Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Being a binding site: Characterizing residue composition of binding sites on proteins

Gábor Iván[1, 2], Zoltán Szabadka[1, 2], Vince Grolmusz[1, 2, *]

[1]Protein Information Technology Group, Department of Computer Science, Eötvös University, Pázmány P. stny. 1/C, H-1117 Budapest, Hungary; [2]Uratim Ltd., Sóstói út 31/b, H-4400, Nyíregyháza, Hungary;
Vince Grolmusz* - E-mail: grolmusz@cs.elte.hu; * Corresponding author

received October 25, 2007; accepted December 29, 2007; published online December 30, 2007

Abstract:
The Protein Data Bank contains the description of more than 45,000 three-dimensional protein and nucleic-acid structures today. 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. We first cleaned and re-structured the PDB data, then analyzed the residue composition of the binding sites in the whole PDB for frequency and for hidden association rules. Main results of the paper: (i) the cleaning and repairing algorithm (ii) redundancy elimination from the data (iii) application of association rule mining to the cleaned non-redundant data set. We have found numerous significant relations of the residue-composition of the ligand binding sites on protein surfaces, summarized in two figures. One of the classical data-mining methods for exploring implication-rules, the association-rule mining, is capable to find previously unknown residue-set preferences of bind ligands on protein surfaces. Since protein-ligand binding is a key step in enzymatic mechanisms and in drug discovery, these uncovered preferences in the study of more than 19,500 binding sites may help in identifying new binding protein-ligand pairs.

Keywords: binding site; functions; structural data; protein; association rules

Background:
The increasing accuracy and size of structural information stored in the Protein Data Bank **[1]** makes possible 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. Since most of the known biologically active compounds are ligands bound to proteins, this study is of considerable importance in the mathematical foundations of drug discovery and drug design. In the present work we apply data-mining techniques for the sets, formed from the residues at each binding sites present in the whole Protein Data Bank. Note that protein-ligand binding is a key step in enzymatic mechanisms, therefore the classification, characterization and the analysis of the binding sites is of special importance in understanding, predicting and designing enzymatic mechanisms. On the other hand, most drugs used today are small ligand molecules, and they act through binding to proteins or enzymes, and modifying their biological roles. Therefore, any large scale study of binding sites on proteins is of special importance in several fields. The rigorous cleaning and re-structuring procedure for the entries in the Protein Data Bank was reported in our earlier work **[2]**. We made use of the following techniques in the creation of that RS-PDB database: Computing the InChI™ code **[3, 4]** applied a graph-isomorphism testing, transforming aromatic notation to Kekule-notation used a non-bipartite graph-matching algorithm **[5]**, breadth-first-search graph traversals **[6]** were used throughout the work **[2]**, depth-first search **[6]** was used in building the ligand molecules and identifying ring structures, Kd-trees **[7]** were applied for computing covalent bonds, and hashing **[6]** were utilized for the fast generation of protein-sequence IDs.

Methodology:

Identification of protein-ligand complexes

It was 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 or atoms of modified residues or even atoms in small molecules added during crystallization or covalently 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 examination 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 **[8]** contains reliable structural information on ligands and binding sites, but, unfortunately, the automatic processing of the database is not possible. The sc-PDB database **[9]** 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 **[10]**. In the PDBbind Database **[11]**, by manual, human-involved search, binding affinities were compiled from hundreds of biochemical publications about protein-ligand complexes, from the PDB. We have chosen a much more reliable, fully automatic mathematical method for identifying complexes in **[2]**. We presented a new pre-processing algorithm that worked on the mmCIF (macromolecular Crystallographic Information File) format of the PDB, and used the International Chemical Identifier (InChI^{TM}) of the International Union of Pure and Applied Chemistry (IUPAC) **[3, 4]**. The result was 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 **[10]**. 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). One of the derived tables, called binding sites, contains the description of all the identified protein-ligand complexes from the whole PDB: this table consists of more than 1.9 million rows, and served as the basis of the further applications in the present work.

Results and discussion:

Dealing with data redundancy

The RS-PDB database is prepared from the Protein Data Bank. In the PDB, important proteins are present in more than one copy: different PDB entries frequently contain 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 appear in more than 100 PDB entries each. In our study we cannot consider all the binding sites on the surface of all the proteins in the world, just those structures what are deposited in the PDB. 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 quantitative 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. Note that this way we examined proteins even with very small differences (e.g., point-mutations) from the PDB as different entities; we did not delete them. The reason for that was that even point-mutations may change the residue-composition of a binding site, and that is our main subject in this study.

Binding sites and residues

After the rigorous ligand identification and redundancy-deleting procedure, we gained 19,581 pair-wise different binding sites. 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. 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 identified 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.

Our goal was to analyze the properties of these ligand-binding residue-sets by finding hidden association rules **[12]**. The frequent item-sets were collected using the apriori algorithm **[12]**. The frequencies of the individual amino acids (i.e., the 1-element subsets) are given in Table 1 (supplementary material). The numbers in Table 1 (supplementary material) 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).

Association rules

From the observations in Table 1 (see supplementary material), the distribution of the amino acids in the binding sites is far from being uniform. We would like to reveal hidden rules of the residue-composition of binding sites of the protein-ligand complexes. We are interested in implication-like rules such as: (ALA, LEU) → (ILE, VAL), that is, if a binding site contains amino acids leucine and alanine, it will "likely" contain also valine and isoleucine. Here the word "likely" needs further clarification: in our study we have found, that 14.68% of all binding sites contain all the four residues leucine, alanine, valine and isoleucine. Moreover, from all the binding sites, containing both alanine and leucine, 41.69% contain also isoleucine and valine. These relations are called association rules in data mining **[12]**, and are believed to describe hidden rules or implications in large enough data sets **[13, 14, 15]**.

We give here the terminology and the definitions needed to

describe the results, following the data mining terminology **[12]**. Let I denote a finite set, its elements are called items, (in our study the items are residues). Let T denote a set, consisting of several subsets of set I, where multiplicities are also allowed: subsets of I can be contained in T more than once. The elements of T (which are subsets of I in the same time) are called transactions. Note, that in our present study the transactions are the sets of amino acids in binding sites. Let U and V be two disjoint subsets of I. Association rule $U \rightarrow V$, has support α, if the union $U \cup V$ is a subset of exactly the α-fraction of all transactions (or with the probability-notation $Pr_X (U \cup V \subseteq X) = \alpha$); it has confidence β, if among the transactions, containing the whole set U , a β fraction contains also set V (or, with conditional probabilities: $Pr_X (V \subseteq X \mid U \subseteq X) = \beta$). Rule $U \rightarrow V$ has lift γ, if γ equals to the confidence, divided by the probability of containing V, (or, in other words, $\gamma = Pr_X (V \subseteq X \mid U \subseteq X)/Pr(V \subseteq X) = Pr(U \cup V \subseteq X)/(Pr(U \subseteq X)Pr(V \subseteq X))$. The support of an association rule characterizes its frequency, the confidence and the lift its interest. In our example association rule (ALA, LEU) → (ILE, VAL), the support is 14.68%, the confidence is 41.69%, its lift is 1.53. The meaning of the lift can be explained as follows: the support of the set {ALA, LEU} is 0.3520, the support of set {ILE, VAL} is 0.2732, the support of {ALA, LEU, ILE, VAL} is 0.1468; this is 1.53-times higher than the value $0.3520 \cdot 0.2732 = 0.09616$. This example shows that the presence of leucine and alanine 1.53 times increase the probability of the presence of valine and isoleucine at a binding site.

Association rules $X \rightarrow Y$, where Y is a very frequently appearing residue-subset, are not interesting generally: Since Y is very frequent, it will appear in a great variety of different residue-subsets, and the $X \rightarrow Y$ rules, with different X's will be valid, but, however, they mean nothing else, that Y is quite frequent. In data mining, this situation is explained as follows: Suppose that the transactions are consumer baskets in a supermarket. Since the great majority of the consumers buy bread, there will be lots of association rules of the form $X \rightarrow$ (bread), and they be valid in the sense that they will have large support and confidence, but they will not generally be interesting and will not generally uncover new facts. On the other hand, if Y is an infrequently appearing set, then the support and the confidence generally will not reach the thresholds to be included in our results. For example, Y = GLY appears very frequently, while Y = CYS or Y = TRP appears very rarely. Association rules of unusually high and unusually low lifts and rules of form $X \rightarrow Y$ with high confidence and not-too-high support for Y are of particular interest **[12]**. Our figures here visualize such remarkable data.

With the parameters of Figure 1, only the arginine is present as the head of 8 arrows of lift less than 1. This may show that arginine, the fourth most frequent residue in binding sites (Table 1 in supplementary material) relatively infrequently appears together with some residue-sets, but its frequency is high enough to give larger than 0.5 confidence. It is interesting that only arginine has this property. The star with a leucine-center may be due that leucine is the second most frequent residue in binding sites after glycine (Table 1 in supplementary material), and rules pointing to glycine were deleted. It is remarkable, however, that the alanine-isoleucine pair is very frequently contained at the base of those arrows in the star. The higher than 1.8 lifts in the lower half of Figure 1 shows a common property: all sets at beginning of the arrows are of low support, and at the end of the arrows are of high support (see Figure 2 for legends). The data show that the frequency of the rarely appearing sets there will be increased relatively strongly if they appear together with the sets at the tip of the arrows.

On Figure 2 those rules are listed, which has high lift, high support and high confidence, but the $X \rightarrow$ GLY rules are deleted for clarity. Note, that all stars with in-degree larger than 3 contains GLY, which is not really surprising, by the data of Table 1 (supplementary material). However, the star of [GLY] + [THR] in the upper right corner has two remarkable properties: (i) alanine (the sixth most frequent residue) is present in almost all bases, and (ii) threonine (the tenth most frequent residue) appears with GLY in the centre (iii) bases contain three or four residues. This result shows that from a very large data set one can derive surprising facts by setting the thresholds properly.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

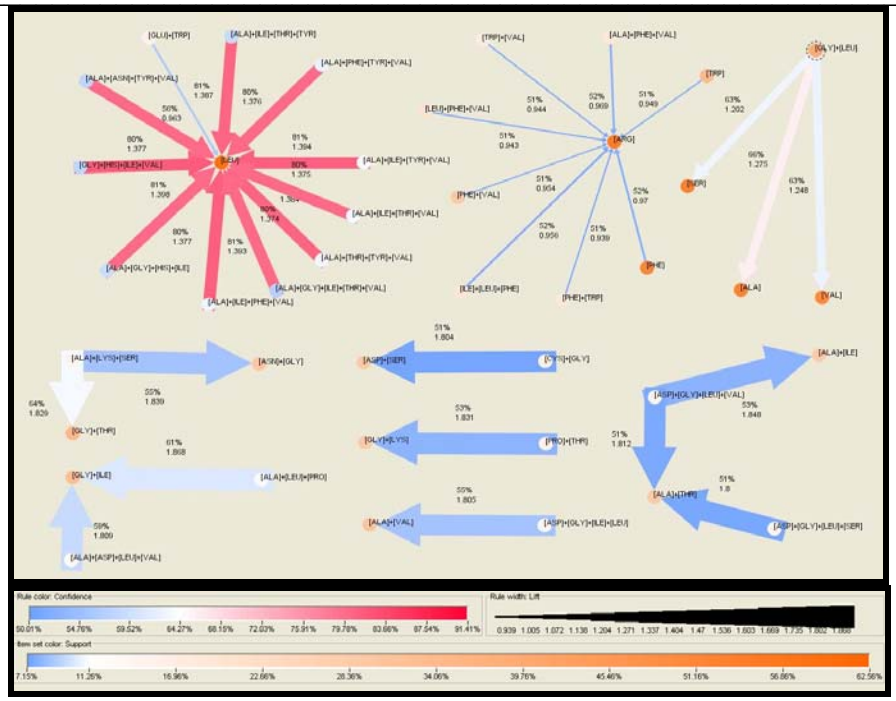

Figure 1: Association rules-Set 1: Figure 1 was created by deleting all $X \rightarrow GLY$ association rules for clarity, and including those rules which satisfy that their supports are at least 7.15% and their confidences are at least 0.5 and, moreover, at least one of the following conditions hold: (a) their confidences are at least 0.8 or (b) their lifts are at least 1.8 or (c) their lifts are at most 0.97 or (d) their supports are at least 24%. The color and width of the arrows corresponds to the lift, the color of residue-sets corresponds to the support, as shown on the figure legend. Four areas are identifiable on the figure: in the lower half the rules of large lifts are shown; in the upper left corner the rules of high confidences (with one exception), in the upper middle part the lower than 0.97 lift rules, and in the upper right corner the high support rules are shown. Note that these rules form almost disjoint classes

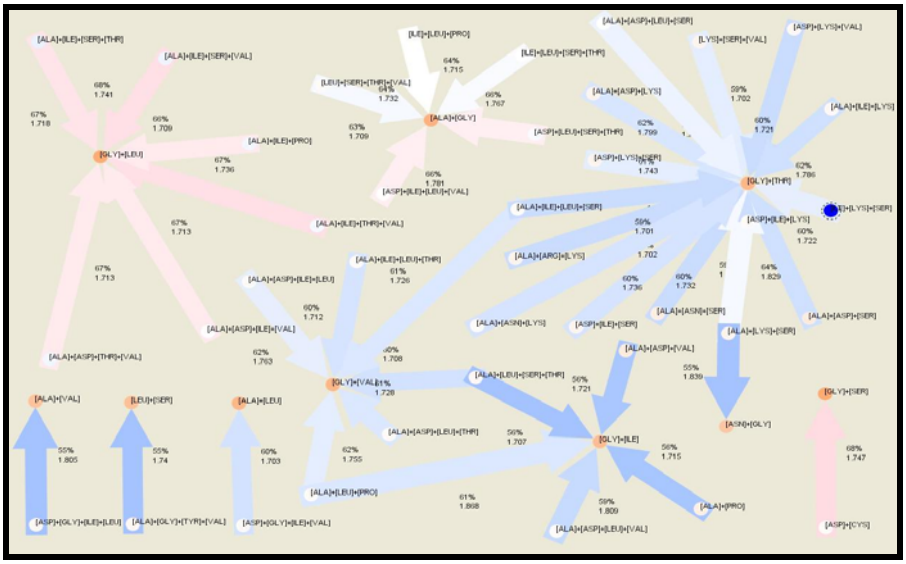

Figure 2: Association rules-Set 2: The figure was created by deleting all $X \rightarrow GLY$ association rules for clarity, and including only those rules which satisfy that their support is at least 7.15% and their confidence is at least 0.55 and their lift is at least 1.7

Conclusion:

We cleaned the largest depository of the three-dimensional structural protein information database, the Protein Data Bank **[1, 2]**, and also identified a non-redundant set of more than 19,500 ligand binding sites in the data. We collected the residues from the binding-site collection, and analyzed the residue-sets by association-rule mining. Hundreds of thousands of association rules can be identified in such a large dataset. It is a challenging task to reduce the number of the association rules by selecting the relevant ones. We applied filtering according to support, confidence and lift, connected by Boolean relations in different ways in Figure 1 and Figure 2. This way we gained easily readable association rule sets. The PDB is the result of the work of tens of thousands scientists (biochemists, physicists, crystallographers, mathematicians) in about forty years. The more than 19,500 binding sites identified by us are clearly contains information of enormous biological value, hard to understand and interpret. The data mining techniques developed in the last decade make possible to gain knowledge from large and hardly manageable datasets. Our results here demonstrate this fact. Most probably, every association rule on our figures has non-trivial biochemical meaning. The explanation even of these very restricted set of rules are out of the scope of the present work. We just would like to mention just one interesting relation from the figures here. On Figure 1 there is a star with an ARG center, containing lower-than-1 lift rules. The confidences are just above the threshold 0.5 there, meaning that, for example, if a binding site contains PHE and VAL, then it will also contain ARG in 51% in all cases. However, the low lift (0.954) means that the (PHE, VAL, ARG) triad is present with less frequency than it can be predicted from the individual (marginal) probabilities of (PHE, VAL) and ARG. The explanation of that fact can be that the pair (PHE, VAL) is over-represented (given by lift) together with other residues, for example with LEU, ALA, and TYR on one arrow on the upper left star on Figure 1. By this avenue, we believe, one can create such large, only machine-handle-able rule-sets, that can evaluate and predict protein-ligand binding by taking into account several points in parallel. We think that our present work is a step in this direction, and can open this line of research.

Acknowledgement:

We thank Balázs Kósa for his help in the early stages of this work. This research was partially supported by the European Commission FP6 program "scrIN-SILICO" and by the Hungarian OTKA agency, under grant Nos: T046234 and N67867. Parts of this work were done in cooperation with Math-for-Health LLC.

References:

[01] H. Berman, *et al.*, *Nucleic Acids Research*, 28: 235 (2000) [PMID: 10592235]

[02] Z. Szabadka & V. Grolmusz, *Proceedings of the 28th IEEE EMBS Annual International Conference*, New York, 5755 (2006)

[03] S. L. Rovner, *Chem. and Eng. News*, 83: 39 (2005)

[04] D. Adam, *Nature*, 417: 369 (2002) [PMID: 12024181]

[05] L. Lovász & M. D. Plummer, *Matching theory*, *North-Holland Publishing Co*., 121 (1986)

[06] T. H. Cormen, *et al.*, *Introduction to Algorithms, MIT Press*, (2001)

[07] J. L. Bentley, *Communications of the ACM*, 18: 509 (1975)

[08] R. A. Laskowski, *Nucleic Acids Research,* 29: 221 (2001) [PMID: 11125097]

[09] N. Paul, *et al., Proteins,* 4: 671 (2004)

[10] R. Wang, *et al.*, *J. Med. Chem.*, 48: 4111 (2005) [PMID: 15943484]

[11] R. Wang, *et al.*, *J. Med. Chem.*, 47: 2977 (2004) [PMID: 15163179]

[12] J. Han & M. Kamber, *Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers* (2000)

[13] I. Artamonova, *et al.*, *Bioinformatics,* 21: iii49 (2005) [PMID: 16306393]

[14] P. Carmona-Saez, *et al.*, *BMC Bioinformatics*, 7: 54 (2006) [PMID: 16464256]

[15] T. Oyama, *et al.*, *Bioinformatics*, 18: 1151 (2002) [PMID: 12050067]

Edited by J. C. Tong, T. W. Tan & S. Ranganathan

Citation: Ivan *et al.*, Bioinformation 2(5): 216-221 (2007)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Supplementary material

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		

Table 1: The frequencies of the 1-element residue-sets in the binding sites. The numbers 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)