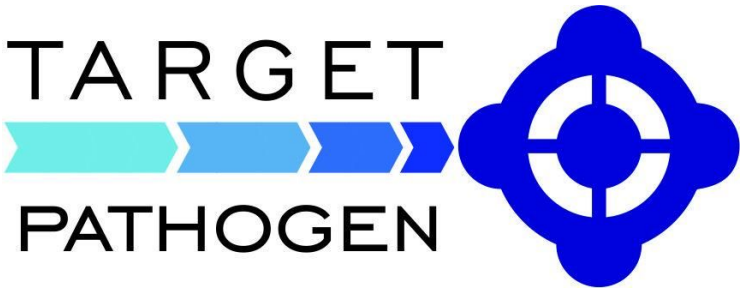


A Structural bioinformatic approach to prioritize drug targets in pathogens.

Tutorial

Introduction

The goal of Target-Pathogen is to become a useful resource for researchers working in the field of drug discovery to translate biological questions in a computational tractable way by exploring, filtering and weighting the vast quantity of genomic-scale data sets that are now available in order to produce a shortlist of suitable targets for further investigation. The main feature of Target-Pathogen is to integrate data from different sources with structural druggability analysis and metabolic network reconstruction in a consistent and effective manner, contributing to a better selection of potential drug targets for screening campaigns and the analysis of targets for structure-based drug design projects.

The general purpose of this tutorial is to show users how to explore data present in Target-Pathogen and how to weight this information in order to identify and prioritize drug targets for pathogens.

The following sections will guide users with examples to browse available genomic information, to obtain a ranked list of putative drug targets and to choose promising pathways from the drug discovery point of view.

Browsing the available genomic information in Target-Pathogen.

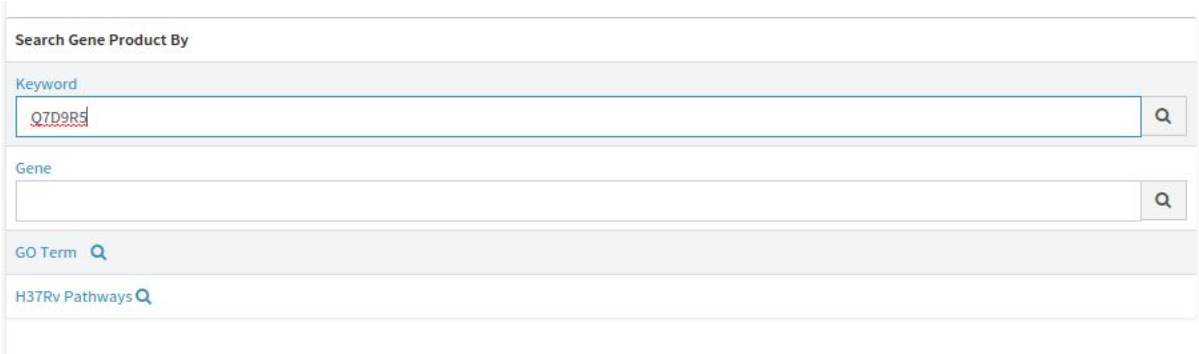
Target-Pathogen function as a database that allow users to rank and prioritize targets for drug development. A large amount of information of each gen and protein within genomes is actually present in the database. The genome browser can be accessed and queried using the web interface at http://target.sbg.qb.fcen.uba.ar/patho/. Here you have to choose one of the genomes already uploaded in Target-Pathogen by clicking Genomes. Suppose you are searching for a particular *Mycobacterium tuberculosis* protein, so you must select H37Rv genome.

The following example will take you through a trip around Target-Pathogen, showing its salient characteristics to search for a protein, that allows accession of the desired record in a fast and intuitive manner.

All searches start in the *main search page*, where you can use a keyword including Go terms, Uniprot ID (1), PFam ID (1, 2) or structure PDB ID (3) or you can specifically search a gene or pathway. Target-Pathogen also allows users to navigate the genome using

JBrowse. Searches may return a single database entry (e.g. when searching by Gene) or multiple entries (e.g. Keyword and pathways). Finally, genomes can be also easily explored hierarchically by EC number (4) or the different categories of Gene Ontology (5) by using Krona.

Let's assume, that in the present example, we already know our target protein ID, thus we simply type "Q7D9R5" in “Keyword”, to retrieve all associated records.



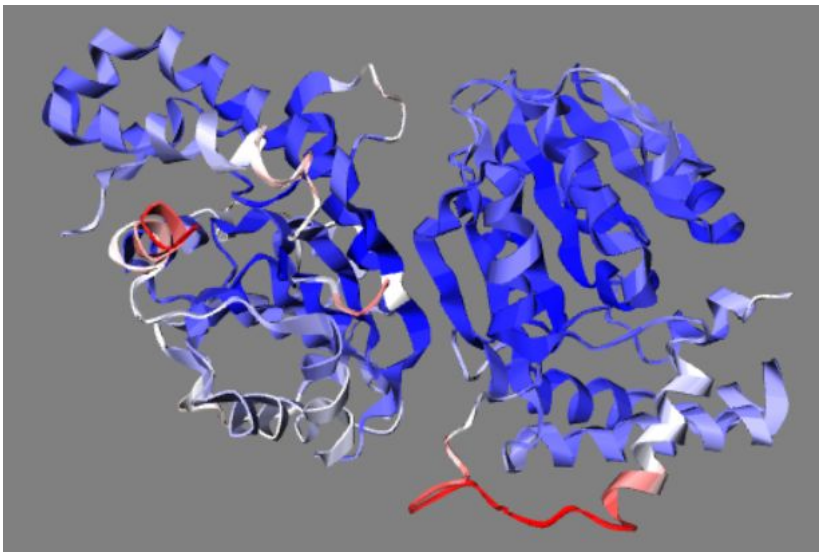
The resulting records are listed in the ranking page. For each record, size, druggability score, gene name and the number of pathways where the proteins are involved is presented.

Protein Product	Size	Druggability	Pathways	Gene	Description
Rv0470c	288 aa	0.551	1PW	Rv0470c pcaA	Cyclopropane mycolic acid synthase 3

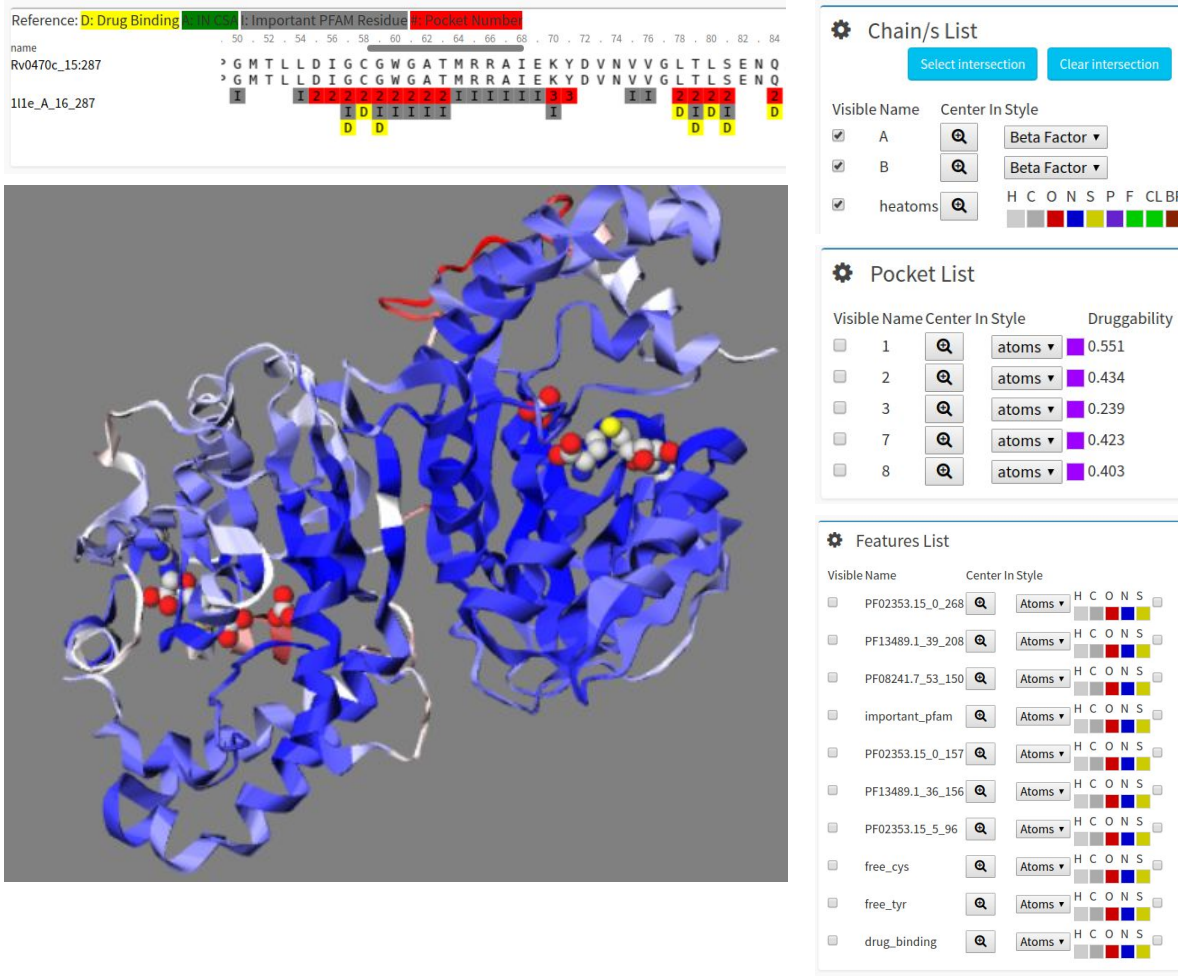
By clicking on the desired row, eight tabs of the corresponding record will be expanded. This tabs, will contain general information, metadata, protein's sequence, ontology and PFAM domains (2, 6), joined with structural and metabolic information.

In the present example, our protein of interest has been crystallized (PDB ID=1I1e). To access the record, click in the in 1I1e structure. Six main tabs will be display.

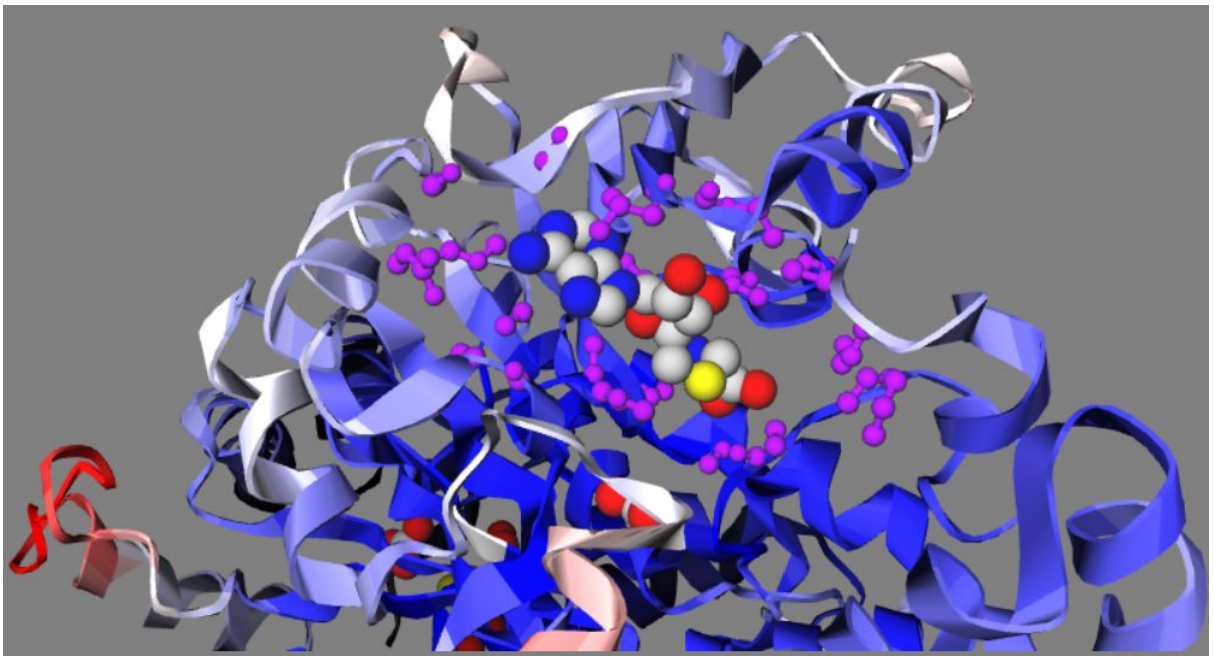
The top tab is where the visualization of the protein can be downloaded for VMD. Clicking in the download button a compressed file is download. This is the visualization for the protein in our example in the GLMol web software (http://webglmol.osdn.jp/index-en.html) used in the server:



Other tabs presents the structure related data, including the interactive pocket visualization module. The visualization module allows i) to select which pocket to show (ticking the corresponding pocket Select field), ii) display present HETATMS (7), assigned CSA (8) or PFAM relevant residues, iii) Display the protein chains in different styles, iv) Display the pocket residues or the alpha spheres (With polar and apolar spheres). In the druggable pocket of the example shown below, we depict polar alpha spheres of pocket "1" in black while its apolar alpha spheres in white. The HETATMS found in the crystal structure are shown as balls and sticks in different colours.



Another possible visualization of the same pocket, could be done by selecting to show the residues lining the pocket (not the alpha spheres) and the residues reported to be part of a domain in PFAM or a drug binding site. In the figure below superposition between the druggable pocket and one of the drug binding site of the crystallized protein is shown.



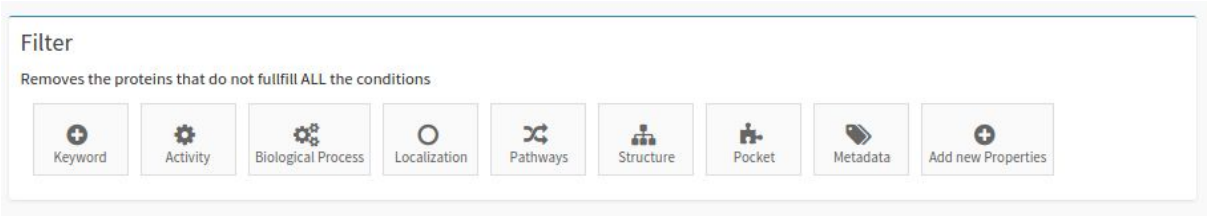
Obtaining a list of drug targets candidates for Mycobacterium tuberculosis.

We have previously defined two important features to select a gene product as a potential target for new drugs development to combat *Mycobacterium tuberculosis (9)*. First, the role of the protein within the metabolism and second, it's ability to bind a drug-like molecule, which in turn inhibits its function. Target-Pathogen allows users to interactively visualize genomic data in order to explore these criteria in genes and proteins of ten genomes now available in our web server.

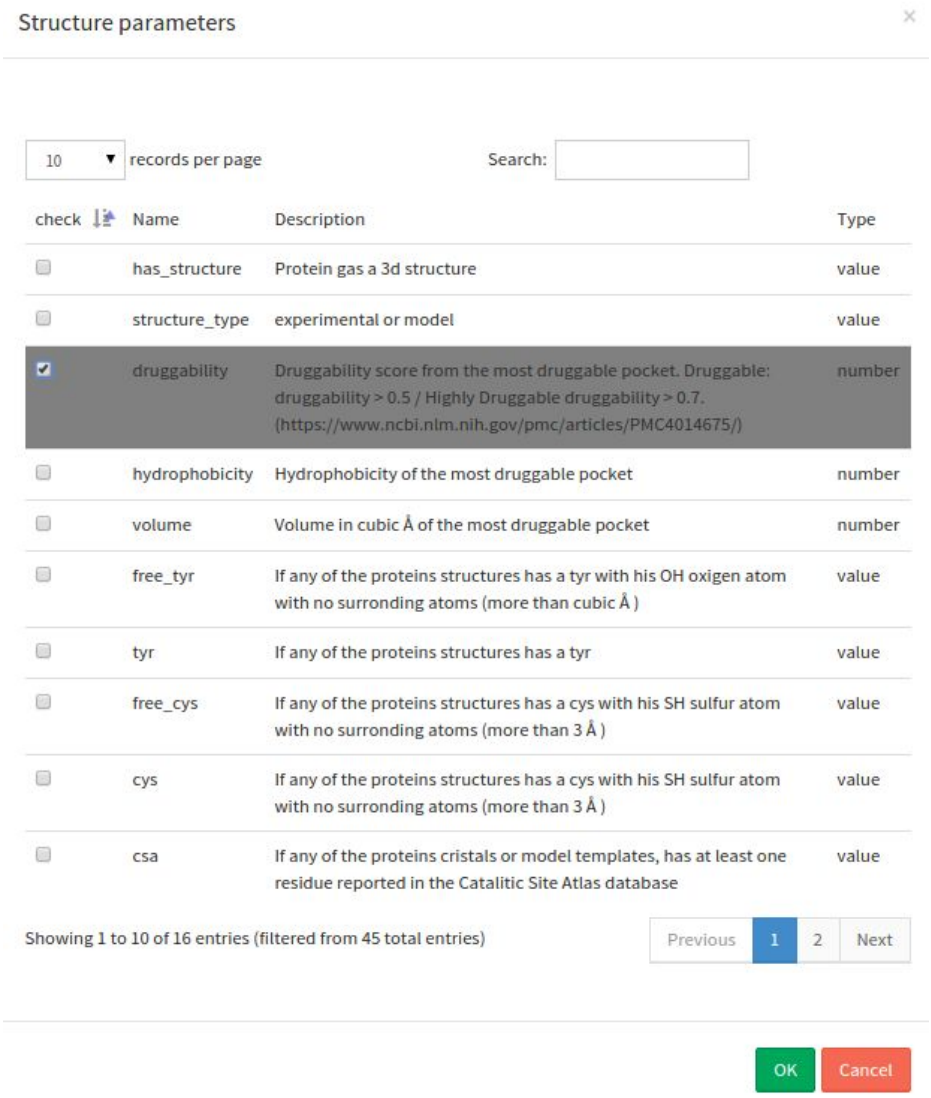
In this example we will guide you to obtain a ranked list of targets for drug development against latent *Mycobacterium tuberculosis.*

To obtain a short list of proteins that could be adequate candidates for drug targets you have to choose H37Rv genome and click “Prioritize Targets” in the protein column of the “Genomes” page. By doing that, you will be directed to a three tabs page, where you can filter and weight the data present in the database to display a set of proteins that fulfill the criteria defined by the user.

Once there, you can filter out all proteins without a druggable pocket and also present possible side effects with human host.



Just by a click in Structure, a window containing all parameters related with the protein structure will be open.



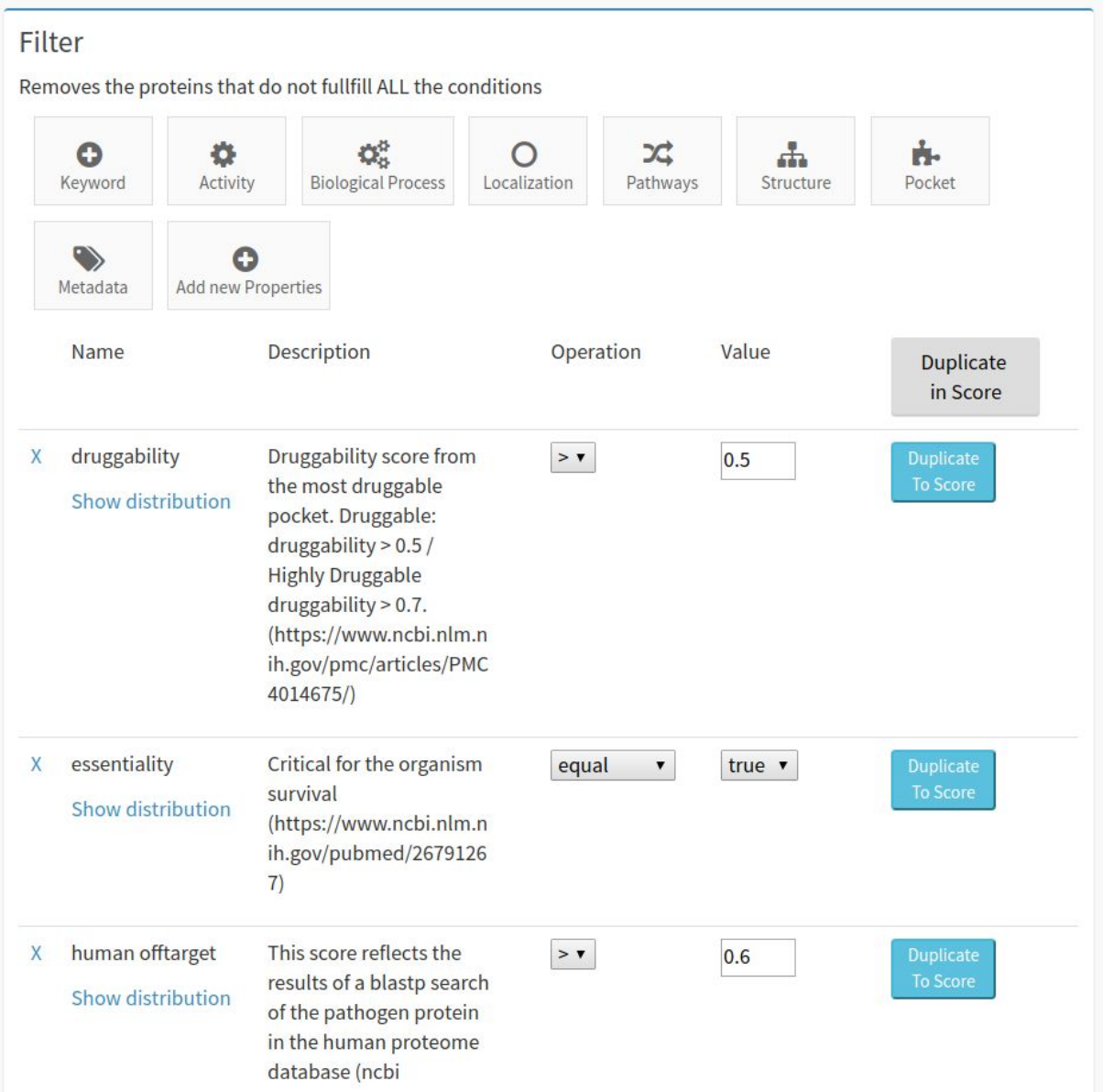
check	Name	Description	Type
☐	has_structure	Protein gas a 3d structure	value
☐	structure_type	experimental or model	value
☑	druggability	Druggability score from the most druggable pocket. Druggable: druggability > 0.5 / Highly Druggable druggability > 0.7. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4014675/)	number
☐	hydrophobicity	Hydrophobicity of the most druggable pocket	number
☐	volume	Volume in cubic Å of the most druggable pocket	number
☐	free_tyr	If any of the proteins structures has a tyr with his OH oxigen atom with no surronding atoms (more than cubic Å)	value
☐	tyr	If any of the proteins structures has a tyr	value
☐	free_cys	If any of the proteins structures has a cys with his SH sulfur atom with no surronding atoms (more than 3 Å)	value
☐	cys	If any of the proteins structures has a cys with his SH sulfur atom with no surronding atoms (more than 3 Å)	value
☐	csa	If any of the proteins cristals or model templates, has at least one residue reported in the Catalitic Site Atlas database	value

If you check druggability you will filter proteins by the druggability score (10). If you want druggable and highly druggable proteins you must keep all proteins with DS>0.5.

	Name	Description	Operation	Value
X	druggability Show distribution	Druggability score from the most druggable pocket. Druggable: druggability > 0.5 / Highly Druggable druggability > 0.7. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4014675/)	>	0.5

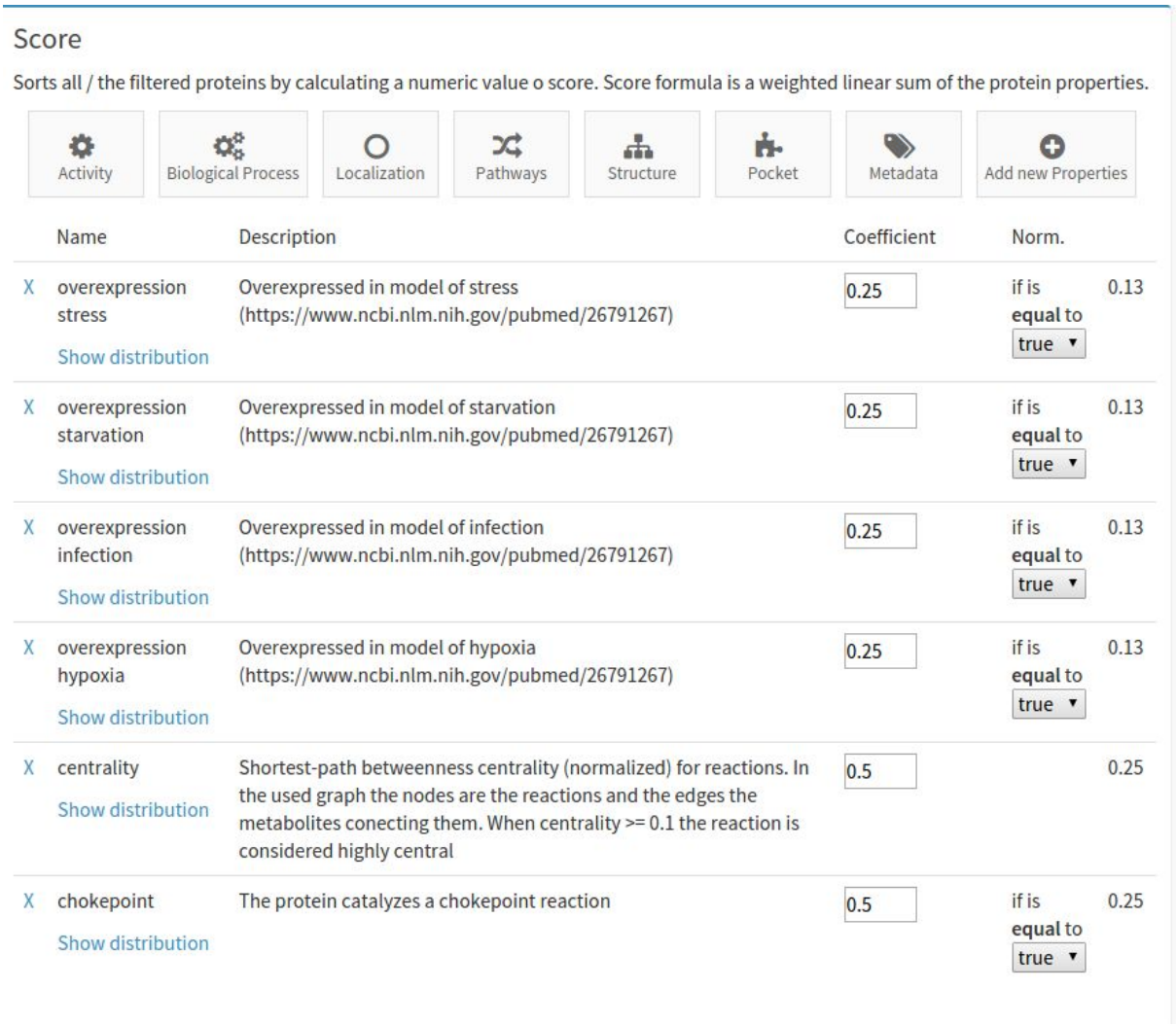
In a similar way, you can select “human_offtarget” in “Metadata” to filter out those protein with an human offtarget score > 0.6. By doing this you will retain 2047 records from a total of 4,023 proteins in *M tuberculosis* genome.

The lack or inhibition of an essential protein will conduce to inhibit growth or to death of the pathogens. So, a key criteria to select a good group of targets in tuberculosis is the essentiality of the proteins. To use these criteria, we can click in “Metadata” in the Filter-Tab and select “essentiality”. In this case gene essentiality was defined as in previous works (11) (10). Another criteria to select a good group of targets is their lack of a close homolog in humans to prevent side effects (human offtarget property in “Metadata”) . By doing this you will keep the 762 druggable, essential and without close human homologous that are actually annotated for *M. tuberculosis* genome.



To further rank the putative targets to specifically combat latent tuberculosis, we use the right panel (Score) to define a scoring function as:

$$SF = \frac{H+S+R+I}{4} + \frac{Ch+Cy}{2}$$



Score

Sorts all / the filtered proteins by calculating a numeric value o score. Score formula is a weighted linear sum of the protein properties.

Activity | Biological Process | Localization | Pathways | Structure | Pocket | Metadata | Add new Properties

	Name	Description	Coefficient	Norm.	
X	overexpression stress Show distribution	Overexpressed in model of stress (https://www.ncbi.nlm.nih.gov/pubmed/26791267)	0.25	if is **equal** to true	0.13
X	overexpression starvation Show distribution	Overexpressed in model of starvation (https://www.ncbi.nlm.nih.gov/pubmed/26791267)	0.25	if is **equal** to true	0.13
X	overexpression infection Show distribution	Overexpressed in model of infection (https://www.ncbi.nlm.nih.gov/pubmed/26791267)	0.25	if is **equal** to true	0.13
X	overexpression hypoxia Show distribution	Overexpressed in model of hypoxia (https://www.ncbi.nlm.nih.gov/pubmed/26791267)	0.25	if is **equal** to true	0.13
X	centrality Show distribution	Shortest-path betweenness centrality (normalized) for reactions. In the used graph the nodes are the reactions and the edges the metabolites conecting them. When centrality >= 0.1 the reaction is considered highly central	0.5		0.25
X	chokepoint Show distribution	The protein catalyzes a chokepoint reaction	0.5	if is **equal** to true	0.25

The first term of the equation integrates available expression data under different conditions mimicking infection. H, S, R, I are variables that defines overexpression in different experimental models: hypoxia, starvation, RNOS stress and mice infection models respectively (9). The second term focus in metabolic context of the proteins. In this way C_h and C_y determines if the reaction associated to the protein is a chokepoint or central in the bacteria metabolism. Note that expression and metabolic terms are divided by two and four respectively assigning the same weight to both components.

Each variable takes the value of 1 if the protein comply the criteria and 0 if not. A high value means that the protein fulfill most of the criteria that defines a promising drug target. The third tab (show below), at the bottom, is where it is displayed the proteins ranked by the

criteria previously set. You can easily download this table in csv format just by clicking on "Download List".

Showing 1 to 100 of 762 entries (filtered from 4,023 total entries)
Export first 100 to CSV

Refresh | Download list | gene... | description...

Protein Product	Size	Druggability	Pathways	Gene	Description	Properties	Score
Rv3206c	393 aa	0.985	1PW	Rv3206c moeB1	Probable adenylyltransferase/sulfurtransferase MoeZ	**overexpression_stress**: true, **overexpression_starvation**: true, **overexpression_infection**: false, **overexpression_hypoxia**: true, **centrality**: 0.02, **chokepoint**: true	1.26
Rv2245	417 aa	0.777	7PW	Rv2245 kasA	3-oxoacyl-[acyl-carrier-protein] synthase 1	**overexpression_stress**: True, **overexpression_starvation**: True, **overexpression_infection**: false, **overexpression_hypoxia**: True, **centrality**: 0.02, **chokepoint**: true	1.26
Rv1285	333 aa	0.652	2PW	Rv1285 cysD	Sulfate adenylyltransferase subunit 2	**overexpression_stress**: True, **overexpression_starvation**: True, **overexpression_infection**: false, **overexpression_hypoxia**: True, **centrality**: 0.01, **chokepoint**: true	1.25
Rv1286	615 aa	0.542	2PW	Rv1286 cysNC	Bifunctional enzyme CysN/CysC	**overexpression_stress**: True, **overexpression_starvation**: True, **overexpression_infection**: false, **overexpression_hypoxia**: True, **centrality**: 0.01, **chokepoint**: true	1.25
Rv2225	282 aa	0.768	1PW	Rv2225 panB	3-methyl-2-oxobutanoate hydroxymethyltransferase	**overexpression_stress**: True, **overexpression_starvation**: false, **overexpression_infection**: True, **overexpression_hypoxia**: True, **centrality**: 0.00, **chokepoint**: true	1.25

Uploading users data

A key feature that distinguish Target-Pathogen from other target prioritization software is that users can upload their own data in an easy way, just by simply uploading a tsv format archive (tab separated values). As an example we show a tsv archive with some antibiotics resistance related genes. Download example1 Download example2

id	resistance	AMI	PAS	EMB	FLQ	INH	SM	RIF	ETH	PZA
Rv0846c	no	no	no	no	no	no	no	no	no	no
Rv0203	no	no	no	no	no	no	no	no	no	no
Rv0343	yes	no	no	yes	no	yes	no	no	no	no
Rv0341	yes	no	no	yes	no	yes	no	no	no	no
Rv0342	yes	no	no	yes	no	yes	no	no	no	no
Rv0483	no	no	no	no	no	no	no	no	no	no
Rv1433	no	no	no	no	no	no	no	no	no	no
Rv1804c	no	no	no	no	no	no	no	no	no	no

The first column in the tsv must be the genes id of the corresponding genomes and must be named "id". Then you can add as many columns as you wish with different values that can be either numeric or strings. In this example each column represent first and second line antibiotics against *M tuberculosis*. For each combination of genes and antibiotics there is a "yes" if the gene has a genetic polymorphisms associated with the respective drug and a "no" if hasn't. To upload this data you should click "Add new properties" in the Filter or Score Tab. Once uploaded this data you can use it to filter or to calculate a new score.

Choosing promising pathways as putative targets of new drugs.

Numerous genomic sequencing projects have provided a nearly complete list of the components that are present in an organism, so post-genomic projects now focus on understanding metabolic and signaling networks, large multimeric complexes or even whole organisms. This emerging field of systems biology provides a key framework for understanding cellular metabolism under different conditions, facilitating the discovery of new drugs. Due to this, the reconstruction, through bioinformatic tools, of pathogens metabolic networks is key to explore possible molecular targets (proteins) of novel drugs. As said before, Target-Pathogen allow users to select and study proteins not only according to properties such as the essential role in the metabolism (essentiality) and / or feasibility of being inhibited (druggable), but also to its contextual role (contextuality) in metabolic pathways. Moreover, it allows you to rank pathways with a user-defined criteria in order to prioritize entire pathways as good candidates for novel therapies. One fundamental advantage of studying the metabolic context of putative targets is that results are expected to allow the design of possible combined therapies (targeting more than one target from the same metabolic pathway).

For example, if we want to determine which pathways are relevant for develop new therapies for polymyxin B-resistant *Klebsiella pneumoniae*, maybe we should be interested in a scoring function as defined in equation 2 in order to assign a score to each pathway:

$$SF = C_x + Chk + C_y + H + E + C_{Kp} + Pb$$

Where C_x = (*pathways.completeness*) is the ratio between the total number of reactions of a pathway associated with a gene and the total number of enzymatic reactions present in the pathway. Chk (*pathways.norm_chokepoint*) is the proportion of reactions that are actually chokepoints in the pathway. C_y (*pathways.max_centrality*) is the ratio between the node centrality and the node with with the biggest centrality in the entire metabolism. C_{Kp} reflects the presence of the different proteins belonging to pathway in pathogenic *Klebsiella pneumoniae*(*metadata.conserved_pathogen_norm*), H is where off-target criteria

(*metadata.human_offtarget*) analysis is taking place and E defines essentiality of the pathway (*metadata.hit_in_deg, metadata.essential in mgh78578*) and, at last, P *(overexpressed in polymixin)* is the ratio between the genes present in the pathway and the overexpressed genes in polymyxin B-induced transcriptomic response (12).

Score

Sorts all / the filtered proteins by calculating a numeric value o score. Score formula is a weighted linear sum of the protein properties.

Activity | Biological Process | Localization | Pathways | Structure | Pocket | Metadata | Add new Properties

	Name	Description	Coefficient	Group		Norm.
X	completeness Show distribution	Proportion of reactions in the pathway with at least one known protein that catalize them	1			0.13
X	max centrality Show distribution	Maximum betweenes centrality of all the reactions in the pathway, normalized by the reaction with max betweenes centrality in the whole graph	1			0.13
X	human offtarget Show distribution	This score reflects the results of a blastp search of the pathogen protein in the human proteome database (ncbi accession GCF_000001405.36) with the scale 1 - max(alignment identity), so when a protein has no hit in the human proteome, the value is 1, and if it has 2 hits, one with an identity of 0.4 and other with 0.6, the score is 0.4 (human_offtarget = 1 - 0.6, uses the max identity).	1	avg		0.13
X	hit in deg Show distribution	Has a hit in Database of Essential Genes	1	avg	if is **equal** to Yes	0.13
X	essential in mgh78578 Show distribution	Hits with an essential gene of Klebsiella pneumoniae MGH78578	1	avg	if is **equal** to Yes	0.13
X	overexpressed in polymyxin Show distribution	Overexpressed in polymyxin B resistance induction (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5088521/)	1	avg	if is **equal** to YES	0.13
X	conserved pathogen norm Show distribution	Hit count in different pathogenic Kp strains divided by the total number of compared bacteria (39 pathogenic Kp).	1	avg		0.13
X	norm chokepoint Show distribution	Chokepoint reactions/reactions in pathway ratio	1			0.13

Overall, a high value would mean that most genes in the pathway and the hole pathway itself fulfill most of the user defined criteria and therefore are attractive from the drug discovery point view. At last, we defined a pathway as druggable if at least one of the proteins involved is druggable and rule out non-druggable pathways (in the Filter, at the left part of the screen, property ” *druggable*” was added as a filter).

Filter

Removes the proteins that do not fullfill ALL the conditions

Pathways

	Name	Description	Operation	Value
X	druggable Show distribution	The pathway has at least one druggable protein	equal	Yes

Top five pathways are shown below.

-	Term	Name	Reactions	Reactions with gene	Genes	Properties	Score
1	PWY-7346	UDP-α-D-glucuronate biosynthesis (from UDP-glucose)	1	1	1	completeness = 1.00 , norm_chokepoint = 1.00 , max_centrality = 0.02 , hit_in_deg = 1.00 (avg), essential_in_mgh78578 = 1.00 (avg), overexpressed_in_polymyxin = 1.00 (avg), conserved_pathogen_norm = 0.97 (avg), human_offtarget = 0.71 (avg)	6.71
2	NAGLIPASYN-PWY	lipid IV_A biosynthesis	6	6	6	completeness = 1.00 , norm_chokepoint = 1.00 , max_centrality = 0.29 , hit_in_deg = 1.00 (avg), essential_in_mgh78578 = 0.83 (avg), overexpressed_in_polymyxin = 0.50 (avg), conserved_pathogen_norm = 0.83 (avg), human_offtarget = 1.00 (avg)	6.45
3	PWY0-1264	biotin-carboxyl carrier protein assembly	4	3	5	completeness = 0.75 , norm_chokepoint = 1.00 , max_centrality = 0.31 , hit_in_deg = 1.00 (avg), essential_in_mgh78578 = 0.75 (avg), overexpressed_in_polymyxin = 0.75 (avg), conserved_pathogen_norm = 0.99 (avg), human_offtarget = 0.75 (avg)	6.31
4	UDPNAGSYN-PWY	UDP-*N*-acetyl-D-glucosamine biosynthesis I	8	8	9	completeness = 1.00 , norm_chokepoint = 0.63 , max_centrality = 1.00 , hit_in_deg = 0.75 (avg), essential_in_mgh78578 = 0.50 (avg), overexpressed_in_polymyxin = 0.75 (avg), conserved_pathogen_norm = 1.00 (avg), human_offtarget = 0.62 (avg)	6.25
5	PWY-6387	UDP-*N*-acetylmuramoyl-pentapeptide biosynthesis I (*meso*-DAP-containing)	8	8	8	completeness = 1.00 , norm_chokepoint = 1.00 , max_centrality = 0.08 , hit_in_deg = 1.00 (avg), essential_in_mgh78578 = 1.00 (avg), overexpressed_in_polymyxin = 0.13 (avg), conserved_pathogen_norm = 0.99 (avg), human_offtarget = 1.00 (avg)	6.19

1. The UniProt Consortium (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.

2. Finn,R.D. (2005) Pfam: the protein families database. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*.

3. Burley,S.K., Berman,H.M., Kleywegt,G.J., Markley,J.L., Nakamura,H. and Velankar,S. (2017) Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods Mol. Biol.*, **1607**, 627–641.

4. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.

5. Gene Ontology Consortium and Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, 258D–261.

6. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.

7. Guruprasad,K., Savitha,S. and Babu,A.V.N. (2005) Computational tools for the analysis of heteroatom groups and their neighbours in protein tertiary structure. *Int. J. Biol. Macromol.*, **37**, 35–41.

8. Furnham,N., Holliday,G.L., de Beer,T.A.P., Jacobsen,J.O.B., Pearson,W.R. and Thornton,J.M. (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.*, **42**, D485–9.

9. Defelipe,L.A., Do Porto,D.F., Pereira Ramos,P.I., Nicolás,M.F., Sosa,E., Radusky,L., Lanzarotti,E., Turjanski,A.G. and Marti,M.A. (2016) A whole genome bioinformatic approach to determine potential latent phase specific targets in Mycobacterium tuberculosis. *Tuberculosis* , **97**, 181–192.

10. Radusky,L., Defelipe,L.A., Lanzarotti,E., Luque,J., Barril,X., Marti,M.A. and Turjanski,A.G. (2014) TuberQ: a Mycobacterium tuberculosis protein druggability database. *Database* , **2014**, bau035.

11. Defelipe,L.A., Do Porto,D.F., Pereira Ramos,P.I., Nicolás,M.F., Sosa,E., Radusky,L., Lanzarotti,E., Turjanski,A.G. and Marti,M.A. (2016) A whole genome bioinformatic

approach to determine potential latent phase specific targets in Mycobacterium tuberculosis. *Tuberculosis* , **97**, 181–192.

12. Ramos,P.I.P., Custódio,M.G.F., Quispe Saji,G.D.R., Cardoso,T., da Silva,G.L., Braun,G., Martins,W.M.B.S., Girardello,R., de Vasconcelos,A.T.R., Fernández,E., *et al.* (2016) The polymyxin B-induced transcriptomic response of a clinical, multidrug-resistant Klebsiella pneumoniae involves multiple regulatory elements and intracellular targets. *BMC Genomics*, **17**, 737.