Cysteine and Tryptophan Anomalies Found when Scanning all the Binding Sites in the Protein Data Bank

February 11, 2010

Abstract

The fast-growing Protein Data Bank contains the description of more than 63,000 structures today, being one of the richest source of structural biological information on the Earth. Started to exist as the computer-readable depository of crystallographic data complementing printed articles, the proper interpretation of the content of the individual files in the PDB still frequently needs the detailed information found in the citing publication. This fact implies that the fully automatic processing of the whole PDB is a very hard task.

Here we show a mathematical and graph theoretical method for automatically repairing, re-organizing and re-structuring PDB data. The most important result of this cleaning procedure is the reliable and automatic identification of all the protein-ligand complexes and binding sites in the PDB. The identification of all binding sites on the surface of all proteins of known 3D structure opens the door for large-scale studies for the characterization of binding sites. Since protein-ligand binding is of special importance in several areas (e.g., enzymology or drug design), in silico analysis of several tens of thousands of binding sites would yield remarkable results. In the residue composition of the binding sites we identified strong cysteine and tryptophan irregularities in the data.

1 Introduction

The increasing size and accuracy of structural information stored in the Protein Data Bank [1] make possible large-scale, fully automated *in silico* studies involving thousands of protein-ligand complexes and binding sites. The most important implication of such studies were the structural classification of binding sites on protein-surfaces, applicable for the prediction and modeling of protein-ligand interactions.

In the present work we structurally analyze and re-build the PDB, identify protein-ligand complexes and binding sites, and we apply datamining techniques for the sets, formed from the residues at each binding sites present in the whole Protein Data Bank.

Directly or indirectly we make use non-trivial mathematical, mainly graph-algorithms: Computing the InChI$^{\mathrm{TM}}$ code [2, 3] applies a graph-isomorphism testing, transforming aromatic notation to Kekule-notation uses a non-bipartite graph-matching algorithm [4], breadth-first-search graph traversals [5] are used throughout this work, depth-first search [5] is used in building the ligand molecules and identifying ring structures, kd-trees [6] are applied for computing covalent bonds, and hashing [5] are utilized for the fast generation of protein-sequence ID's.

Identification of Protein-Ligand Complexes

It is a highly non-trivial problem to automatically identify protein-ligand complexes in the Protein Data Bank [1]. The HET label of atoms in the PDB files may denote metals, atoms of modified residues, even atoms in small molecules added in crystallization and also co-valently bound ions. Consequently, the HET atoms alone will not identify ligands. Small pieces of broken peptide-chains – erroneously – may also be seen to be ligands. Obviously, by careful human examination of the remark fields of the individual PDB entries, together with the thoughtful study of journal publications where the solution of the protein-structure was first reported would solve these problems, but they are definitely inadequate for *automatic processing* of the whole PDB, even by the most powerful textual data-mining techniques.

The PDBsum pictorial data base [7] contains reliable structural information on ligands and binding sites: hand-examination of single entries is comfortable, automatic examination of large sets of graphical data is impossible. The sc-PDB database [8] was made by automatic processing of the whole PDB, using, among others, textual information in the remark and title fields of the entries in deciding if a structure is a complex or not. However, if a protein-ligand complex is not marked with the words "complex" or "ligand" in some remark field, then their method will not find it, as it was remarked [9]. In the PDBbind Database [10], by manual, human-involved search, binding affinities were compiled from hundreds of protein-ligand complexes from the PDB.

It is a highly non-trivial task to assign proper structure for multi-meric small molecules in the PDB. As it was observed in [11], the otherwise high quality RELIBASE [12] has some shortcomings in this area. The creators of the SMID [11] small molecules depository in PubChem (http://pubchem.ncbi.nlm.nih.gov) used the mmCIF information from the PDB for bond descriptions for these small molecules, while the source of the protein data was obtained from the MMDB [13, 14, 15, 16, 17, 18] database.

We choose a more reliable, fully automatic mathematical method for identifying complexes. The reliability comes from the fact that we present a new pre-processing algorithm that works enterily on the mmCIF (macromolecular Crystallographic Information File) format of the PDB, and uses the International Chemical Identifier (InChI$^{\mathrm{TM}}$) of the International Union of Pure and Applied Chemistry (IUPAC) [2, 3]. We found this resource mathematically much more reliable than the *ad hoc* naming conventions in Section 2.2 in [11]. 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. In the process we are using as little as possible the labels and remarks in the file in the PDB. The result is a strictly structured, homogeneous database, called the RS-PDB database, adequate for processing diverse queries and serving intricate data-mining applications. We applied in our database a modification of the definition of the ligands used in the PDBbind Database [9].

- The input of the algorithm is the mmCIF file of the PDB entry and the PDB Chemical Component Dictionary which contains the chemical structure of each monomer in the PDB.
- The output of the algorithm is the RS-PDB database (the abbreviation stands for *Rich Structure PDB*).

We explain step-by-step how an entry from the PDB, given in the mmCIF format, is processed, checked for errors and is finally decomposed into polymer chains and ligand molecules in Section 1.

Results and Discussion

After the rigorous ligand identification and redundancy-deleting procedure, we gained 19,581 different binding sites, and analyzed the residues found at the binding site.

We, for each ligand L identified in the RS-PDB database, a description of the residues in the binding site was generated by the following method: we went through the ligand atoms one-by-one and found those protein atoms which were closer to them than 1.05 times the sum of the Van der Waals radii of the two atoms scanned. (See Figure 1 for an example). Note, that covalently bound ligands are already filtered out at this point, so all binding is non-covalent. After identifying the atoms in the protein, we identify the residues containing these atoms: for every binding site a subset of the 20 amino acids were created. If the same residue appeared more than once, we inserted it only once into the residue-set. We have made this choice since we were mostly interested in the joint patterns of appearance of the residues in binding sites:

handling the multiplicities of the residue-appearances in the collection phase is not difficult, but results were hard to analyze.

Our goal is to analyze the properties of these ligand-binding residue-sets by identifying the frequency of the residue-sets.

Residue-set frequency: the cysteine and the tryptophan anomalies

We counted the frequencies of all the subsets of the 20 different residues, which appeared in at least 450 different binding sites. Note that small subsets, containing one or two residues, appear each at least 450 binding sites; we have cut at 450 since we intended to analyze residue-subsets with large enough frequency. The individual amino acids (i.e., the 1-element subsets) in the binding sites appeared with the frequencies, given in Table 1.

The analysis of our results are based on the individual frequencies in Table 1. A central quantity of interest is the *lift* [19] of subsets of the 20 amino acids. If the residue-composition of the binding sites were distributed randomly and independently with the probabilities equal to the frequencies in Table 1, then, for example, the probability of the two element subset (CYS, SER) were $0.2094 \times 0.5227 = 0.1095$.

In the database, however, the (CYS, SER) pair appears with much higher frequency: 0.1425. The lift is the quotient of the real, counted frequency and the computed frequency. In our example, the lift of the (CYS, SER) pair is 1.3014.

It is a remarkable observation that a relatively infrequent amino acid, the cysteine, appears with very high to high lifts in pairs with the following residues in binding sites: (ordered in decreasing lift): SER, GLN, GLY, HIS, ASP, VAL, ALA, ILE, TRP, THR, LEU.

On the other hand, another, relatively infrequent amino acid, the tryptophan, appears with lifts less than one with several residues: LYS, THR and with lower statistical significance with ALA, ARG and ILE. Tryptophan is a large, apolar amino acid, mostly found in the inside of protein molecules. Note, that this fact does not explain the anomaly: atoms in tryptophan was counted in 33.56 % of all binding sites in binding distance to ligands, so tryptophans in the analysis are on the surface of the proteins.

It is also an interesting fact that the (TRP, TYR) pair has the fourteenth largest lift, and also has a large frequency. Consequently, tryptophan is unusually rare only when in pair with certain residues; tryptophan can be found with tyrosine (and also with histidine or glutamine) much more frequently.

Phenylalanine is also frequently present in infrequently appearing pairs; e.g., in Table 2 it appears three times in the right column.

Table 2 shows the ten amino acid pairs of the largest and ten pairs of the smallest lifts.

Note also, that the cysteine containing triples CYS,LYS,MET and CYS,LYS,TRP are in the 20 triples of the lowest lift, while cysteine containing pairs and triples are among those with the highest lift.

Turning to even larger subsets, it is worth mentioning that the five-tuple ALA-CYS-ILE-PRO-THR appears with more than 4 times higher frequency in binding sites as expected from the individual frequencies of these five residues.

Methods

Creating the RS-PDB database

The PDB entry [1] in mmCIF format consists of several tables, called "data categories", and the attributes in a table are called "data items". The most important mmCIF data categories are:

- struct_asym: List of the components in the asymmetric unit. Each component has an asym id.
- pdbx_poly_seq_scheme: Describes the sequence of monomers in a polymer entity.
- pdbx_nonpoly_scheme: List of the monomers belonging to the non-polymer entities.

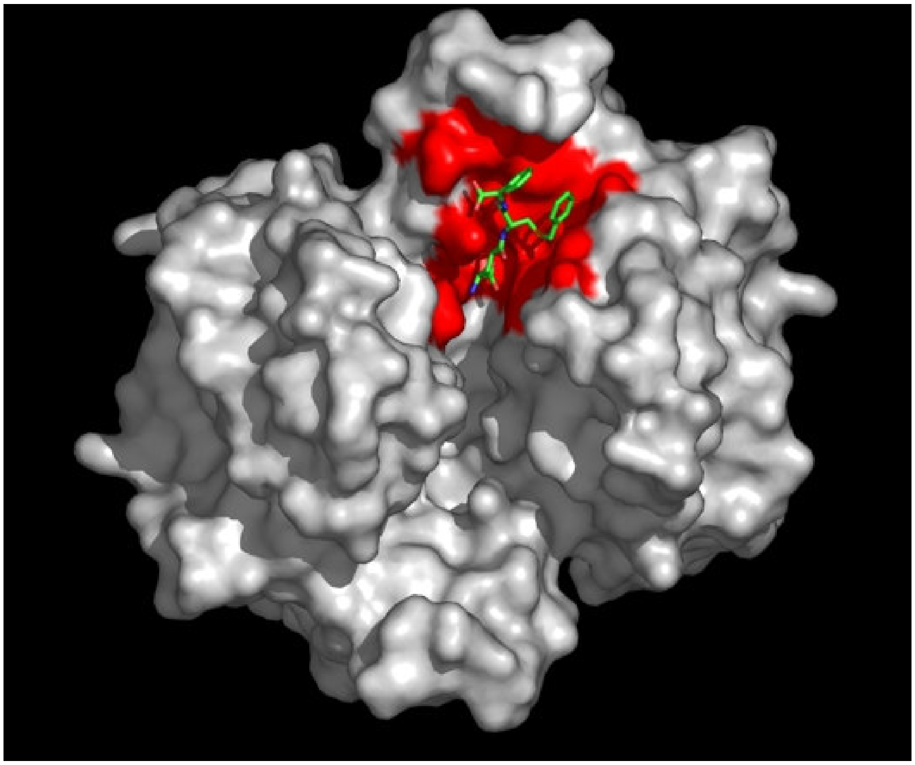

Figure 1: *We collected the residues, forming the binding sites for all PDB entries. On this figure, red colored area contains the residues of a binding site on PDB entry 10GS (Human glutathione s-transferase p1-1, complex with ligand ter117 (Gamma-glutamyl-(l)-(s-benzyl)cysteinyl-(d)-phenylglycine)).*

- atom_site: Coordinate data for atoms, whose positions could be experimentally determined.

The following entry header information is stored in the RS-PDB database:

- The species of the source organism(s) from which the structure was obtained. This can be found in the entity_src_gen table if the source was genetically manipulated, otherwise in the entity_src_nat table.
- The method used in the experiment (exptl.method), e.g., X-Ray Diffraction, NMR.
- The resolution of the structure (refine.ls_̣_res_high).

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.

From now on, we call the elements of the PDB Chemical Component Dictionary (formerly the HET Group Dictionary), found on the location

ftp://ftp.rcsb.org/pub/pdb/data/monomers/components.cif, "monomers".

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 (formerly HET Group Dictionary, HGD).

Connecting the monomers to obtain the covalent structure of the whole chain is performed by adding the monomers to the chain one-by-one:

In the case of protein chains, when we add a new amino acid (i.e., a monomer), we remove the atoms OXT and HXT from the end of the chain, and the atom HN2 (it is sometimes denoted by 2HN) from the new monomer, and add a covalent bond between the atoms C and N. In the case of PRO, we remove both HT1 and HT2. If, in the case of a non-standard amino acid (i.e., protein monomer), the above mentioned atoms are not present, we refuse to make chain. In this way we can ensure that only proteins with standard peptide bonds will be processed, without excluding the numerous modified amino acid monomers that can be found in the HGD. We use a similar protocol in the case of DNA/RNA chains as well.

Selecting the initial set of ligands: After creating the connection table for the polymer chains, the list of monomers from the table pdbx_nonpoly_scheme will be read. The initial set of ligand molecules will be these, plus the monomers from the polymer entities that were not long enough (these will form the oligopeptide ligands, for example). We obtain their connection table from the HGD. If it cannot be found there, an error is given. In this way, we will work only with previously checked components.

What we now have, is the covalent structure of few polymer chains, and the initial set of ligands. At this step the 3D atomic coordinates of the atoms are not processed yet.

Inserting atomic coordinates: The coordinates of the atoms can be found in the mmCIF table atom_site. By going through each row of this table, we need to identify the atom in our previously built covalent model, that this row is referring to. This is not always an easy task, because there are four different numbering schemes used simultaneously, but we try a few combinations before giving up, and each time we also check that the monomer id of this row matches the monomer id of the atom. If we fail to find the atom for a certain row, we give an error message and stop processing. After reading the table atom_site, there will be several atoms, whose coordinates are known. But unfortunately, there will still be several atoms whose coordinates are unknown. We will refer to them as ”missing atoms” hereafter. There are three different reasons for an atom to be missing.

- First, the hydrogen atoms can not be “seen” on the electron density maps, so they are usually missing, this is a completely normal case.
- Second, there can be flexible chain segments, a few residues at the beginning or at the end, or a longer loop at the middle of the chain. The position of these flexible parts can not be determined, so the atoms in them are all missing, not only the hydrogen atoms.
- The third reason: there can be atoms, that are in this initial ligand set only and will not be part of the final structure as we shall see later.

The next step is verifying distances. Now that the largest part of the atoms of our molecules are located in the space, we can check whether the bond lengths are correct. It is done by taking all pairs of atoms that are in the same monomer, and check whether their distance is in accordance with the connectivity information in the HGD. That is, if they are covalently bound, the proportion of their distance to the sum of their covalent radii should be 0.75 to 1.25 (in this case we call the atoms to be in ”covalent range”), and if they are not, then it should be more than 1.25. The bound atom-pairs, what are not bound covalently, should be closer than 1.05-times the sum of their Van der Waals radii. The lengths of the peptide bonds, that were added later, are also checked. The deviances are recorded in a separate table in the database, so that one can use this information, when selecting the most exact structures.

Building multi-monomer ligands: At this point, we still have the initial set of ligands. A molecule in the final set can consist of two or more such monomers (with three-letter HGD-code), bound covalently. To identify such covalent bonds, we select all pairs of atoms in the entry that are situated closer than 6 Å. This is achieved by building a kd-tree [6] on the atoms, avoiding the examination of all pairs, and saving a considerable amount of computational time. As a byproduct, the pairs of atoms that are too close to each other are also obtained and recorded as a warning in the database. For the covalent bonds we consider the atom pairs that are in covalent range, as defined above. To actually add a covalent bond, one of the following three conditions has to be met:

- There is a missing hydroxyl group on one atom, and at least one missing hydrogen on the other atom. In this case we remove these three atoms, and add the covalent bond. This way, the number of missing heavy atoms can decrease.

- Both atoms are sulfur of a CYS residue. In this case we remove a hydrogen from each atom, and add the covalent bond.
- One atom is a metal, and the other atom is a non-metal that can donate a lone pair. In this case we simply add a coordinate covalent bond.

If a covalent bond was made between the atoms of two ligand molecules in the above process, then these two molecules will be merged into one. If the bond was made between the atom of a polymer and an atom of a ligand, then we do not merge them to preserve the linear structure of the polymer, but record this bond in a separate table in the database. If no bond was made between the atoms, that were closer than 6 Å, then we take the two molecules they were part of, and add a logical "near-to" relation between them. This will define a graph on the molecules as the nodes, and these near-to relations as the edges. We will call it the component graph of the entry.

Finalizing the set of ligands: After creating all possible covalent bonds, we have the set of ligands. The set of ligands will be further filtered as follows.

First, we apply and slightly extend the monomer characterization of Wang et al. [10, 9], given in their Table 2. They had a list of "special" monomers, consisting of biologically important cofactors/coenzymes, as well as "junk" molecules, that were found in many PDB entries. We defined further types of monomers, such as modified amino acids, modified nucleic acids, metals, and organo-metallic molecules.

- The modified amino/nucleic acids were identified with the help of the mon_nstd_parent_comp_id attribute of the monomers in the mmCIF format of the HGD. If this parent attribute was one of the standard amino/nucleic acids, then the monomer was assumed to be its modified form.
- Metal monomers are those that contain at least one metal atom, but no carbon atom.
- Organo-metallic monomers are those containing both metal and carbon atoms.
- The monomers that had less than six heavy atoms and were not classified before, were assigned the "tiny" type.
- The other monomers were classified as standard ligand building stones, called pro-ligand-monomers.

See Table 3 for the distribution of the monomers.

Based on this classification of the monomers, we define the following categories of ligands:

- Peptide: contains at least one amino acid monomer, and all of its monomers are (modified) amino acids.
- Cofactor/coenzyme: contains at least one such monomer.
- Junk: contains at least one junk monomer and all of its monomers are junk, metal or tiny
- Metal: contains at least one metal monomer and all of its monomers are metal or water.
- Tiny: it consists of a single tiny monomer or it is not classified above and has less than six heavy atoms.
- Pro-ligand: it is not classified above and it contains at least one (modified) amino-or nucleic acid or organo-metallic or ligand monomer.

Table 1 - The frequencies of the 1-element residue-sets in the binding sites

The numbers in Table 1 give the fractions of binding sites where the amino acid in question appears. For example, GLY is present in 62.56% of all binding sites (that is, 0.6256 fraction of the all binding sites).

Residue	Frequency	Residue	Frequency	Residue	Frequency
GLY	0.6256	LEU	0.5823	TYR	0.5568
ARG	0.5386	SER	0.5227	ALA	0.5184
PHE	0.5101	VAL	0.5016	ASP	0.4817
THR	0.4748	ILE	0.4660	ASN	0.4336
HIS	0.4254	LYS	0.4221	GLU	0.4110
TRP	0.3356	GLN	0.3111	PRO	0.2859
MET	0.2830	CYS	0.2094		

Our definition of ligands

Molecules, classified as peptides, cofactors and pro-ligands above were inserted in the database, and only binding sites, containing these ligands connected to proteins (non-covalently) are further investigated in this work. The word "ligand" means one molecule from the three categories above in what follows.

Finally, the coordinates of the hydrogen atoms are computed using standard hybridization rules for the entries in the database.

The resulting database

The output of the algorithm, the RS-PDB database consists of three parts:

- The Small Molecule Database. For any atom only its element symbol, its formal charge, and its model coordinates are stored. The bonds can be of order one, two or three, and we store the Kekule representation of the aromatic systems. If a new molecule is inserted into this part of the database, then first its InChITM [2, 3] 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$^{\text{TM}}$ is found.
- PDB Chemical Component Dictionary (formerly the HET Group Dictionary). The second part of the database contains the PDB Chemical Component Dictionary. Its mmCIF format, components.cif can be downloaded from Protein Data Bank [1]. It consists of the connection tables of the molecules and functional groups associated with the three letter monomer codes found in the PDB. These connection tables are inserted into the small molecule database after conversion and error-checking. Since the mmCIF format contains four types of bonds: single, double, triple and aromatic, the latter had to be converted to the Kekule representation. Because the aromatic bonds are special types of ring bonds, each bond in a molecule marked as aromatic is checked to be a part of a ring. This was done with a depth-first-search graph-traversal of the molecule [5]. Next, with a graph matching algorithm [4], those bonds were selected from the aromatic ones which need to be double bonds. Before inserting a monomer into the database the standard "valence rules" are verified as described in the InChI Technical Manual.
- The Main Part of the Database. The results of processing of the PDB entries are stored in the third, and most important part of the database. 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 the case that was already there, then we obtain its molid. For every ligand it is recorded whether they bind covalently to a polymer chain, and also the id of the binding site they are in. For the polymers we record only their type ("P"=protein chain and "N"=DNA/RNA chain) and their length in monomers. The table listing the atomic coordinates will contain an attribute called the status of the atom. This can be 'M'= missing,

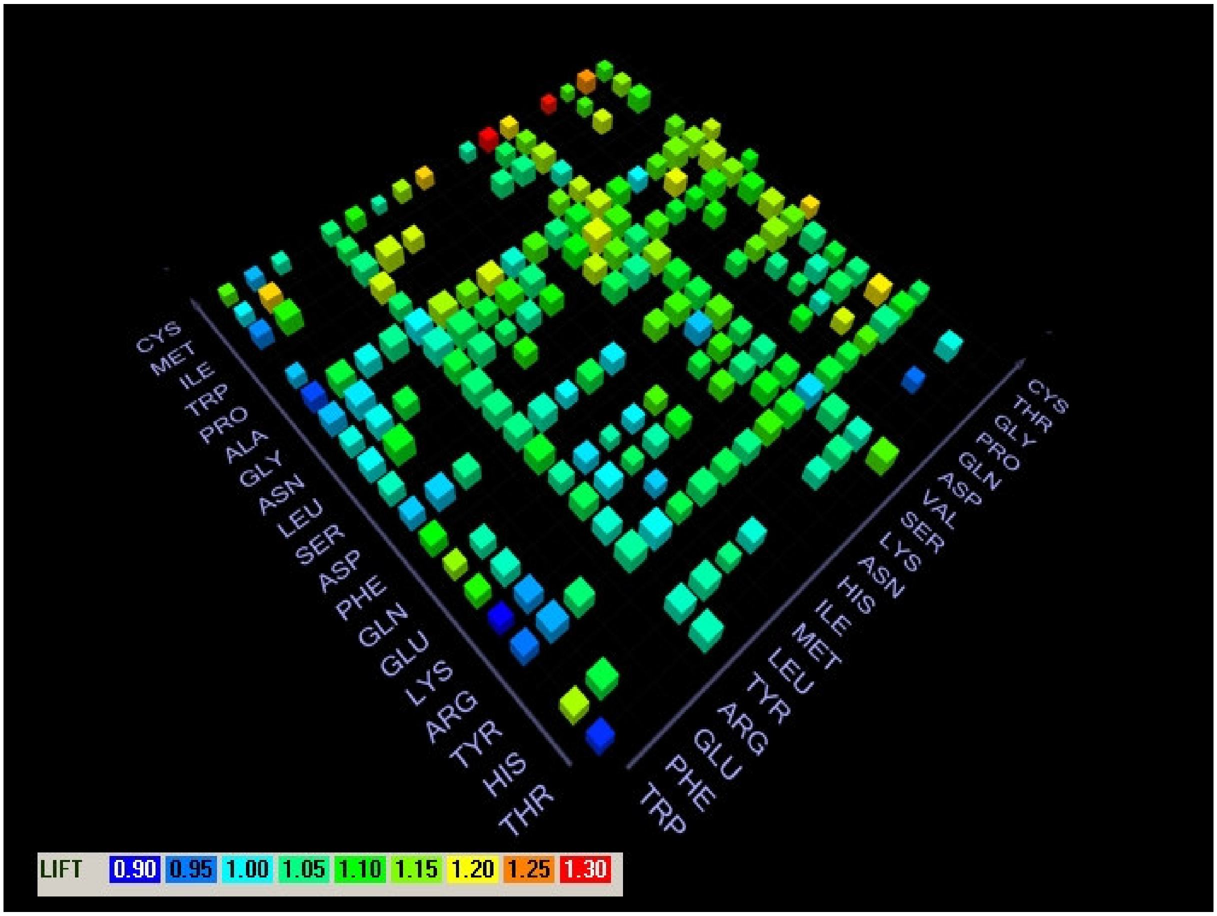

Figure 2: *Here we give the lifts of each residue-pairs color-coded according to the legend. Each pair is present only once. The cysteine and the tryptophane anomalies are clearly visible.*

'E'= experimentally determined and 'C'= computed (as in the case of hydrogen atoms). The cnum attribute in each table means the canonical number of the atom in the given molecule. For the polymer chains, we define the canonical number of an atom by adding 1000*(residue sequence number) to its canonical number within the monomer residue. Because the covalent bonds within the ligands are stored in the small molecule database, we only have to record the bonds between the ligands and the polymer chains here, as well as the disulphide bonds between the polymer chains. The type of these bonds can be covalent and coordinated covalent denoted by COV and COORD.

Table 2 - Ten pairs of the largest and of the smallest lifts.

Residues	Lift	Residues	Lift
CYS,SER	1.301	LYS,TRP	0.902
CYS,GLN	1.294	THR,TRP	0.921
CYS,GLY	1.245	ALA,TRP	0.931
CYS,HIS	1.220	HIS,PRO	0.947
MET,PHE	1.219	ARG,TRP	0.949
ASP,CYS	1.214	ILE,TRP	0.957
CYS,VAL	1.212	LYS,PHE	0.966
LYS,THR	1.209	ARG,PHE	0.970
GLY,PRO	1.194	CYS,PHE	0.974
GLY,SER	1.189	GLY,TRP	0.977

Getting rid of redundancies

The RS-PDB database is prepared from the Protein Data Bank. In the PDB, important (or just popular) proteins are present in more than one copies: different PDB entries frequently contains the same protein sequence with different ligands, co-factors or with different resolution. For example, the protein chain of bovine trypsin is present in 165 different PDB entries, and three other protein sequences appears in more than 100 PDB entries each.

The composition of the PDB is clearly biased to the direction of "important" proteins. Since popular or important structures were deposited with a large multiplicity, it is essential to count them only once if we aim to have correct results concerning the frequency of appearance of certain subsets of residues on binding sites.

All the different (protein-surface area, ligand molecule) pairs were identified, and the redundancies were deleted from our database: if at the same area two different ligands were bound in two different PDB entries, then they were counted twice; if the same ligand appeared twice on the same area in two different PDB entries, they were counted only once.

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.

However, this count still contains numerous redundancies: the very same structures may be listed many times. For example, the PDB id 1R15 is present 24 times in the our database, 16 times bound to the ligand nicotinamide. Multiple copies of the nicotinamide ligand is bound to the atoms of residues ASN, GLU, SER, TRP. We believe that these binding sites are essentially the same, and we aim to exclude such redundancies from our analysis. This goal is accomplished as follows.

First we consider the following data items from the RS-PDB database:

- the ligand participating in the bond;

- identifier of the protein from the PDB database;

For these items the following objects are created:

- number of protein chains the ligand binds to, by counting the distinct protein-sequence id's for the same ligand;
- for each protein chain, a binary vector (seq_vector) identifying the amino acids that participate in the bond within the protein chain. The vector's length equals to the length of the protein chain (measured in amino acid count), and the vector's n-th coordinate is set to X, if the n-th amino acid of the protein chain, participating in the bond with the current ligand is X.
- the amino acid profile of the binding site. This is a 21-bit value, the bits of which correspond to the 20 amino acids that can occur at a binding site, plus one entry for all the modified amino acids. A given bit is set to 1, if the appropriate amino acid occurs at the binding site; otherwise, the bit is set to zero. Data mining is performed by using amino acid profile; the rarely occurring modified amino acids are disregarded in the present study.

Two binding sites are considered to be "equal" here, if the following criteria are met:

- the ligand is the same at the two binding sites;
- the number of protein chains are the same at each of the binding sites;
- the set of the protein chains is equal, which means that for each protein chain from the first binding site, there can be found a protein chain at the second binding site with the same length and seq_vector value.

Note, that the exact correspondence of the amino-acid sequence of the protein chains is not a requirement: this increases the error-tolerance of the method, since even one erroneously given residue in a protein chain will alter the result. After several experiments, we found that the correspondence of the seq_vector value more reliably identify the identical binding sites than the correspondence of amino-acid sequence.

One binding site is chosen from each class of the "equal" ones, defined above. After the filtering process, we are left with 19,581 different binding sites.

Table 3 - Distribution of the monomers.

Monomer type	Number in HGD
Amino acid	20
modified amino acid	376
nucleic acid	6
modified nucleic acid	236
Cofactor/coenzyme	131
Junk	29
Metal	185
Organo-metallic	94
Tiny	221
Pro-ligand-monomer	4713

Conclusions

By giving a strictly mathematical (more exactly, graph-theoretical) re-structuring of the full PDB, we were able to identify and analyze all the binding sites contained in the Protein Data Bank. We identified subsets of residues on the binding sites and found statistical phenomena not observed before, concerning cysteine and tryptophan frequencies.

We believe that such exact re-structuring of the richest three-dimensional structural biology repository of the Earth, the Protein Data Bank, would yield much more interesting results in the future.

Equipment and settings

A custom-built dual OPTERON^{TM} server with Debian LINUX operational system was used for data-base building and analysis. The database engine was the MySQL^{TM} 4.1 server of the MySQL AB.

Figure 1 in the main text was created by using PyMOL http://pymol.sourceforge.net/ with the raytracing option.

Frequent datasets were identified by the apriori algorithm [19], with our own implementation.

Figure 2 in the main text was produced by the MineSet^{TM} suite's visualization component of the Purple Insight Company.

The Microsoft Excel^{TM} tables were created from multiple conversions from MySQL^{TM} format, through ODBC from Microsoft Access^{TM} database engine.

References

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[2] Sofie L. Rovner. Chemical 'naming' method unveiled. *Chem. & Eng. News*, 83:39–40, 2005.

[3] David Adam. Chemists synthesize a single naming system. *Nature*, 417(369), 2002.

[4] L. Lovász and M. D. Plummer. *Matching theory*, volume 121 of *North-Holland Mathematics Studies*. North-Holland Publishing Co., Amsterdam, 1986. Annals of Discrete Mathematics, 29.

[5] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to algorithms*. MIT Press, Cambridge, MA, second edition, 2001.

[6] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[7] R. A. Laskowski. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Research*, 29:221–222, 2001.

[8] Nicodeme Paul, Esther Kellenberger, Guillaume Bret, Pascal Muller, and Didier Rognan. Recovering the true targets of specific ligands by virtual screening of the Protein Data Bank. *Proteins: Structure, Function, and Bioinformatics*, 54(4):671 – 680, 2004. http://bioinfo-pharma.u-strasbg.fr/scpdb/scpdb_form.html.

[9] Renxian Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The PDBbind database: methodologies and updates. *J. Med. Chem.*, 48:4111–4119, 2005.

[10] Renxian Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.*, 47:2977–2980, 2004.

[11] Howard J Feldman, Kevin A Snyder, Amy Ticoll, Greg Pintilie, and Christopher W V Hogue. A complete small molecule dataset from the protein data bank. *FEBS Lett*, 580(6):1649–1653, Mar 2006.

[12] Manfred Hendlich, Andreas Bergner, Judith GÃ$\frac{1}{4}$nther, and Gerhard Klebe. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J Mol Biol*, 326(2):607–620, Feb 2003.

[13] Yanli Wang, Kenneth J Addess, Jie Chen, Lewis Y Geer, Jane He, Siqian He, Shennan Lu, Thomas Madej, Aron Marchler-Bauer, Paul A Thiessen, Naigong Zhang, and Stephen H Bryant. Mmdb: annotating protein sequences with entrez's 3d-structure database. *Nucleic Acids Res*, 35(Database issue):D298–D300, Jan 2007.

[14] Jie Chen, John B Anderson, Carol DeWeese-Scott, Natalie D Fedorova, Lewis Y Geer, Siqian He, David I Hurwitz, John D Jackson, Aviva R Jacobs, Christopher J Lanczycki, Cynthia A Liebert, Chunlei Liu, Thomas Madej, Aron Marchler-Bauer, Gabriele H Marchler, Raja Mazumder, Anastasia N Nikolskaya, Bachoti S Rao, Anna R Panchenko, Benjamin A Shoemaker, Vahan Simonyan, James S Song, Paul A Thiessen, Sona Vasudevan, Yanli Wang, Roxanne A Yamashita, Jodie J Yin, and Stephen H Bryant. Mmdb: Entrez's 3d-structure database. *Nucleic Acids Res*, 31(1):474–477, Jan 2003.

[15] Yanli Wang, John B Anderson, Jie Chen, Lewis Y Geer, Siqian He, David I Hurwitz, Cynthia A Liebert, Thomas Madej, Gabriele H Marchler, Aron Marchler-Bauer, Anna R Panchenko, Benjamin A Shoemaker, James S Song, Paul A Thiessen, Roxanne A Yamashita, and Stephen H Bryant. Mmdb: Entrez's 3d-structure database. *Nucleic Acids Res*, 30(1):249–252, Jan 2002.

[16] Y. Wang, K. J. Addess, L. Geer, T. Madej, A. Marchler-Bauer, D. Zimmerman, and S. H. Bryant. Mmdb: 3d structure data in entrez. *Nucleic Acids Res*, 28(1):243–245, Jan 2000.

[17] A. Marchler-Bauer, K. J. Addess, C. Chappey, L. Geer, T. Madej, Y. Matsuo, Y. Wang, and S. H. Bryant. Mmdb: Entrez's 3d structure database. *Nucleic Acids Res*, 27(1):240–243, Jan 1999.

[18] H. Ohkawa, J. Ostell, and S. Bryant. Mmdb: an asn.1 specification for macromolecular structure. *Proc Int Conf Intell Syst Mol Biol*, 3:259–267, 1995.

[19] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2000.

[20] Torben Hagerup and Christine Rüb. A guided tour of Chernoff bounds. *Inform. Process. Lett.*, 33(6):305–308, 1990.