Bioinformatics Advance Access published May 21, 2013

Structural Bioinformatics

TiPs: A database of therapeutic targets in pathogens and associated tools

Rosalba Lepore[1], Anna Tramontano[1,2,3,*] and Allegra Via[1,*]

[1]Department of Physics, [2]Center for Life Nano Science @Sapienza, Istituto Italiano di Tecnologia and [3]Istituto Pasteur, Fondazione Cenci Bolognetti, Sapienza University, 00185 Rome, Italy

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Motivation: The need for new drugs and new targets is particularly compelling in an era that is witnessing an alarming increase of drug resistance in human pathogens. The identification of new targets of known drugs is a promising approach, which has proven successful in several cases. Here, we describe a database that includes information on 5153 putative drug-target pairs for 150 human pathogens derived from available drug-target crystallographic complexes.

Availability and implementation: The TiPs Database is freely available at http://biocomputing.it/tips

Contact: anna.tramontano@uniroma1.it, allegra.via@uniroma1.it

1 INTRODUCTION

Novel mechanisms to escape therapy are constantly emerging among human pathogen populations and this clearly urges the development, on one hand, of new drugs for the treatment of the diseases and, on the other hand, of rapid and effective methods to help expanding the landscape of available treatment options (Hopkins, et al., 2011). In this context, computational studies are called upon to help identify novel therapeutic targets and characterise their interactions, and indeed a number of such efforts are described in the literature (Aguero, et al., 2008; Kinnings, et al., 2010; Lepore, et al., 2011; Orti, et al., 2009). However, these are mostly devoted to the analysis of single targets or specific tropical disease pathogens.

The TiPs Database has been developed with the aim of facilitating the identification of new therapeutic targets in more than 150 organisms responsible for human infections. We performed a large-scale analysis to systematically identify candidate targets in the proteomes of such organisms. The rationale of our approach is based on the intrinsic polypharmacological behaviour of compounds targeting homologous proteins (Paolini, et al., 2006). We considered all drug-target pairs for which the three-dimensional (3D) structure of the complex is experimentally known and used the sequence of the target to identify its homologues in human pathogens. The evolutionary conservation of such homologues and their 3D structures (available or predicted) were used to verify whether the original drug was in principle able to bind them as it does the original target. To this aim, stringent filters were applied to ensure that predicted binding sites and their interactions with the drug are as accurate as possible. Pathogen proteins predicted with high confidence to be therapeutic targets and the putative drugs interacting with them were collected and annotated in TiPs.

2 METHODS

More than 400 human pathogen species were obtained from "The Approved List of Biological Agents" provided by the Advisory Committee on Dangerous Pathogens. In order to unambiguously assign an identifier (ID) to human pathogens, the names of the organisms were mapped onto the NCBI Taxonomy Database records (http://www.ncbi.nlm.nih.gov/Taxonomy/).

Drug compounds and information on their molecular targets were obtained from DrugBank (http://www.drugbank.ca). The SMILE IDs of drugs annotated either as "inhibitor", "agonist" or "antagonist" were used to associate them with ligands present in the PDB structure entries (Berman, et al., 2012). Only identical compounds were considered (Tanimoto coefficient = 1). A total of 308 distinct drugs were observed in complex with at least one PDB structure. About 40% of these (119/308) occur in complex with their actual pharmaceutical target. These were used as starting points to predict potential drug targets in pathogens. The search for homologues in pathogens was performed using BLAST+ (Camacho, et al., 2009) with default parameters against the nr database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/). We only retained highly reliable hits, i.e. those showing at least 40% sequence identity to the original target and e-value<10^{-6}. Pathogen taxonomic IDs were retrieved by matching the gi numbers of BLAST hits to the NCBI Taxonomy database.

For each known drug-target complex, we defined the binding site as the subset of target residues having at least one atom within 3.5 Å distance from any atom of the drug. The drug binding site residues in the predicted pathogen sequences were retrieved through a multiple sequence alignment (MSA) of the original target sequence with its homologues generated with T-coffee (Taly, et al., 2011). The number and type of aligned residues were used to classify the binding site local conservation, both in terms of sequence coverage (percentage of binding site residues in the original target that could be aligned to the pathogen sequence) and identity (percentage of identical residues among the aligned binding site residues). Coverage and identity percentages were calculated separately for each pathogen sequence in the alignment. Only pathogen proteins showing at least 80% coverage in their binding sites were further considered (4215 in total). Among these 4215 reliable putative targets, only 41 have a solved structure in the PDB. Homology modelling

*To whom correspondence should be addressed.

(Kopp and Schwede, 2004) was used to predict the structure of the remaining ones as follows: For each pathogen sequence, an MSA was generated using three iterations of HHblits (Remmert, et al., 2012) (with default parameters) on the non-redundant Uniprot database. The MSA was used as HHsearch query to search for templates in the PDB70 database. We only selected templates with at least 40% sequence identity (and e-value lower than 10^{-5}) with the pathogen query sequence. If more than one template was found, the one with the highest coverage to the pathogen sequence was selected. Models were generated using the Modeller software. Note that the best template used to build the model corresponds to the original structure in the drug-target complex only in 153 cases, in all the others the best template was a different structure.

The binding site residues of the original complex and of the predicted target were structurally superimposed using the LGA software (Zemla, 2003). Subsequently, the ligands were transferred into the structure or model of the pathogen proteins that could be successfully superimposed under 5 Å distance to the known target. Binding sites in the modelled structures were analysed for the occurrence of nearby insertions/deletions. These cases are suitably highlighted in the TiPs database search output. This allows users to analyse them to establish the likelihood that their presence affects the conformation of the binding site.

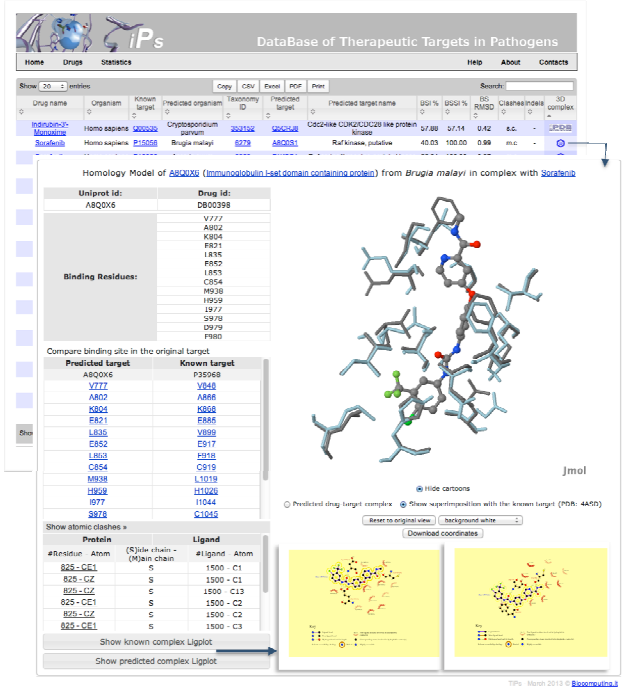

Figure 1 - The figure shows the results of a "all pathogens" filtered by the "ATP binding" GO term query in the TiPs database. The output table lists all putative pathogen targets. Each table row reports the known and predicted target UniProt IDs, their overall sequence identity, their binding site identity and rmsd, whether or not there are clashes between the known drug and the predicted target, and whether there are insertions or deletions nearby the binding site in the alignment used to model the protein. For each hit, the system also shows details of the structure(s) and the binding site(s) in a Jmol window and the corresponding Ligplot drawings.

3 RESULTS

TiPs currently contains 4071 candidate pathogen target structures involved in 5153 different drug-target complexes in 150 pathogens. All entries are thoroughly annotated with both sequence and functional information. The database can be queried by organism name (genus or specie name), protein family or function (EC number, GO terms, Pfam), as well as UniProt ID. The query returns a sortable table providing information about both known and predicted drug-target pairs and links to visualise specific information on the drug(s) (physicochemical properties, structure, indication, side effects), the target(s) (UniProt annotation and PDB structure(s)) and to visually analyse or download their 3D complexes. Ligplot (Laskowski and Swindells, 2011) drawings of both the known and inferred binding sites in complex with the drug are available as well (Figure 1).

ACKNOWLEDGEMENTS

We are grateful to all member of the group for useful suggestions.
Funding: King Abdullah University of Science and Technology (KAUST), award number KUK-I1-012-43, PRIN 20108XYHJS and FIRB RBIN06E9Z8_005.

REFERENCES

Aguero, F., *et al.* (2008) Genomic-scale prioritization of drug targets: the TDR Targets database, *Nature reviews. Drug discovery*, **7**, 900-907.

Berman, H.M., *et al.* (2012) The Protein Data Bank at 40: reflecting on the past to prepare for the future, *Structure*, **20**, 391-396.

Camacho, C., *et al.* (2009) BLAST+: architecture and applications, *BMC bioinformatics*, **10**, 421.

Hopkins, A.L., *et al.* (2011) Rapid analysis of pharmacology for infectious diseases, *Current topics in medicinal chemistry*, **11**, 1292-1300.

Kinnings, S.L., *et al.* (2010) The Mycobacterium tuberculosis drugome and its polypharmacological implications, *PLoS computational biology*, **6**, e1000976.

Kopp, J. and Schwede, T. (2004) Automated protein structure homology modeling: a progress report, *Pharmacogenomics*, **5**, 405-416.

Laskowski, R.A. and Swindells, M.B. (2011) LigPlot+: multiple ligand-protein interaction diagrams for drug discovery, *Journal of chemical information and modeling*, **51**, 2778-2786.

Lepore, R., *et al.* (2011) Identification of the Schistosoma mansoni molecular target for the antimalarial drug artemether, *Journal of chemical information and modeling*, **51**, 3005-3016.

Orti, L., *et al.* (2009) A kernel for open source drug discovery in tropical diseases, *PLoS neglected tropical diseases*, **3**, e418.

Paolini, G.V., *et al.* (2006) Global mapping of pharmacological space, *Nature biotechnology*, **24**, 805-815.

Remmert, M., *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment, *Nature methods*, **9**, 173-175.

Taly, J.F., *et al.* (2011) Using the T-Coffee package to build multiple sequence alignments of protein, RNA, DNA sequences and 3D structures, *Nature protocols*, **6**, 1669-1682.

Zemla, A. (2003) LGA: A method for finding 3D similarities in protein structures, *Nucleic acids research*, **31**, 3370-3374.