BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 3 2013, pages 395–397
doi:10.1093/bioinformatics/bts691

Structural bioinformatics

Advance Access publication November 29, 2012

PLI: a web-based tool for the comparison of protein-ligand interactions observed on PDB structures

Anna Maria Gallina[1], Paola Bisignano[1,2], Maurizio Bergamino[3] and Domenico Bordo[1,*]

[1]IRCCS Azienda Ospedaliera Universitaria San Martino—IST—Istituto Nazionale Ricerca sul Cancro, 16132 Genova, Italy, [2]Istituto Italiano di Tecnologia, 16163 Genova, Italy and [3]Dipartimento di Fisica, Universita' di Genova, 16146 Genova, Italy

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: A large fraction of the entries contained in the Protein Data Bank describe proteins in complex with low molecular weight molecules such as physiological compounds or synthetic drugs. In many cases, the same molecule is found in distinct protein-ligand complexes. There is an increasing interest in Medicinal Chemistry in comparing protein binding sites to get insight on interactions that modulate the binding specificity, as this structural information can be correlated with other experimental data of biochemical or physiological nature and may help in rational drug design.

Results: The web service protein-ligand interaction presented here provides a tool to analyse and compare the binding pockets of homologous proteins in complex with a selected ligand. The information is deduced from protein-ligand complexes present in the Protein Data Bank and stored in the underlying database.

Availability: Freely accessible at http://bioinformatics.istge.it/pli/.

Contact: domenico.bordo@istge.it

Received on September 19, 2012; revised on November 21, 2012; accepted on November 25, 2012

1 INTRODUCTION

About one fourth of the entries deposited with the Protein Data Bank (PDB; Dutta *et al.*, 2009) represent proteins in complex with small molecules (not including common salts and compounds commonly used in molecular and structural biology). The number of these ligands, in the PDB referred to as heterogeneous compounds, is currently ~14 000. Binding specificity is achieved by the formation of a network of interactions between the protein and the ligand, which depends on the shape and on the physicochemical nature of the amino acids forming the binding pocket, as well as on structure flexibility of both ligand and protein. In particular, ~17% of the ligands present in the PDB are found in complex with distinct proteins. The proteins involved may either share homology with each other or belong to distinct families. In the first case, it is possible to establish a relationship of structural equivalence among the amino acids forming the binding pockets; this allows to analyse the residue conservation and to identify interactions with the ligand, which are maintained despite mutated residues. If the proteins belong to distinct structural families, it might still be useful to compare the amino acid composition of the distinct binding pockets and the involved protein-ligand interactions. In the current release of the PDB database, a total number of 1593 ligands are found in complexes with at least two homologous proteins (adopting a maximum value of 95% identical residues to avoid similar or just point-mutated proteins), and 1049 are found in complex with at least two unrelated proteins. The PLI database and the associated web server described here have been developed to provide an easily usable tool to compare the binding pockets interacting with a selected ligand as observed in experimentally determined PDB protein complexes.

2 METHODS

The web tool is based on an underlying database, which is updated regularly with an automated procedure to include new PDB entries and heterogeneous compounds. For each heterogeneous compound, identified in the PDB with a Het_id code, the list of PDB entries that contain the compound is obtained from the PDBsum website (Laskowski *et al.*, 2005). To avoid the inclusion of small molecules having low specificity and therefore high frequency of occurrence, only ligands found in a maximum number of 25 PDB entries are included in the database. This resulted in the exclusion of ~3% of the entries (412 instances on the current PDB release). Owing to intrinsic complexity, also the instances with binding pocket located at the protein:protein interface or those with the ligand not in direct contact with protein atoms (e.g. in the presence of a cluster of ligands) are not included in the database. The current database includes 9372 distinct ligands, of which 574 classified as Drugs according to KEGG and DrugBank (Kaneisha *et al.*, 2012; Knox *et al.*, 2001). The data acquisition and processing, repeated for each ligand, consist in four steps. In the first step, the PDB entries associated with the specific ligand are sorted in homologous families (Fig. 1a). In the second step, for each PDB entry, the amino acids in interaction with the bound ligand are deduced from the LigPlot output of PDBsum (Wallace *et al.*, 1995). The third step is carried out on each group of homologous proteins identified in step one: the pair-wise sequence alignments deduced from the SAS section of PDBsum are used to identify the structurally equivalent amino acids of the binding pockets (Fig. 1b). The fourth step consists of computing, for each pair of homologous protein structures, the degree of conservation of the binding pocket residues. This value is then compared with the overall amino acid conservation to identify putative evolutionary constraints involving binding pocket residues (Fig. 1c; see supplementary materials for details).

*To whom correspondence should be addressed.

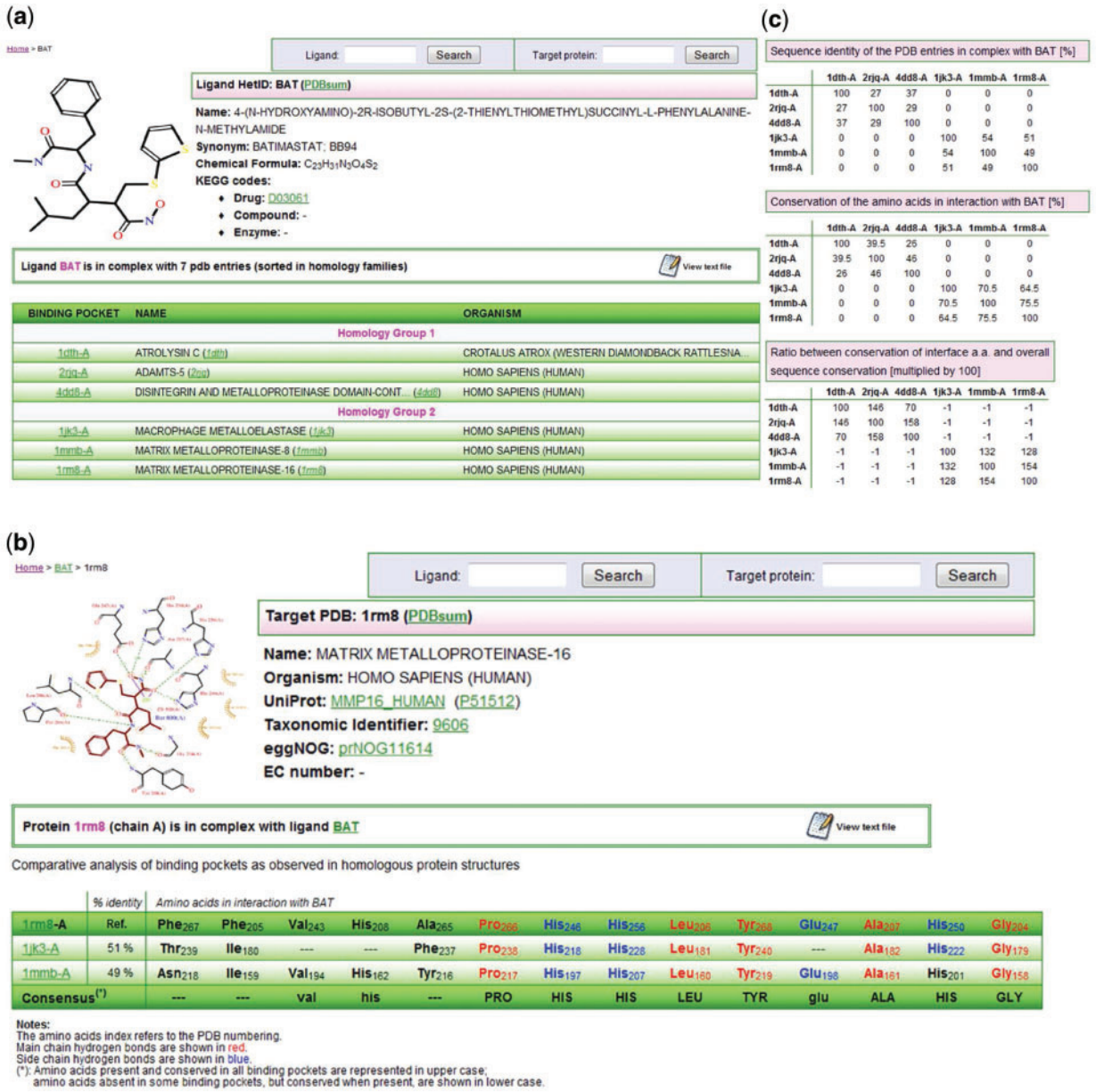

Fig. 1. Example of web pages describing the ligand Batimastat (Het_id: BAT). (**a**) PDB entries in which this ligand is found in complex with protein chains. (**b**) Binding pocket of Batimastat with Matrix Metalloproteinase-16 (PDB: 1rm8, chain A) and comparison of the binding pockets of the homologous proteins 1jk3 and 1mmb. (**c**) Matrices describing the sequence identity (top), the conservation of the binding pocket (middle) and the conservation index (bottom), defined as described in the supplementary materials

3 RESULTS

Drug design and lead optimization often rely on information obtained by structural biology methods; this information may integrate that obtained with other approaches such as those used to generate pharmacophore models (e.g. Ortuso *et al.*, 2006). In spite of the abundant, and rapidly growing, structural information in the PDB describing the interaction of ligands with distinct target proteins, there is a substantial lack of tools for the comparison of the binding pockets of homologous proteins not requiring the use of programs of structural superposition and significant expertise in the field of structural biology. Furthermore, the structural superposition and comparison becomes rapidly time-consuming if the number of homologous protein-ligand complexes exceeds four. The PLI web service provides and aids to carry out this comparison (see Fig. 1).

ACKNOWLEDGEMENTS

D.B. is grateful to Peer Bork and Michael Kuhn for many fruitful discussions and suggestions.

Funding: Italian Ministry of Health and the Regione Liguria, grant 'Identification of tumor biomarkers through a biology-driven integrated approach'.

Conflict of Interest: none declared.

REFERENCES

Dutta,S. *et al.* (2009) Data deposition and annotation at the worldwide Protein Data Bank. *Mol. Biotechnol.*, **42**, 1–13.

Laskowski,R.A. *et al.* (2005) PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res.*, **33**, D266–D268.

Ortuso,F. *et al.* (2006) GBPM: GRID-based pharmacophore model: concept and application studies in protein-protein recognition. *Bioinformatics*, **22**, 1449–1455.

Wallace,A.C. *et al.* (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.*, **8**, 127–134.

Kaneisha,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular sets. *Nucleic Acid Res.*, **40**, D109–D114.

Knox,C. *et al.* (2001) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acid Res.*, **39**, D1035–D1041.