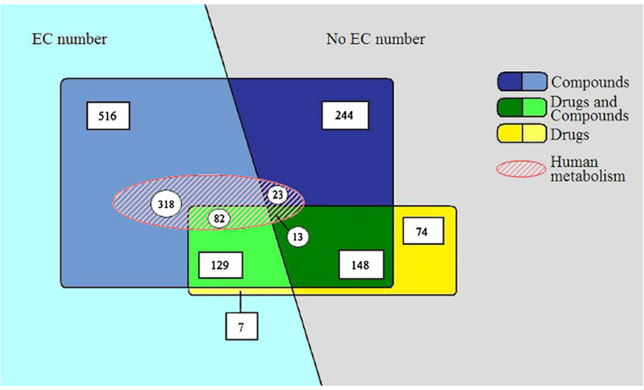

Figure 1. Classification of the 1118 ligands included in this study. The number of ligands having a KEGG classification in drug and/or compound, or having an assigned EC number are schematically represented. Ligands present in the human metabolism and in complex with human proteins are also shown.

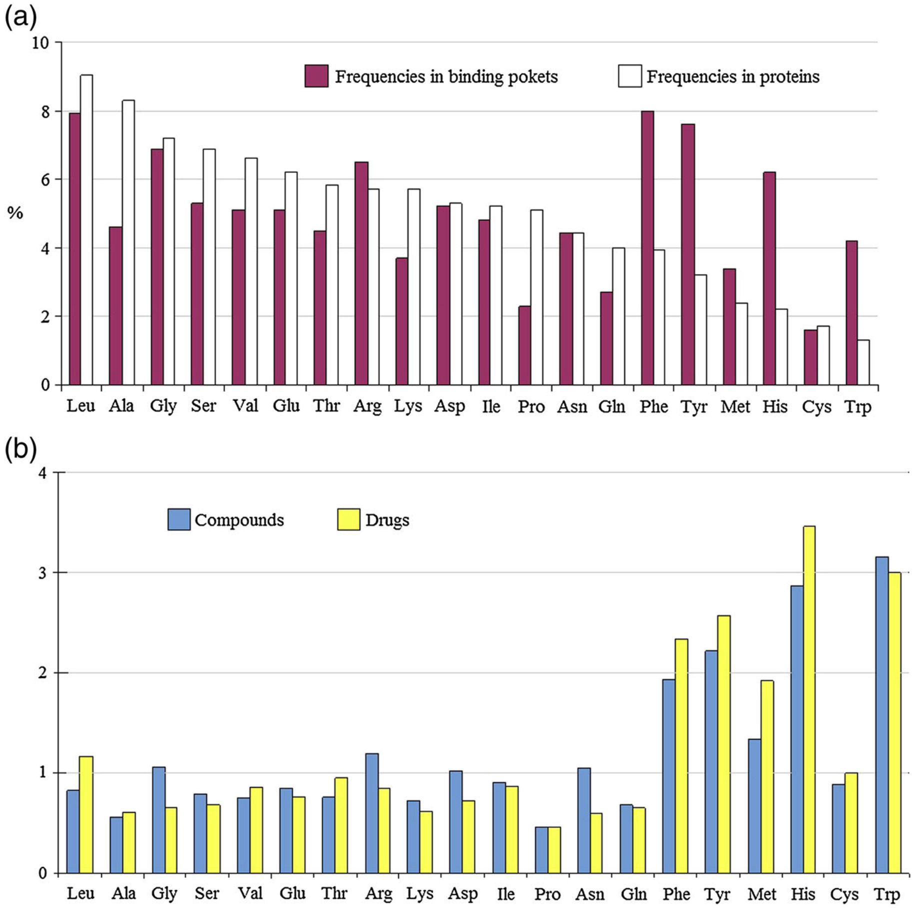

Figure 2. Frequencies of amino acid types in binding pockets. (a) The histogram shows the observed frequencies in binding pockets and the average amino acid composition of protein sequences. The amino acids are listed along the abscissa following the decreasing frequency of occurrence in proteins. (b) Propensity of occurrence in binding pockets, defined for each residue type as the ratio between the observed frequency in binding pocket and the average frequency in protein sequences.

spots (i.e., amino acids with relevant binding energy) on protein interfaces (Bogan and Thorn, 1998). Of the three amino acids displaying significant enrichment in hot spots, Trp, Tyr, and Arg, the first two are also observed with high frequency also in protein-small molecule interaction. However, significant differences in amino acid preferences between the results presented in this analysis and those of Bogan and Thorn indicate that protein-protein interaction are likely to obey to a more complex criteria than those found in the comparatively small interface area involved in protein-small molecule interactions.

Interaction types in protein-ligand complexes

The number of amino acids involved in protein-ligand complex formation depends on the conformation of both ligand and protein. In general, protein-ligand interactions include a significant number of van der Waals contacts, involving on average about six amino acids for drugs and eight for compounds (see Table 1). Ligand classified as both drugs and compounds displayed intermediate behavior (not shown). As the expected, the number of protein-ligand interactions depends also on the molecular size of the bound ligand, and both compounds and drugs display a wide variability in molecular weight, the number of intermolecular interactions was normalized to the molecular weight of the ligand (Table 1). Notably, although the number of van der Waals contacts per kDa formed by drugs (25.2 kDa^{-1}) and compounds (28.8 kDa^{-1}) are similar, the number of hydrogen bonds with either protein side or main chains is for compounds about twice as many as that observed for drugs; in detail, the average number of side chain hydrogen bonds is 12.1 kDa^{-1} and 6.4 kDa^{-1} for compounds and drugs, respectively, and that of main

Table 1. Average number of protein-ligand interactions

	Compounds	Drugs
Main chain hydrogen bonds	3.9 kDa^{-1} (0.9; 0.06)*	2.0 kDa^{-1} (0.6; 0.03)
Side chain hydrogen bonds	12.1 kDa^{-1} (2.6; 0.19)	6.4 kDa^{-1} (1.8; 0.10)
Van der Waals interactions	28.8 kDa^{-1} (6.0; 0.43)	25.2 kDa^{-1} (7.7; 0.37)

* Average number per kDa molecular weight of the ligand. The two numbers within parentheses show the average number of bonds per binding pocket and per ligand atom, respectively.

chain hydrogen bonds is 3.9 kDa^{-1} and 2.0 kDa^{-1}, respectively. The same analysis restricted to human endogenous compounds in complex with human proteins showed the same trends (not shown). As expected, ligands classified as both drugs and compounds displayed compounds behavior. Furthermore, for both drugs and compounds, about three-quarter of the hydrogen bonds with the ligand are carried out by side chains and only one-quarter by main chain atoms. Among the amino acids performing the main chain hydrogen bonding with the ligand, a large fraction (28% and 16% for compounds and drugs, respectively, see Table 2) involves Gly. The amino acids more frequently involved in side chain hydrogen bonds with drugs are His (21% of the total), and Tyr and Arg (13% each). Also for compounds, side chain hydrogen bonds carried out by Arg and His occur more frequently (17% and 12% of the total, respectively), together with Asp (12%). No significant variation with respect to compounds was observed in the subset of human endogenous compounds in complex with human proteins (Table 2).

Homologous proteins in complex with the same ligand

An increasingly large number of PDB entries describe complexes of a selected ligand with distinct homologous proteins. We analyzed the variability of binding pocket residues observed in homologous proteins in complex with the same ligand and compared this with the overall primary sequence variability with the aim of identifying general constraints affecting the variability of each type of binding pocket residue. This analysis was carried out on 299 distinct ligands having each at least two homologous PDB entries with sequence identity smaller than or equal to 95%. This limit in sequence identity was adopted to minimize statistical biases due to very similar, usually intensely studied, proteins. The pair wise comparison was carried out within each group of homologous proteins. Due to the paucity of statistics, for each residue type, the ratio between observed and expected variability was deduced as the average value calculated on seven similarity bins, each having 10% width and spanning the residue identity range 25% - 95% (see Methods Section). Subsequently, it was calculated for each residue type the ratio between the observed and the expected variability, the latter being the middle value of each similarity shell. This ratio, referred to as the Conservation Index, summarizes for each residue type the degree of conservation when part of a binding pockets with respect to the overall sequence conservation (Figure 3). The amino acids displaying highest Conservation Index were Gly (1.5), Arg (1.4), Glu (1.4), His (1.4), Asp (1.4), and Thr (1.4). Apart from Met (0.7), Ile (0.8), and Val (0.9), which tend to mutate more often than the average protein residues, no special conservation trends were observed for the remaining amino acids for both compounds and drugs.

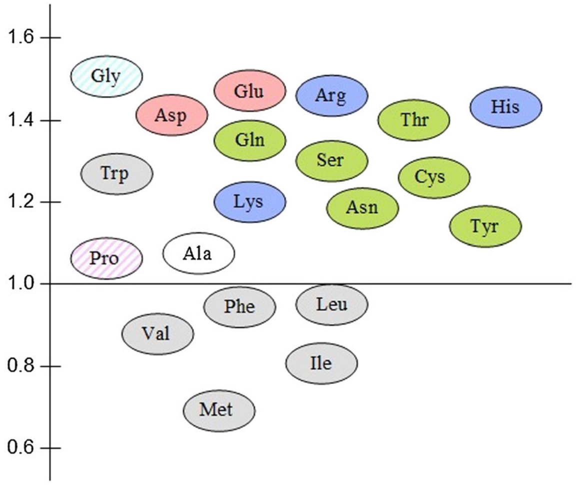

Figure 3. Conservation Index for ligands bound to homologous proteins. The Conservation Index is shown on the vertical axis. The horizontal axis has no dimension. To facilitate the readability of the figure the value 1 of the conservation index is shown by a horizontal line. Basic and acidic amino acids are displayed in blue and red background, respectively, and polar and hydrophobic residues in green and gray, respectively. The peculiarities of Pro and Gly are emphasized by hatched backgrounds.

Conservation of the binding pocket in non-homologous proteins

Of the 1118 ligands included in this study, 449 were found in complex with proteins belonging to distinct homology families. Of them, 318 were compounds, 20 drugs, and 111 had both KEGG classifications. In spite of the lack of structural equivalence among residues forming the binding pockets of unrelated proteins, it was still possible to search for maintained patterns by looking at the co-occurrence of the same amino acid type in the interaction with the conserved ligand. This analysis, carried out on a total number of 1567 protein complexes as described in the Methods Section, showed that amino acid types are more conserved than expected in the absence of constraints. The probability to find an amino acid in the binding pocket of one protein, knowing that the same amino acid is present in the binding pocket of a protein belonging to a different homology family, but in complex with the same ligand, was referred to as probability of co-occurrence. This probability is approximately 40% for Leu, Arg, Asp, Lys, Phe, Tyr, and His (Figure 4). In general,

Table 2. Frequency of hydrogen bonds in protein-ligand complexes

	Leu	Ala	Gly	Ser	Val	Glu	Thr	Arg	Lys	Asp	Ile	Pro	Asn	Gln	Phe	Tyr	Met	His	Cys	Trp
Main chain hydrogen bonds (%)																				
Drugs*	7	10	16	4	9	3	7	2	7	9	3	6	0	3	3	4	4	2	1	0
Compounds	7	10	28	6	6	3	3	2	3	3	4	3	4	2	5	2	3	2	2	2
Human endogenous compounds	5	10	30	5	8	2	5	2	3	2	4	2	3	3	7	2	5	0	2	0
Side chain hydrogen bonds (%)																				
Drugs	—	—	—	10	—	8	7	13	7	8	—	—	6	4	—	13	—	21	2	1
Compounds	—	—	—	10	—	9	7	17	9	12	—	—	9	5	—	8	—	12	0	2
Human endogenous compounds	—	—	—	11	—	10	7	20	11	12	—	—	6	6	—	9	—	7	0	1

*On each line, the values add up to 100.

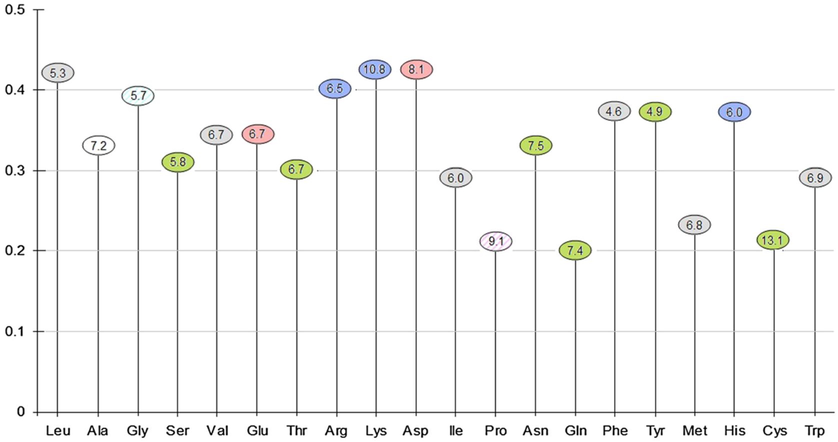

Figure 4. Probability of co-occurrence in the binding pocket of proteins belonging to distinct homology families in complex with the same ligand. For each residue type, the ratio between the observed and the expected frequency is also shown. Residues are color-coded as in Figure 3.

amino acids display conservation probabilities ranging between 42% and 20%, values which are significantly higher than those expected on a random basis, as shown by the ratio between observed and expected frequencies (Figure 4). The latter were calculated using the average amino acid composition of all binding pockets included in this study. The ratio between observed and expected frequencies can be as large as 13.1 and 10.8 for Cys and Lys, respectively (Figure 4). No significant difference was observed between drugs and compounds. We measured also the degree of conservation of the basic, acidic, or aromatic character of binding pocket residues. The associated indices measure the probability of the concomitant presence, in the binding pockets of two unrelated target proteins, of amino acids with conserved chemico-physical character. For aromatic residues (Phe, Tyr, Trp), the probability of co-occurrence is 54% (Table 3), for acidic or basic residues, the respective values of 54% and 59% were found. Also in this case, no distinct behavior was observed between drugs and compounds.

DISCUSSION

In this work, the interactions between proteins and small molecules were analyzed by scrutinizing the protein-ligand complexes available in the PDB. The rather large number of protein-ligand complexes found in the PDB allowed to analyze the average composition of the binding pocket and to identify differences between artificial drugs and physiological compounds. Former studies (Zvelebil and Sternberg, 1988; Bartlett *et al*., 2002) were focused on the residues directly involved in the catalytic process in enzymes. Here, a more general approach was adopted, and all amino acids in interaction with the bound substrate were considered. In order to identify differences between artificial drugs and physiological compounds, the 1118 distinct ligands included in this study were sorted, according to the KEGG classification, in drugs, compounds, drugs, and compounds. Drugs are molecules which have been approved in either USA, Europe, or Japan, they include nutriaceuticals and OTC drugs. Compounds are defined in KEGG as molecules with biological role and are the product of biological processes. Ligands classified uniquely as drugs are for the most part synthetic compounds, whereas examples of ligands which are classified as both compounds and drugs include antibiotics, antifungal agents, and vitamins.

Table 3. Probability of co-occurrence in non-homologous protein

Aromatics	54%
Basics	59%
Acidics	54%

The present analysis showed significant biases in the composition of the binding pocket with respect to the average frequency observed in protein sequences. Five of the six rarer amino acids in protein sequences, Phe, Tyr, Met, His, and Trp, are instead the most frequently observed in binding pockets of both drugs and compounds. Similar results have been obtained, on a smaller test set, by Soga *et al*. (2007). Together, the five amino acid types account for the 26% of the residues forming the binding pocket, whereas on a statistical basis, they sum up to a mere 10.6% of the total number of amino acid in protein sequences. The same analysis restricted to enzymes showed similar trends (data not shown). Among the five residues mentioned above, His is known to be frequently involved in the catalytic process in enzymes (Bartlett *et al*., 2002). Notably, the three amino acids scoring just below His in the frequency distribution, Phe, Tyr, and Trp are aromatic. The enhanced presence in the binding pocket of aromatic and hydrophobic residues suggests that the van der Waals interactions and primarily the aromatic staking interactions are relevant in establishing the non-covalent but specific binding of the ligand to the target protein(s) (Tewari and Durbey, 2008; Pyrkov *et al*., 2009). In particular, the directionality of the aromatic stacking interactions is likely to play a relevant role in substrate specificity.

Although drugs and compounds do not display large variation in the number of van der Waals interactions with the bound protein, significant differences were observed in the number of hydrogen bonds, which are on average twice as many for compounds than for drugs. This trend was found for both side chain and main chain hydrogen bonds. The same patterns were observed when the analysis was restricted to endogenous human compounds in complex with human proteins. This might suggests that physiological compounds have optimized substrate selectivity by establishing, under the pressure of natural selection, an adequate number of hydrogen bonds. By comparison, the smaller number of hydrogen bonds carried out by synthetic drugs may be linked to the technological limits of the process of drug design and optimization, and in some cases, it could be associated to the reduced selectivity of lead compounds for the primary target with respect to secondary targets. The comparison with natural compounds indicates that there should still be leeway for improvement in drug specificity by increasing, during the process of lead optimization, the number of hydrogen bonds established with the desired target protein. For both compounds and drugs, hydrogen bonds involve mostly side chains (three-quarters of the instances). Among the main chain bonds, about one-third involve Gly, likely due to the lack of the steric hindrance of the side chain. Among the residues carrying out side chain hydrogen bonding with either drugs or compounds, His and Arg, respectively, are more frequently observed. As pointed out by one of the reviewers, the distinct behavior of these two residues in complexes with drugs and compounds is rather unexpected. One interpretation of the large number of His involved in side chain hydrogen bonding with drugs could be that these are especially optimized to bind the catalytic site, which is often a His residue. The involvement of Arg side chain in the catalytic process may be the reason of the increased frequency of side chain hydrogen bonds carried out by the Arg guanidinium group in protein-compound complexes. No significant biases were found among the other polar amino acids involved in side chain hydrogen bonds with the ligand. These trends were maintained unchanged if the analysis was restricted to mammals, to enzymes, or to human endogenous compounds (data not shown).

The PDB contains many instances of ligands bound to homologous proteins. In these cases, structurally equivalent amino acids forming the binding pockets could be compared, and the conservation of binding pocket residues could be estimated relatively to the overall sequence similarity. The Conservation Index provides a coarse estimation of the conservation of binding pocket amino acids with respect to the average conservation of the overall primary sequence. The maximum value attainable by the Conservation Index is associated to the full conservation of binding pocket residues, and depends on the similarity shell. The maximum value varies between 3.33, for the similarity bin 0.25-0.35, and 1.11 for the shell 0.85-0.95. Assuming an identical number of instances in each similarity shell, the maximum value attainable by the Conservation Index, which is the average value of the seven similarity shells, would be 1.9. The amino acid types displaying higher degree of conservation with respect to the rest of the protein are Gly (1.51), Arg (1.45), Glu (1.44), His (1.43), Asp (1.41), and Thr (1.40). These values, compared to the maximum of 1.9, display substantial conservation with respect to the overall sequence similarity. From the structural viewpoint, Gly is likely to be conserved because of the reduced steric hindrance that makes it difficult for its replacement with a larger residue at the interface with the ligand, especially in the (frequent) case in which Gly is hydrogen bonded with the ligand. The conservation of the charged residues His, Arg, Glu, and Asp (but not Lys) is likely to be due to the involvement in side chain hydrogen bonding with the ligand, possibly enhanced by a Coulomb component. The Conservation Index may provide the basis for the definition of a scoring function aimed at identifying novel binding sites in proteins homologous to the primary target of molecules of pharmacological interest (numerical values are given in the Table S1).

The PDB contains also numerous cases of ligands in complex with proteins belonging to distinct homology families. In spite of the lack of structural equivalence among residues involved in the binding with the conserved ligand, a systematic comparison of the amino acid composition of the distinct binding pockets could still be carried out. The analysis showed that different proteins display a marked conservation of the amino acid composition of the binding pockets of the same ligand. For Cys and Lys, the conservation could be as high as thirteen (Cys) or eleven (Lys) times higher, respectively, than that expected in the absence of any bias. It should be noted that the high probability of co-occurrence of Lys in the binding pocket of protein belonging to distinct homology families is not in contrast with the moderate conservation observed for the same amino acid among homologous proteins, in fact in the first case Lys may occupy distinct positions in the binding pocket, whereas the statistics for homologous proteins refer to the conservation occurring at the same structural position. Also, the presence of an aromatic, basic, or acidic side chain is maintained with more than 50% probability. These observations could be useful in the field of drug repurposing. In this case, the amino acids of the binding pocket form a structural signature that might be used, together with an appropriate scoring function deduced from the analysis discussed above, to scan for putative binding site in new potential, and unrelated protein targets.

REFERENCES

Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. 2002. Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324**: 105–121.

Benson ML, Smith RD, Khazanov NA, Dimcheff B, Beaver J, Dresslar P, Nerothin J, Carlson HA. 2007. Binding MOAD, a high-quality protein-ligand database. *Nucleic Acid Res.* **36**: D674–D678.

Bogan AA, Thorn KS. 1998. Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**: 1–9.

Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. 2008. Drug target identification using side-effect similarity. *Science* **321**: 263–266.

Creighton TE. 1993. Proteins. Structures and molecular properties. 2nd Edn. W.H. Freeman and Company: New York.

Dakshanamurthy S, Issa NT, Assefnia S, Seshasayee A, Peters OJ, Madhavan S, Uren A, Brown ML, Byers SW. 2012. Predicting new indications for approved drugs using a proteochemometric method. *J. Med. Chem.* **9**: 6832–6848.

Dutta S, Burkhardt K, Young J, Swaminathan GJ, Matsuura T, Hendrick K, Nakamura H, Berman HM. 2009. Data deposition and annotation at the worldwide Protein Data Bank. *Mol. Biotechnol.* **42**: 1–13.

Ekins S, Williams AJ, Krasowski MD, Freundlich JS. 2011. *In silico* repositioning of approved drugs for rare and neglected diseases. *Drug Discov. Today* **16**: 298–310.

Fersht A. 1993. Enzyme structure and mechanism. 2nd Edn. W.H. Freeman and Company: New York; 452–459.

Fuller JC, Burgoyne NJ, Jackson RM. 2009. Predicting druggable binding sites at the protein-protein interface. *Drug Discov. Today* **14**: 155–161.

Gallina AM, Bisignano P, Bergamino M, Bordo D. 2013. PLI: a web-based tool for the comparison of protein-ligand interactions observed on PDB structures. *Bioinformatics* **29**: 395–397.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**: D109–D114.

Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS. 2011. DrugBank 3.0: a comprehensive resource for “omics” research on drugs. *Nucleic Acids Res.* **39**: D1035–D1041.

Laskowski RA, Chistyakov VV, Thornton JM. 2005. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res.* **33**: D266–D268.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.* **85**: 2444–2448.

Pyrkov TV, Pyrkova DV, Balitskava ED, Efremov RG. 2009. The role of stacking interactions in complexes of proteins with adenine and guanine fragments of ligands. *Acta Naturae.* **1**: 124–127.

Reddy AS, Amarnath HS, Bapi RS, Sastry GM, Sastry GN. 2008. Protein ligand interaction database (PLID). *Comput. Biol. Chem.* **32**: 387–390.

Sardana D, Zhu C, Zhang M, Gudivada RC, Yang L, Jegga AG. 2011. Drug repositioning for orphan diseases. *Brief. Bioinform.* **12**: 346–356.

Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R, Yeats C, Thornton JM, Orengo CA. 2013. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.* **41**: D490–D498.

Singh J, Petter RC, Baillie TA, Whitty A. 2011. The resurgence of covalent drugs. *Nat. Rev. Drug Discov.* **10**: 307–317.

Soga S, Shirai H, Kobori M, Hirayama N. 2007. Use of amino acid composition to predict ligand-binding sites. *J. Chem. Inf. Model.* **47**: 400–406.

Tewari AK, Durbey R. 2008. Emerging trends in molecular recognition: utility of weak aromatic interactions. *Bioorg. Med. Chem.* **16**: 126–143.

Wallace AC, Laskowski RA, Thornton JM. 1995. LIGPLOT: a program to generate schematic. *Protein Eng.* **8**: 127–134.

Wang R, Fang X, Lu Y, Yang C-Y, Wang S. 2005. The PDBbind database: methodologies and updates. *J. Med. Chem.* **48**: 4111–4119.

Zvelebil M, Sternberg M. 1988. Analysis and prediction of the location of catalytic residues in enzymes. *Prot. Eng.* **2**: 127–138.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.