Target-Pathogen Methodology

All data present in Target Pathogen database is based either on the *in-silico* calculation of selected properties for each protein or, on the integration and meta-analysis of publicly available data. The pipeline-engine which we call Target Pathogen is schematically shown below.

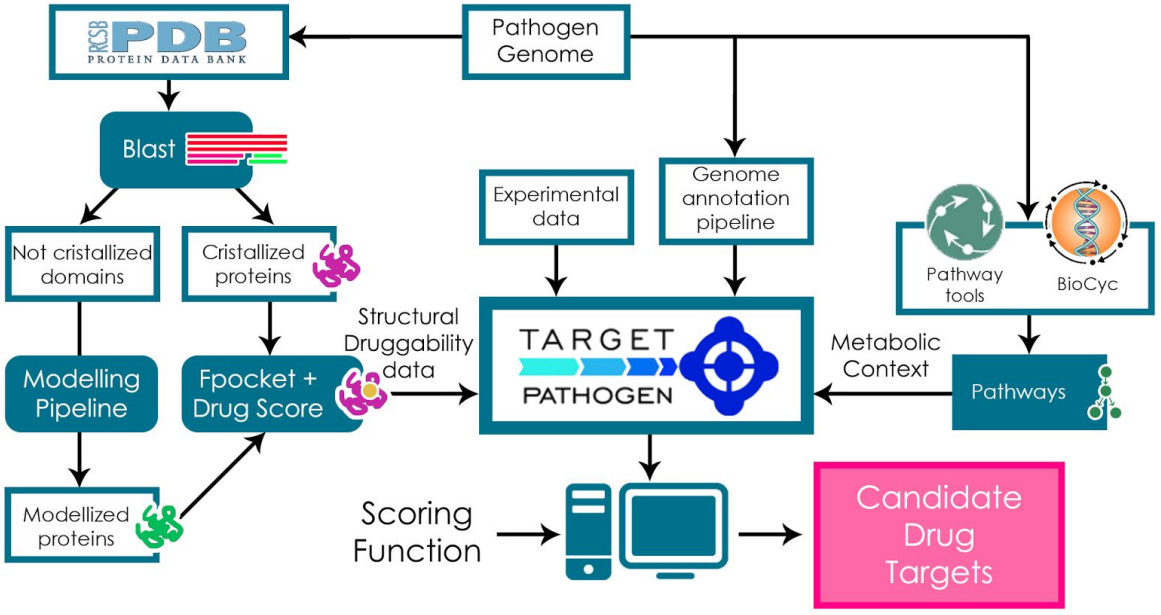

Figure 1. A general sketch of the pipeline. Structural druggability and metabolic analyses are integrated with available experimental and in silico data. After all data is integrated in Target-Pathogen, an user designed scoring function is used to weight differents features in order to obtain a ranked list of candidate drug targets

Generation of Structural homology based models

For all ORFs in *Mycobacterium tuberculosis, Klebsiella pneumoniae* and Shigella dysenteriae genomes we attempted to build homology-based models using the following structural genomic pipeline. The first step consists in performing a psi-blast search against a template library, which includes all sequences from every individual protein chain in the PDB, grouped at 95% sequence identity threshold using CD-hit (Li and Godzik 2006). Then, every target structure was built with the MODELLER software (Eswar et al. 2008), using local alignment derived from the above-described psi-blast search (Altschul et al. 1997). For each target sequence, 5 different models were built, and their quality measures were assigned using the GA341 (Melo and Sali 2007) and QMEAN (Melo and Sali 2007; Benkert,

Tosatto, and Schomburg 2008) methods. The best model (max QMean Score) for each protein was kept. Previosuly, only the models with GA341 score above 0.7 and over 60% coverage were retained. Other models were obtained from Modbase database (Pieper et al. 2014).

For or all the structures (crystals and models) we then compute several structural properties like: i) the DS for each pocket using fpocket, ii) the similarity with human protein (to evaluate potential off-target effects), iii) The active site residues (if available) by using data from Catalytic Site Atlas (CSA) (Furnham et al. 2013) and iv) The PFAM conserved or family relevant residues.

Structural Assessment of Druggability

Structural druggability of each potential target was assessed by determining (and characterizing) the ability of putative pockets to bind a drug-like molecule by using the fpocket program (Schmidtke et al. 2010) and DrugScore (DS) index (Schmidtke et al. 2010; Schmidtke and Barril 2010). Briefly, the method is based on Voronoi tessellation algorithm to identify pockets and computes suitable physicochemical descriptors (hydrophobic density, polar and apolar surface area, hydrophobic and polarity score) that are combined to yield the DS, which ranges between 0 to 1. Figure 2 shows a histogram for the druggability score computed for those pockets present in all unique protein in the Protein Data Bank, which were crystallized in complex with a drug like compound that correspond effectively to the binding pocket is shown:

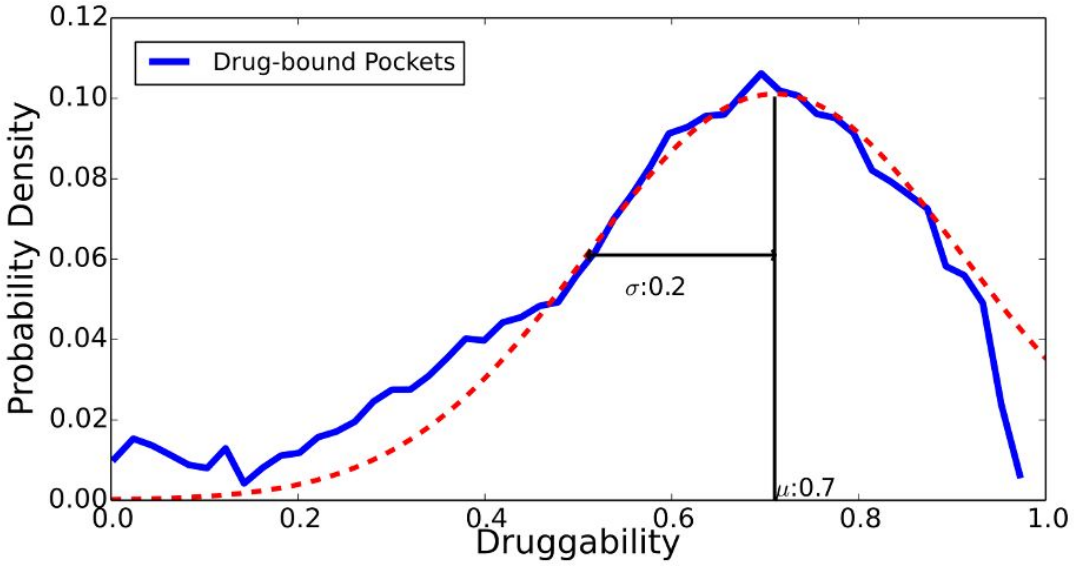

Figure 2: In blue, histogram for the druggability score computed for all those pockets present in all unique protein in the Protein Data Bank, which were crystallized in complex with a drug like compound inside the corresponding pocket. In red, the gaussian fitting of the pocket classification sets.

Fitting the resulting histogram to a gaussian distribution results in a mean of 0.7 with a standard deviation of 0.2. As expected the DS computed for all pockets in the PDB, except those having inside a drug like molecule, show a distribution that peaks at DS of zero, and falls rapidly (See Green plot in Figure 3)

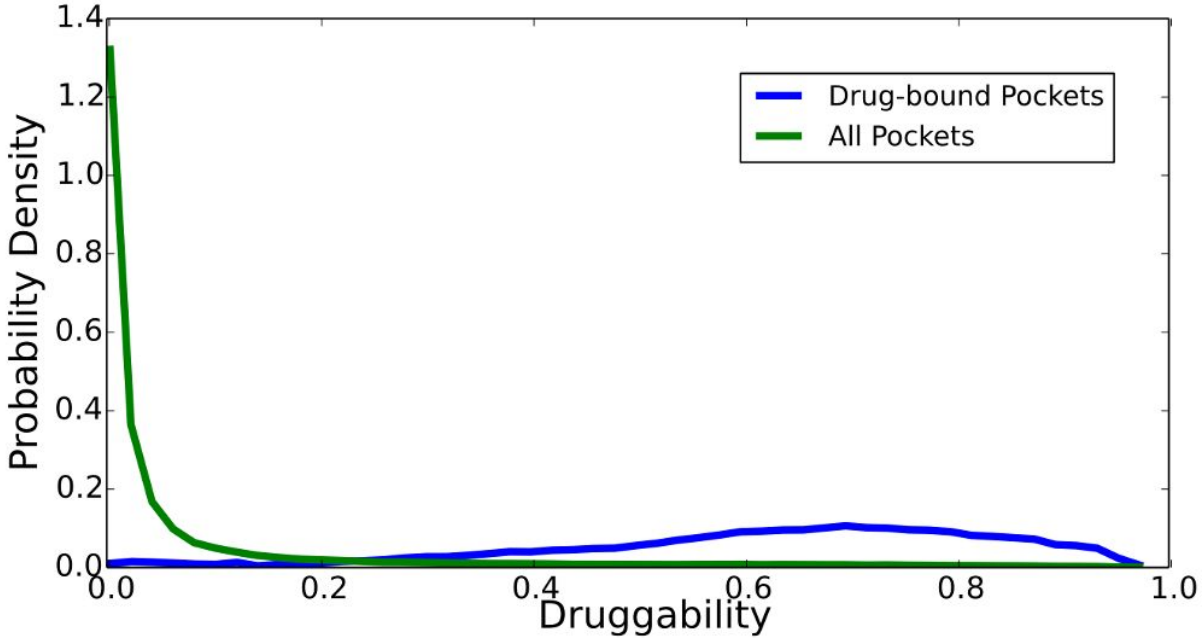

Figure 3: In green, the druggability score computed by fpocket software for all the pockets in all the structures of the PDB database except those having inside a drug like molecule which are shown in blue.

Based on this analysis we classified each pocket according to the following four categories, non druggable (ND), with DS between 0 and 0.2, poorly druggable (PD), with DS between 0.2 and 0.5, Druggable (D) with DS between 0.5 and 0.7, and highly druggable (HD) with DS between 0.7 and 1. In the web application, visualization is available for the first three categories but statistics are available for all the detected pockets regardless of its druggability score.

Active site pocket identification

To identify the active site pocket and/or determine the relevance of a given pocket to protein function, Target-Pathogen uses TuberQ `s methodology (Radusky et al. 2014). It consist in two different analyses, that rely on, (i) the information from the CSA (Catalytic Site Atlas) and (ii) a PFAM position site importance criteria

The data from CSA (downloaded from http://www.ebi.ac.uk/thornton-srv/databases/CSA/) consists of a list of PDB linked to a number of residues, which comprise the corresponding protein active site. To map the active sites to as many protein and/or domains as possible, each PDB in CSA was assigned to a protein with that PDB hit.

As an alternative approach to determine the relevance of a given pocket (or residue), we looked for residues of a given PFAM family/domain that are located in an important position and are well conserved. Important positions were defined as those positions in the corresponding HMMer model whose information content was larger than a defined importance cutoff value (*icov*). The nature of the conserved amino acids in the corresponding position was determined by comparing each residue type emission probability (*ep*) with *icov*. If the ratio between *ep* and *icov* was larger than a conserved type cutoff value (*ctcov*), the corresponding residue type was assumed to be conserved. Optimal values of *icov* and *ctcov* were 0.27 and 0.24, respectively.

Off-target and essentiality criteria

All proteins in the database were subjected to NCBI-BLASTp (*e-value* smaller than 1e-07) against human proteome to identify non-host homologs targets. The criteria for regarding a protein as a human homologue were a sequence similarity of greater than 30% using a BLOSUM62 matrix, for a length of more than 30% of the bacterial query protein with an E-value less than 10-4.

Furthermore, all proteomes were submitted to the Database of Essential Genes (Zhang 2004a; Luo et al. 2014)(DEG, which contains experimentally validated genes under different conditions in three domains of life) for homology analyses (Barh et al. 2013; Zhang 2004b) . The BLASTp cut-off values used were: *e-value* = 1e-05, *bit score* ≥100, *identity* ≥ 35% (Barh et al. 2011). Only Mycobacterium tuberculosis essential genes were defined as in previous works (Defelipe et al. 2016; Radusky et al. 2014)

Metabolic network construction

Metabolic networks (MN) were built by using the PathoLogic algorithm within Pathway Tools v. 19.0 (Karp et al. 2016) (or a previously existing PGDB was used when it was available) . PathoLogic creates a Pathway/Genome Database (PGDB) containing the predicted metabolic pathways of a given organism using as input a Genbank file with the corresponding product annotations and gene coordinates along the genome. The Genbanks entries were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/) and were used as initial input for MN reconstructions. The steps involved in the reconstruction include determining gene-protein-reaction associations, which are based in either the availability of the

corresponding enzyme commission (EC) number or alternatively in the gene product annotation, using a custom dictionary within Pathway Tools which links products to reactions. The reconstructed metabolic network was exported in systems biology markup language (SBML) and format for downstream analyses. Reactions involving macromolecules (such as DNA, RNA and proteins, as per the BioCyc ontology) were filtered, and only the small-molecule complement of the MNs was considered. After MN reconstruction, we generated a list of all compounds present in the network, and we collected their frequency as reaction participants using a Python script. Those who most frequently appeared as reaction participants are considered currency compounds (such as ATP, cofactors, water) and were disregarded from the network since they may create artificial links on the graph-based representation of the network as they are involved in many reactions which are not necessarily related.

Metabolic network analysis.

After MN reconstruction, we generated a reaction graph, where nodes represent reactions (i.e usually enzymes) and there is an edge between two nodes if the product of one reaction is used as substrate on the reaction that follows. Cytoscape v. 2.8.3 was used for data visualization and further MN analyses (Karp et al. 2016; Russell and Cohn 2012). Choke-point analysis was conducted in order to identify potential drug targets from the metabolic perspective. We also calculated the betweenness centrality of every node in MN, using the betweenness_centrality function in the NetworkX python package. The betweenness centrality of a given node v, $C_B(v)$, in the graph $G = (V, E)$, where V is a set of vertices or nodes and E a set of edges is given by: $C_B(v) = \sum s \neq t \neq v \in V \frac{\sigma_{st}(v)}{\sigma_{st}}$ (Brandes 2001) where σ_{st} is the number of shortest paths from $s \in V$ to $t \in V$ and $\sigma_{st}(v)$ represents the number of shortest paths from s to t that some node $v \in V$ lies in.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. “Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs.” *Nucleic Acids Research* 25 (17): 3389–3402.

Barh, Debmalya, Krishnakant Gupta, Neha Jain, Gourav Khatri, Nidia León-Sicairos, Adrian Canizalez-Roman, Sandeep Tiwari, et al. 2013. “Conserved Host–pathogen PPIs : Globally Conserved Inter-Species Bacterial PPIs Based Conserved Host-Pathogen Interactome Derived Novel Target in C. Pseudotuberculosis, C. Diphtheriae, M. Tuberculosis, C. Ulcerans, Y. Pestis, and E. Coli Targeted by Piper Betel Compounds.” *Integrative Biology* 5 (3): 495.

Barh, Debmalya, Neha Jain, Sandeep Tiwari, Bibhu Prasad Parida, Vivian D’Afonseca, Liwei Li, Amjad Ali, et al. 2011. “A Novel Comparative Genomics Analysis for Common Drug and Vaccine Targets in Corynebacterium Pseudotuberculosis and Other CMN Group of Human Pathogens.” *Chemical Biology & Drug Design* 78 (1): 73–84.

Benkert, Pascal, Silvio C. E. Tosatto, and Dietmar Schomburg. 2008. “QMEAN: A Comprehensive Scoring Function for Model Quality Assessment.” *Proteins* 71 (1): 261–77.

Brandes, Ulrik. 2001. “A Faster Algorithm for Betweenness Centrality*.” *The Journal of Mathematical Sociology* 25 (2): 163–77.

Defelipe, Lucas A., Dario Fernández Do Porto, Pablo Ivan Pereira Ramos, Marisa Fabiana Nicolás, Ezequiel Sosa, Leandro Radusky, Esteban Lanzarotti, Adrián G. Turjanski, and Marcelo A. Marti. 2016. “A Whole Genome Bioinformatic Approach to Determine Potential Latent Phase Specific Targets in Mycobacterium Tuberculosis.” *Tuberculosis* 97 (March): 181–92.

Eswar, Narayanan, David Eramian, Ben Webb, Min-Yi Shen, and Andrej Sali. 2008. “Protein Structure Modeling with MODELLER.” In *Methods in Molecular Biology*, 145–59.

Furnham, Nicholas, Gemma L. Holliday, Tjaart A. P. de Beer, Julius O. B. Jacobsen, William R. Pearson, and Janet M. Thornton. 2013. “The Catalytic Site Atlas 2.0: Cataloging Catalytic Sites and Residues Identified in Enzymes.” *Nucleic Acids Research* 42 (D1): D485–89.

Karp, Peter D., Mario Latendresse, Suzanne M. Paley, Markus Krummenacker, Quang D. Ong, Richard Billington, Anamika Kothari, et al. 2016. “Pathway Tools Version 19.0 Update: Software for Pathway/genome Informatics and Systems Biology.” *Briefings in Bioinformatics* 17 (5): 877–90.

Li, W., and A. Godzik. 2006. “Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences.” *Bioinformatics* 22 (13): 1658–59.

Luo, Hao, Yan Lin, Feng Gao, Chun-Ting Zhang, and Ren Zhang. 2014. “DEG 10, an Update of the Database of Essential Genes That Includes Both Protein-Coding Genes and Noncoding Genomic Elements.” *Nucleic Acids Research* 42 (Database issue): D574–80.

Melo, Francisco, and Andrej Sali. 2007. “Fold Assessment for Comparative Protein Structure Modeling.” *Protein Science: A Publication of the Protein Society* 16 (11): 2412–26.

Pieper, Ursula, Benjamin M. Webb, Guang Qiang Dong, Dina Schneidman-Duhovny, Hao Fan, Seung Joong Kim, Natalia Khuri, et al. 2014. “ModBase, a Database of Annotated Comparative Protein Structure Models and Associated Resources.” *Nucleic Acids Research* 42 (Database issue): D336–46.

Radusky, Leandro, Lucas A. Defelipe, Esteban Lanzarotti, Javier Luque, Xavier Barril, Marcelo A. Marti, and Adrián G. Turjanski. 2014. “TuberQ: A Mycobacterium Tuberculosis Protein Druggability Database.” *Database: The Journal of Biological Databases and Curation* 2014 (0): bau035.

Russell, Jesse, and Ronald Cohn. 2012. *Cytoscape*. Book on Demand Limited.

Schmidtke, Peter, and Xavier Barril. 2010. “Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites.” *Journal of Medicinal Chemistry* 53 (15): 5858–67.

Schmidtke, Peter, Vincent Le Guilloux, Julien Maupetit, and Pierre Tufféry. 2010. “Fpocket: Online Tools for Protein Ensemble Pocket Detection and Tracking.” *Nucleic Acids Research* 38 (Web Server issue): W582–89.

Zhang, R. 2004a. “DEG: A Database of Essential Genes.” *Nucleic Acids Research* 32 (90001): 271D – 272.

———. 2004b. “DEG: A Database of Essential Genes.” *Nucleic Acids Research* 32 (90001): 271D – 272.