[1]On the Asymmetry of the Residue Compositions of the Binding Sites on Protein Surfaces

Gábor Iván, Zoltán Szabadka, and Vince Grolmusz

Protein Information Technology Group, Eötvös University; Pázmány P. stny. 1/C, H-1117 Budapest, Hungary

and Uratim Ltd. Sóstói út 31/b, Nyíregyháza, Hungary

ABSTRACT By screening all the ligand binding sites in the Protein Data Bank, we have found that while it is geometrically possible that a loop, formed from a protein-chain with residues ZYX would "impersonate" another chain-loop with residues XYZ by a simple twisting of either the loop or the bound ligand, it almost never happens. This fact is rather surprising, and implies a notable asymmetry, since (i) loops in the folded proteins sometimes *can be* flexible enough to be twisted, but (ii) ligands are almost *always* extremely mobile before binding to the protein, therefore they can turn around and bind to residue-sequence ZYX as well.

Data availability: The on-line supporting Table 3 lists the appearances of the residue-sequences and their inverses in the binding sites of the whole PDB.

Introduction: The fast growing Protein Data Bank[1] contains the three-dimensional structures of more than 50,000 entries today. Its open access and easy availability make possible refined structural studies for the research community. In a recently established database, called RS-PDB (Rich-Structure PDB) database[2], starting from the mmCIF format of the data, we cleaned numerous inconsistencies and re-built the database in a strictly logical and an easily searchable way. The main result of our work was the reliable identification and description of the protein-ligand complexes found in the PDB. Note, that by "ligand" we mean a non-covalently-bound, non-crystallization-artifact, InChI-identifed entity[2].

In the June 1, 2006 version of the PDB we identified more than 25,000 protein-ligand pairs. However, these pairs may contain numerous redundancies: the same polypeptide sequence may be present even in more than 160 PDB entries[2]. We filtered out the redundancies from the dataset by considering every (polypeptide sequence, ligand) pair only once[2]. After filtering out redundancies in this sense, we were left with 19,581 distinct binding sites on protein surfaces. For each binding site those residues were collected that were closer to any ligand atoms than 1.05 times the sum of the Van der Waals radii of the two atoms involved. Next we identified the residues containing these atoms: for every binding site an ordered sequence of the residues were created. Then we identified these residues in the amino-acid sequence of the chain of the protein, residues not present in the given binding site were simply substituted by a "-" mark.

For every ligand-binding site we have built this way one or more sequence like the following one in PDB entry 2BZ6 (for brevity we have cut off the - - - segments from the start and the end of the sequence). Note, that the sequence is listed from N-terminal at the left to the C-terminal on the right.

H -
- - - - - - - TT - - D -
- -
- - - - - - - - - - P - - - - - - - - - - - - - - - - DSCK - - S - - - - -
- - - - - - - - - - - - VSWGQGC - - - - - - G

More than one such sequence was generated for a given binding site if more than one chain took part in forming the site.

One should note that due to the secondary structure of the protein, rough periodicity can be observed by such description of the binding sites. Note also, that this representation is a hybrid of a sequential and a spatial representation of ligand binding sites on proteins.

Results: We gathered all the contiguous, maximal residue-subsequences of length at least 3 in the sequences built for all the ligand binding sites. For example, we collected DSCK and VSWGQGC from the sequence above. Pairs were not considered, since our intentions were to have clear-cut, emphasized evidences. Next the palindromes, if present, were deleted.

As the next step, we counted the number of appearances of these subsequences and also their inversed ones in binding sites. For example, we counted the frequencies of both XYZ and ZYX subsequences, and compared the results.

Clearly, from the 20 residues, 8000 residue-triplets are possible (we filtered out the rarely appearing[3] modified amino acids, too). We have found that from these 8000, only 1125 triplets are present in binding sites, and from these triplets only 137 have their inverse also present in binding sites. It is striking that only 8 such pairs bind the same ligands. (see Table 1 for the list of these 8 triplets).

Residue triplet	Ligand(s) binding to the triplet and also to its inverse	PDB code straight	PDB code inverse
FGI	SPIRILLOXANTHIN	IEYS	1EYS
GMS	FLAVIN-ADENINE DINUCLEOTIDE	1B37 1B5Q	1EL5
SGG	URIDINE-5\'-DIPHOSPHATE	1F6D	2B6U 1C3J
GAV	NICOTINAMIDE-ADENINE-DINUCLEOTIDE	1EZ4 1F8F	2A65
VDL	NICOTINAMIDE-ADENINE-DINUCLEOTIDE	1CYD	1EQU
NGL	NADPH	2DB V	2BL9
GAA	FLAVIN-ADENINE DINUCLEOTIDE	2C3D	1BHY 2BRA 2BRY
GGI	1-[2-HYDROXY-4-(2-HYDROXY-5-METHYL-CYCLOPENTYLCARBAMOYL)5-PHENYL-PENTYL]-4-(3-PYRIDIN-3-YL-PROPIONYL)-PIPERAZINE-2-CARBOXYLIC ACID TERT-BUTYLAMIDE	2BPV 2BPW	2BPW 2BPV
	N-[2(S)-CYCLOPENTYL-1(R)-HYDROXY-3(R)METHYL]-5-[(2(S)-TERTIARY-BUTYLAMINO-CARBONYL)-4-(N1-(2)-(N-METHYLPIPERAZINYL)-3-CHLORO-PYRAZINYL-5-CARBONYL)-PIPERAZINO]-4(S)-HYDROXY-2(R)-PHENYLMETHYL-PENTANAMIDE	2BPY 2BPZ	2BPZ 2BPY
	ARGININE + VALINE + 2-AMINO-4-METHYL-PENTAN-1-OL + PHENYLALANINE + GLUTAMIC ACID + ALANINE + NORLEUCINE AMIDE	1EBK	1BW6 1EBK

TABLE 1: The listing of those 8 residue-triplets, whose inverses bind the same ligands.

By analyzing the data of Table 1, we have found the following phenomena:

In the case of residue-triplet FGI, the bound ligand is a long-chain spirilloxanthin, bound to the M chain's PHE 176 – GLY 177 – ILE 178 (giving the FGI pattern) and also to the ILE 70 – GLY 71 – PHE 72 (yielding the IGF pattern).

In the case of residue-triplet IGG-GGI, most of the appearances are in the HIV-1 protease inhibitor structures (2BPV, 2BPW, 2BPY, 2BPZ, 1EBK); and there, for example, on the B chain, the IGGIGG binding profile can be found in residue-positions 47,48,49,50,51 and 52, respectively. The GGIGG is a length-5 palindrome, consequently, it will not appear in our dataset of length-5 sequences, but its parts (GGI and IGG) appear there.

From the 160,000 possible residue-quadruplets, 675 are present in binding sites, 10 of them appears also in their inverse form, but no such ligand exists what is present

bound to any subsequences and also to its inverse of length 4 or greater.
Table 2 lists those residue-quadruplets, whose inverses are also present in binding sites.
The on-line supporting Table 3 lists the appearances of the residue-sequences and their inverses in the binding sites of the whole PDB.

Methods: We presented a new pre-processing algorithm that works on the mmCIF (macromolecular Crystallographic Information File) format of the PDB[1]. Our method uses the InChI chemical identifier of the IUPAC[4,5] for the unique identification of ligand molecules. Our method checks the entries for errors and inconsistencies, marks missing atoms, decomposes the structures into protein-, nucleic acid- and polysaccharide chains as well as various types of ligand molecules (e.g. peptides, cofactors/coenzymes, metals, etc.), distinguishes between covalently and non-covalently bound ligands, and identifies different binding sites. The result is a strictly structured, homogeneous database, adequate for processing diverse queries and serving intricate data-mining applications.

The first part is the small molecule database, where molecules in a format similar to that used in SD files are stored. We use the canonical numbering generated together with the IUPAC/NIST chemical identifier of the molecule (InChI)[5]. To generate this, our program uses the API that was obtained from IUPAC's web-site[4]. If a new molecule is inserted into this part of the database, then first its InChI is generated, and its presence is checked. The key of this part of the database is a sequence number called molid that is incremented each time a new InChI is found.

The second part of the database contains the PDB Chemical Component Dictionary. Its mmCIF format, components.cif can be downloaded from RCSB's web-site[6]. It consists of the connection tables of the molecules and functional groups associated with the three letter monomer codes found in the PDB. Here we found inconsistencies (e.g., molecules, marked as "aromatic", without any rings). We corrected these errors, and before we inserted a monomer into the database, the standard "valence rules" were verified as described in the InChI Technical Manual[5].

Residue quadruplet	Ligand(s) binding to the quadruplet
ISWG	GLUTAMIC ACID + GLYCINE + ARGININE
GWSI	2-AMINO-5,6-DIMERCAPTO-7-METHYL-3,7,8A,9-TETRAHYDRO-8-OXA-1,3,9,10-TETRAAZA-ANTHRACEN-4-ONE GUANOSINE DINUCLEOTIDE
ALGG	HISTIDYL-ADENOSINE MONOPHOSPHATE
GGLA	7-DEAZA-7-AMINOMETHYL-GUANINE
LHGG	N-TOSYL-L-LYSINYL METHYL KETONE
GGHL	5-FORMYL-6-HYDROFOLIC ACID and 5-HYDROXYMETHYLENE-6-HYDROFOLIC ACID
SGGI	NICOTINAMIDE-ADENINE-DINUCLEOTIDE
IGGS	6-PHOSPHOGLUCONIC ACID and 5-PHOSPHOARABINONIC ACID
SGDL	NADP NICOTINAMIDE-ADENINE-DINUCLEOTIDE
LDGS	2,5-ANHYDROMANNITOL-1,6-DIPHOSPHATE
GEGS	ADENOSINE-5'-TRIPHOSPHATE
SGEG	URANYL (VI) ION + URANYL (VI) ION + URANYL (VI) ION + ADENOSINE MONOPHOSPHATE + WATER
KAGV	FLAVIN-ADENINE DINUCLEOTIDE
VGAK	NICOTINAMIDE-ADENINE-DINUCLEOTIDE
MLGG	GUANOSINE-5'-MONOPHOSPHATE and INOSINIC ACID
GGLM	N-[3-CARBOXY-2-HYDROXY-PROPIONYL]-L-HOMOPHENYLALANYL-AMINO-2-METHYLBUTANE
VLGG	PHOSPHOAMINOPHOSPHONIC ACID-ADENYLATE ESTER + MANGANESE (II) ION + MANGANESE (II) ION + WATER + WATER
GGLV	INOSITOL 1,3,4,5-TETRAKISPHOSPHATE
GGAL	SERINE + GLUTAMIC ACID + HISTIDINE + PHENYLALANINE + ASPARAGINE + GLUTAMIC ACID + TYROSINE and VALINE + VALINE + SERINE + HISTIDINE + PHENYLALANINE + ASPARAGINE + ASPARTIC ACID
LAGG	THYMIDINE and THYMIDINE-5'-TRIPHOSPHATE and 2'-DEOXY-THYMIDINE-BETA-L-RHAMNOSE

Table 2: Listing of quadruplets of amino acids appearing with their inverses in binding sites. Note, that no ligand binds to the quadruplet and to its inverse.

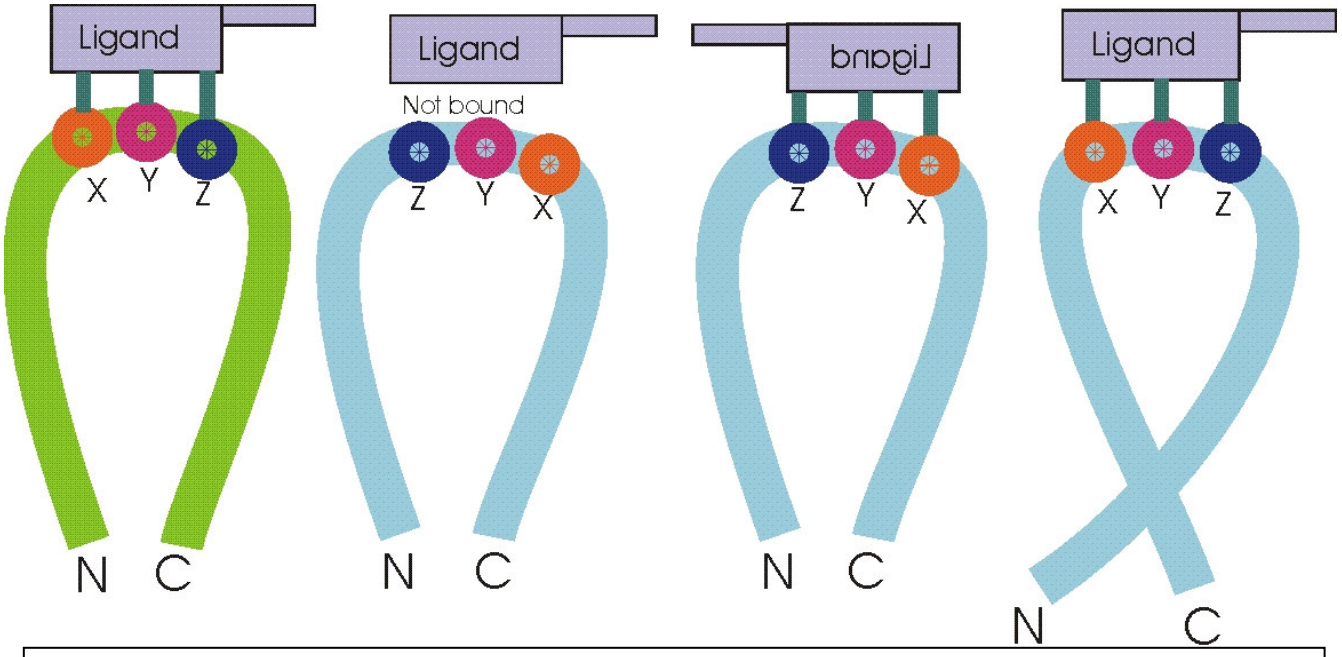

Figure 1. The loop on the left contains residue-sequence XYZ with a bound ligand. The second one contains the sequence ZYX. Most probably the same ligand will not bind to this, unless either the ligand (as on the third loop) or the loop itself (the fourth loop) is twisted. We have found that the latter two cases almost never happen.

The most important part of the database is the third one. After an entry is successfully read by our program, it will consist of polymer and ligand molecules, and each one will have a unique number within the entry. This number is denoted by mol# in the database tables. For each ligand molecule, we create its InChI identifier and insert it into the small molecule database as well. In an mmCIF file, the contents of the asymmetric unit are listed in the table struct_asym. Each item (also called entity) in this list has an asym id. The type of an entity can be polymer, non-polymer or water. Each polymer entity has also a polymer type.

We define a protein chain as a polymer entity of type "polypeptide(L)", if it is at least ten monomers long, and a DNA/RNA chain as a polymer entity, which is at least 5 monomers-long and its type is either "polydeoxiribonucleotide", "polyribonucleotide", or more than half of its monomers are nucleic acids (A,C,G,I,T,U monomer id).

At this point the list of monomers that make up a polymer chain is identified. The covalent structure of these monomers (the so-called "connection table") is read from the PDB Chemical Component Dictionary[6] (formerly HET Group Dictionary, HGD). The next step, i.e., connecting the monomers to obtain the covalent structure of the whole chain is performed by adding the monomers to the chain one-by-one.

After creating the connection table for the polymer chains, we read the list of monomers from the polymer entities that were not long enough (these will form the oligopeptide ligands, for example). Their connection table is obtained from the HGD[6].

At this point, we still have the initial set of ligands. A molecule in the final set of ligands consists of two or more such monomers, bound covalently. To identify such covalent bonds, we select all pairs of atoms in the entry that are closer to each other than 6 Å. We achieve this by building a Kd-tree[7] on the atoms, thus avoiding the examination of all pairs, and saving a lot of computational time.

Finally, the coordinates of the hydrogen atoms are computed using standard hybridization rules. We also calculate partial charges for the ligand molecules using the Gasteiger algorithm, and assign the Kollman partial charges for the atoms in protein chains.

At this point we identified all the ligands in the PDB.

The list of binding sites are created as follows: every binding site is represented as a set of atom-pairs that are "close" to each other: the distance of the two atoms is at most 1.05-times the sum of their respective Van der Waals radii, but the distance should be larger than 1.25 times the sum of their respective covalent radii, since we do not consider covalently bound ligands. These pairs of atoms - one belonging to a ligand the other to the protein chain - are considered to be bond to each other. We have found approximately 1.9 million such pairs in the PDB.

The atom-pairs that came from the same PDB entry, and, moreover, their ligand-atoms are the part of the same ligand molecule, belong to the same binding site. There are 25,552 unique binding sites in our database.

Next, we need to handle the numerous multiplicities, occurring in the PDB: we intend to take into account every (ligand, protein) pair only once. After this step we were left with 19,581 distinct protein-ligand pairs.

Next we identified the residues containing the protein-atoms in the binding sites: for every binding site an ordered sequence of the residues were created. Then we identified these residues in the amino-acid sequence of the chain of the protein, residues not present in the given binding site were simply substituted by a “-" mark.

For every ligand-binding site we have built this way one or more sequence like the following one in PDB entry 2BZ6 (for brevity we have cut off the - - - segments from the start and the end of the sequence). Note, that the sequence is listed from N-terminal at the left to the C-terminal on the right.

H -
- - - - - TT - - D -
- -
- - - - - - - - - P - - - - - - - - - - - - - - - - DSCK - - S - - - - - -
- - - - - - - - - - - VSWGQGC - - - - - - G

More than one such sequence was generated for a given binding site if more than one chain took part in forming the site.

Conclusions: All the ligand binding sites of the PDB were scanned for residue-patterns in the polypeptide sequence.
It was concluded that in binding sites on protein surfaces, it is very unlikely that loops, formed from protein chains, with - say - residue-subsequence ZYX, by twisting either the loop or the ligand, take part in binding the same ligands that another loop with residue-subsequence XYZ, as it is visualized on Figure 1.

We find this result highly surprising, since even if we accept that polypeptide loops are extremely rarely twisted in protein structures (as the fourth loop on Figure 1), due to evolutionary determination of the tertiary structure, the very mobile and small ligand molecules could assume any transitional position relative to the protein before binding, therefore the conformation depicted on the third loop needs to be much more frequent than observed.

The authors are not aware of any similar observation on a complete data-set as ours.

Acknowledgements:

The authors acknowledge the partial support of an OTKA grant N 67867 and NKTH Jedlik grant TB-INTER.

References:

(1) Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P., Nucleic Acids Research 28, 235--242 (2000).

(2) Szabadka, Z., Grolmusz, V.: Proceedings of the 28th IEEE EMBS Annual International Conference, New York City, Aug. 30-Sept 3, 2006., pp. 5755-5758.

(3) Szabadka, Z., Grolmusz, V.: High throughput processing of the structural information in the Protein Data Bank, Journal of Molecular Graphics and Modeling, 25, (2007) pp. 831-836

(4) http://www.iupac.org/inchi/index.html

(5) Adam, D., Chemists synthesize a single naming system, Nature 417, (2002), No. 369

(6) The PDB Chemical Component Dictionary (formerly the Het Group Dictionary) is available at http://deposit.pdb.org/public-component-erf.cif

(7) Bentley, J. L., Multidimensional Binary Search Trees Used for Associative Searching, Communications of the ACM, 18, (1975) No. 18, pp. 509-517

(8) A publicly available PDB-file ligand decomposer is on-line at http://decomp.pitgroup.org