A Structural bioinformatic approach to prioritize drug targets in pathogens.

User Guide
Using Target-Pathogen

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.

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 upload in Target-Pathogen by clicking Genomes.

<table>
<thead>
<tr>
<th>Name</th>
<th>Proteins</th>
<th>Pathways</th>
<th>WBTS</th>
<th>Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mycobacterium tuberculosis MTBv - Overview</td>
<td>4533</td>
<td>205</td>
<td>182</td>
<td>1597</td>
</tr>
<tr>
<td>Mycobacterium avium AAFCC - Overview</td>
<td>5065</td>
<td>221</td>
<td>50</td>
<td>704</td>
</tr>
<tr>
<td>Plasmodium falciparum Y2 - Overview</td>
<td>1736</td>
<td>66</td>
<td>422</td>
<td>1772</td>
</tr>
<tr>
<td>Leishmania major PSL - Overview</td>
<td>8490</td>
<td>430</td>
<td>120</td>
<td>4030</td>
</tr>
<tr>
<td>Wolbachia endosymbiont of Drosophila melanogaster - Overview</td>
<td>996</td>
<td>210</td>
<td>6</td>
<td>956</td>
</tr>
<tr>
<td>Trypanosoma brucei gambiense BS10 Greenwood - Overview</td>
<td>995</td>
<td>254</td>
<td>123</td>
<td>1286</td>
</tr>
<tr>
<td>Salmonella enterica Subgroup F4 - Overview</td>
<td>6996</td>
<td>396</td>
<td>1181</td>
<td>1786</td>
</tr>
<tr>
<td>Shigella sonnei - Overview</td>
<td>11302</td>
<td>150</td>
<td>34</td>
<td>7107</td>
</tr>
<tr>
<td>Taenia solium N150 - Overview</td>
<td>6222</td>
<td>246</td>
<td>56</td>
<td>1280</td>
</tr>
<tr>
<td>Plasmodium vivax - Overview</td>
<td>5888</td>
<td>207</td>
<td>79</td>
<td>4987</td>
</tr>
</tbody>
</table>

Once you select your genome, you will be redirected to the genome page. Here, you will find three ways to explore the genome. First, the genome can be navigated in a fast,
smooth scrolling and zooming way using JBrowse. Using this browser you can scale from a entire genome to a single base view. Another interface offers a main search menu with several options to retrieve the protein structural druggability records. The options include the use of (i) Keyword (e.g. UniProt id, Protein Kinase PknB), (ii) Gene (NCBI locus tag or gene name) or (iii) Pathways. Searches may return a single database entry (e.g. when searching by Gene) or multiple entries (e.g. Keyword and pathways).

Also, genomes can be easily explored hierarchically by EC number (1) or the different categories of Gene Ontology by using Krona.
Let's assume that you are searching for proteins annotated with the go term mycothiol biosynthetic process, thus simply type it in the keyword field.

The resulting records are listed in the ranking page below the Filter and Score tab. For each record, size, druggability score, gene name and the number of pathways where the proteins are involved is presented. By clicking on the desired row, eight tabs of the corresponding record will be expanded.

In the Overview tab, general information and some metadata (if available), are presented. This tab also shows the higher druggability score detected in a protein pocket.
Other tabs show data about the protein's sequence, pathways, ontology and families (PFAM) (2, 3).

<table>
<thead>
<tr>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organism/Strain</td>
</tr>
<tr>
<td>Protein</td>
</tr>
<tr>
<td>Gene</td>
</tr>
<tr>
<td>Status</td>
</tr>
<tr>
<td>Length</td>
</tr>
<tr>
<td>Description</td>
</tr>
<tr>
<td>essentiality</td>
</tr>
<tr>
<td>dbRefLinks</td>
</tr>
<tr>
<td>metadata_ART_in_seq</td>
</tr>
<tr>
<td>overexpression - hypoxio</td>
</tr>
<tr>
<td>overexpression - infection</td>
</tr>
<tr>
<td>overexpression - starvation</td>
</tr>
<tr>
<td>overexpression - stress</td>
</tr>
<tr>
<td>pathways - centrality</td>
</tr>
<tr>
<td>pathways - checkpoint</td>
</tr>
<tr>
<td>Druggability</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Function/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIOCYC_COMP</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>EC</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>molecular_function (GO)</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>cellular_component (GO)</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>biological_process (GO)</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>PFAM</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>BIOCYC_PW</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>
At last, by clicking one model or crystal in the **Structure-Tab**, you will be directed to a page that presents the structure-related data, including the interactive pocket visualization module.

The visualization module allows the user (i) to select a pocket for graphical display, (ii) to display present heteroatoms, assigned PFAM or CSA relevant residues, (iii) to display the protein chains in different styles (iv) to display the alpha spheres or the pocket residues defining the pocket.
The displayed protein is available to download as a VMD file (4). Additional information about Model quality assessments and templates are shown at the top of this page. By clicking on the template chain, it's structure will be open.

Pockets predicted by fpocket (5) with a druggability score greater than 0.2 are also accessible in the right margin of the page in the 'Pocket list tab'. This tab shows all the pockets ordered by their druggability score. Finally, there is a Features List tab. By clicking in the different features you can stick out residues of interest as residues that bind crystallized ligands or conserved residues in PFAM.

**Querying Target-Pathogen to rank and prioritize drug targets.**

To obtain a short list of proteins that could be good candidate drug targets you have to choose your genome and click in Prioritize targets.
By doing that, you will be directed to a three tabs page. The first one, the “Filter tab”, allow researchers to filter molecular targets based on whether or not they fulfill a set of user defined criteria. For example if a protein is druggable or highly druggable.

![Filter Tab](image)

The second one is the “Score tab”. This tab allows you to assign different weights to different features of the proteins to set an user defined score function (SF). In the SF each selected parameter must have a coefficient, to weight the importance of such value. It can be fractional or even negative. It is important to always verify the distribution of values of the property, for example, the “chokepoints” prop is the number of chokepoints in a pathway, and “chokepoints_norm” is the number of chokepoints in a pathway divided by the number of reactions.

![Score Tab](image)

At last, you can combine different filters with the SF and Target-Pathogen will return a particular list of ranked genes according to an user predefined criteria. The third tab, 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 by clicking on “Export first 100 results” (for a quick view) or the “Download list” button (for a complete filtered list of proteins with the selected scores).
When scoring pathways, properties can belong to the proteins (for example “pocket with csa”) or to the pathway (for example “completeness”). When using the later, there is a direct map between the property and the SF. However when protein properties are used, there are many proteins in one pathway, so there are many ways to combine the values. The group column assigns an operator:

- max (Ex: druggability of the most druggable protein of the pathway)
- sum (Ex: protein count in the pathway with binding activity)
- min (Ex: protein with the lowest centrality)
- avg (Ex: hit in deg proteins / total proteins in the pw)
Score = druggability + hit_in_deg

**Properties Distribution.**

You can access the histogram of Target-Pathogen built in properties (for the selected genome), through the links in the filter of score table of the selected parameters:

**Score**

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

- **Activity**
- **Biological Process**
- **Localization**
- **Pathways**
- **Structure**
- **Pocket**
- **Metadata**
- **Add new Properties**

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Coefficient</th>
<th>Norm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>X centrality</td>
<td>Shortest-path betweenness centrality (normalized) for reactions. In the used graph the nodes are the reactions and the edges the metabolites connecting them. When centrality &gt;= 0.1 the reaction is considered highly central</td>
<td>0.5</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Properties for filtering and scoring proteins

- **EC (ENZYME number)** or **GO (Gene ontology)** Molecular Function terms. Results are sorted from particular to general.

- **GO Biological Process terms.** Results are sorted from particular to general.

- **GO Cellular Component terms.** Results are sorted from particular to general.

- **General protein data.** This category allows the use of the uploaded properties. A user can only view his/her own properties.

---

**Pocket with free tyr**

One of the protein structures has a tyr inside a pocket without atoms surrounding the oxygen of the OH group (more than 3 cubic Å).

**Pocket with tyr**

One of the protein structures has a tyr inside a pocket.

**Pocket with free cys**

One of the protein structures has a cys inside a pocket without atoms surrounding the S atom of the SH group (more than 3 Å).

**Pocket with cys**

One of the protein structures has a cys residue inside a pocket.

**Pocket with csa**

One of the protein structures has at least one residue inside a pocket reported in the Catalytic Site Atlas database.

**Pocket with domain extended**

One of the protein crystal has a pfam domain with residues inside a pocket near a drug or cofactor.

**Pocket with pfam imp residue**

One of the protein structures, has at least one residue, inside a pocket and a pfam domain, in contact (less than 3 Å) with a drug or cofactor.

**Pocket with drug binding**
One of the protein structures has at least one pocket residue in contact with a drug (less than 3 Å)

**Pocket with lipid binding**

One of the protein structures has at least one pocket residue in contact with a lipid (less than 3 Å)

**Pocket with metal binding**

One of the protein structures has at least one pocket residue in contact with a metal (less than 3 Å)

**Pocket with nucleotide binding**

One of the protein structures has at least one pocket residue in contact with a nucleotide (less than 3 Å)

**Pocket with sugar binding**

One of the protein structures has at least one pocket residue in contact with a sugar (less than 3 Å)

---

**Human offtarget**

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).

**Hit in deg**

Has a hit in Database of Essential Genes

**Centrality**

Shortest-path betweenness centrality (normalized) for reactions.

In the used graph the nodes are the reactions and the edges the metabolites connecting them.

When centrality $\geq 0.1$ the reaction is considered highly central
**Chokepoint**
The protein catalyzes a chokepoint reaction

**Chokepoint type**
Chokepoint reaction type (consume, production or double)

Has structure
Protein has a 3D structure

**Structure type**
Experimentally obtained or modelled

**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/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4014675/))

**Hydrophobicity**
Hydrophobicity of the most druggable pocket

**Volume**
Volume in cubic Å of the most druggable pocket

**Free tyr**
One of the protein structures has a tyr without atoms surrounding the oxygen of the OH group (more than 3 cubic Å)

**Tyr**
One of the protein structures has a tyr residue

**Free cys**
One of the protein structures has a cys without atoms surrounding the S atom of the SH group (more than 3 Å)

**Cys**
One of the protein structures has a cys residue

**Csa**
One of the protein structures has at least one residue reported in the Catalitic Site Atlas database

**Domain extended**
One of the protein crystals has a pfam domain with residues near a drug or cofactor

**Drug binding**
One of the protein structures has at least one residue in contact with a drug (less than 3 Å)

**Lipid binding**
One of the protein structures has at least one residue in contact with a lipid (less than 3 Å)

**Metal binding**
One of the protein structures has at least one residue in contact with a metal (less than 3 Å)

**Nucleotide binding**
One of the protein structures has at least one residue in contact with a nucleotide (less than 3 Å)

**Sugar binding**
One of the protein structures has at least one residue in contact with a sugar (less than 3 Å)

(Pathway properties)

**Reactions**
Number of reactions in the pathway

**Norm reactions**
Number of reactions normalized by the highest pathway (with more reactions)

**Reactions with gene**
Number of reactions in the pathway with at least one known protein that catalize them

**Completeness**
Proportion of reactions in the pathway with at least one known protein that catalize them

**Chokepoints**
Number of chokepoints reactions in the pathway

**Norm chokepoint**
Chokepoint reactions/reactions in pathway ratio

**Druggable**
The pathway has at least one druggable protein

**Max centrality**
Maximum betweenness centrality of all the reactions in the pathway, normalized by the reaction with max betweenes centrality in the whole graph
Only H37Rv

Essentiality

Overexpression stress

Overexpression starvation

Overexpression infection

Overexpression hypoxia

Only Kp13

Essential in mgh78578
Hits with an essential gene of Klebsiella pneumoniae MGH78578

Conserved pathogen
Hit count in different pathogen Klebsiella subspecies

Conserved no pathogen
Hit count in different non pathogen Klebsiella subspecies

Overexpressed in polymyxin

Overexpressed in polymyxin all conditions
Overexpressed in presence of Polymyxin in different conditions ([https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5088521/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5088521))

Conserved pathogen norm
Hit count in different pathogenic Kp strains divided by the total number of compared bacteria (39 pathogenic Kp).

Gut microbiome
Number of gut microbiome organisms that have at least one hit with the Kp13 protein (blast: identity > 40% evalue 1e-5)

**Gut microbiome norm**

gut_microbiome normalized by the total number of compared bacteria (226)

**Choosing promising pathways as putative targets of new drugs.**

Target Pathogen allow user 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). To Prioritize pathways you can combine different filters with the SF similar to the prioritization of proteins.

A high value would mean that most genes in the pathway and the pathway itself is attractive from the drug discovery point of view.

**Properties for filtering and scoring pathways**

- **EC (ENZYME number) o GO (Gene ontology)** Molecular Function terms. Results are sorted from particular to general

- GO Biological Process terms. Results are sorted from particular to general

- GO Cellular Component terms. Results are sorted from particular to general

- General protein data. This category allows the use of the uploaded properties. A user can only view his/her own properties.
Pocket with free tyr
One of the protein structures has a tyr inside a pocket without atoms surrounding the oxygen of the OH group (more than 3 cubic Å)

Pocket with tyr
One of the protein structures has a tyr inside a pocket

Pocket with free cys
One of the protein structures has a cys inside a pocket without atoms surrounding the S atom of the SH group (more than 3 Å)

Pocket with cys
One of the protein structures has a cys residue inside a pocket

Pocket with csa
One of the protein structures has at least one residue inside a pocket reported in the Catalytic Site Atlas database

Pocket with domain extended
One of the protein crystal has a pfam domain with residues inside a pocket near a drug or cofactor

Pocket with pfam imp residue
One of the protein structures, has at least one residue, inside a pocket and a pfam domain, in contact (less than 3 Å) with a drug or cofactor

Pocket with drug binding
One of the protein structures has at least one pocket residue in contact with a drug (less than 3 Å)

Pocket with lipid binding
One of the protein structures has at least one pocket residue in contact with a lipid (less than 3 Å)

Pocket with metal binding
One of the protein structures has at least one pocket residue in contact with a metal (less than 3 Å)

Pocket with nucleotide binding
One of the protein structures has at least one pocket residue in contact with a nucleotide (less than 3 Å)

Pocket with sugar binding
One of the protein structures has at least one pocket residue in contact with a sugar (less than 3 Å).

**Human offtarget**

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).

**Hit in deg**

Has a hit in Database of Essential Genes

**Centrality**

Shortest-path betweenness centrality (normalized) for reactions.

In the used graph the nodes are the reactions and the edges the metabolites connecting them.

When centrality >= 0.1 the reaction is considered highly central

**Chokepoint**

The protein catalyzes a chokepoint reaction

**Chokepoint type**

Chokepoint reaction type (consume, production or double)
Protein has a 3D structure

**Structure type**
Experimentally obtained or modelled

**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/)

**Hydrophobicity**
Hydrophobicity of the most druggable pocket

**Volume**
Volume in cubic Å of the most druggable pocket

**Free tyr**
One of the protein structures has a tyr without atoms surrounding the oxygen of the OH group (more than 3 cubic Å)

**Tyr**
One of the protein structures has a tyr residue

**Free cys**
One of the protein structures has a cys without atoms surrounding the S atom of the SH group (more than 3 Å)

**Cys**
One of the protein structures has a cys residue

**Csa**
One of the protein structures has at least one residue reported in the Catalitic Site Atlas database

**Domain extended**
One of the protein cristals has a pfam domain with residues near a drug or cofactor

**Drug binding**
One of the protein structures has at least one residue in contact with a drug (less than 3 Å)

**Lipid binding**
One of the protein structures has at least one residue in contact with a lipid (less than 3 Å)

**Metal binding**
One of the protein structures has at least one residue in contact with a metal (less than 3 Å)

**Nucleotide binding**
One of the protein structures has at least one residue in contact with a nucleotide (less than 3 Å)

**Sugar binding**
One of the protein structures has at least one residue in contact with a sugar (less than 3 Å)

(Pathway properties)

Reactions
Number of reactions in the pathway

Norm reactions
Number of reactions normalized by the highest pathway (with more reactions)

Reactions with gene
Number of reactions in the pathway with at least one known protein that catalize them

Completeness
Proportion of reactions in the pathway with at least one known protein that catalize them

Chokepoints
Number of chokepoints reactions in the pathway

Norm chokepoint
Chokepoint reactions/reactions in pathway ratio

Druggable
The pathway has at least one druggable protein

Max centrality
Maximum betweenness centrality of all the reactions in the pathway, normalized by the reaction with max betweeness centrality in the whole graph

(Only H37Rv)

Essentiality

Overexpression stress

Overexpression starvation
Overexpression infection

Overexpression hypoxia

Only Kp13

Essential in mgh78578
Hits with an essential gene of Klebsiella pneumoniae MGH78578

Conserved pathogen
Hit count in different pathogen Klebsiella subspecies

Conserved no pathogen
Hit count in different non pathogen Klebsiella subspecies

Overexpressed in polymyxin
Overexpressed in polymyxin B resistance induction (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5088521/)

Overexpressed in polymixin all conditions
Overexpressed in presence of Polymyxin in different conditions (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5088521/)

Conserved pathogen norm
Hit count in different pathogenic Kp strains divided by the total number of compared bacteria (39 pathogenic Kp).

Gut microbiome
Number of gut microbiome organisms that have at least one hit with the Kp13 protein (blast: identity > 40% evalue 1e-5)

Gut microbiome norm
Gut_microbiome normalized by the total number of compared bacteria (226)


