NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

Systems Biology MSc Computational Biology and Bioinformatics
2022/2023
Professor Isabel Rocha

Identification of putative drug targets on *Mycobacterium tuberculosis*

Organism choice and its relevance in terms of public health

The model that was chosen to perform the drug target analysis was *Mycobacterium tuberculosis*, an aerobic bacillus *mycobacteria* that is responsible for the tuberculosis disease in humans. [1]

The research for an effective drug target concerning this pathogen is urgent given that tuberculosis is a major cause of death worldwide and one of the most lethal diseases at the moment. According to the Global tuberculosis report 2022 published by World Health Organization (WHO), an estimated 10.6 million people were infected worldwide, being this disease, the leading cause of death from a single infectious disease in the world, even deadlier than HIV, thus the importance for finding a suitable drug target to fight this disease. [2]

The standard treatment for tuberculosis patients consists, mainly of four drugs, lasting over 6 months, namely 2 months of isoniazid, rifampicin, ethambutol and pyrazinamide, followed by 4 months of isoniazid and rifampicin with a 85% rate of success. Although, the rise of several drug-resistant strains over last years, not only due to the prescription of wrong medicines or poor quality drugs or drug unavailability poses the need to research new drugs that can neutralize this lethal disease. [3,4]

Given that, in this project we aim to be able to identify proteins that can be potential drug targets for *Mycobacterium tuberculosis* and finally identify suitable drugs or inhibitors that can act on the targets through BRENDA and DrugBank.

Students: Frederico Água nº 64737
Pedro Duarte nº 64897
Rodrigo Bernardino nº63842

Model Used

We use the updated and standardized model of *Mycobacterium tuberculosis* H37Rv, iEK1011, that can be found at https://github.com/erolkavvas/iEK1011_materials. This is an updated version of the iEK1008 model that can be found in the BIGG Models database. Although the model is not available in SBML format we were able to perform the whole pipeline with JSON format available using cobrapy.

Target Identification Pipeline

For a preliminary identification of proteins that could be potential targets for drugs we simply identified the essential genes in the model using the find_essential_genes function from cobrapy. Having identified the essential genes we crossed the data with the *Mycobacterium tuberculosis* (strain ATCC 25618 / H37Rv) proteome that was accessed from UniProt (https://www.uniprot.org/proteomes/UP000001584) to get the sequences for each protein in a FASTA format. We were able to cross this information because the gene codes in the model matched the ones on the used proteome except for the last digit.

Having the protein sequences, we did protein BLAST against the human proteins in the NCBI protein ref seq database. The BLAST step was automated using the Bio Python package (https://biopython.org/) which contains tools that allow to run the commands in a python script but do the search over the web using the NCBI server. After getting the BLAST results we selected the proteins that had no hits in the search.

Omega Fold (https://github.com/HeliXonProtein/OmegaFold) is a pLm (protein language model) similar in architecture to Alphafold 2 but specialized in single sequence reads, which determines protein structures from sequence. We used the following Omega Fold commands –num_cycle 2 and –subbatch_size 148 in order to reduce the computational requirements (2 repetitions per fasta instead of the standard 10) and cap the GRAM (graphical processing unit ram) usage respectively. A batch script was created to automate the process.

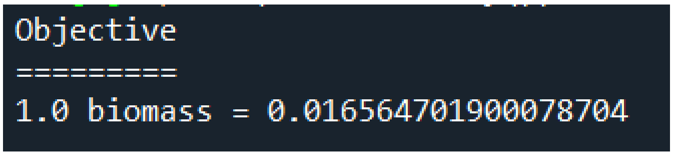
These structures were then used in another tool, P2Rank (https://github.com/rdk/p2rank), that predicts ligand-binding pockets in a protein structure. We ran the standard predict -f for a pdb file generated by Omega Fold and automated it for the entire folder using a bash script. The output data collected was then interpreted programmatically using python. A batch script was created to automate the process.

Using this tool we can identify in the given proteins which ones may have ligand-binding pockets where, possibly, an inhibitor could bind. The output will give us scores for the probability of the protein having a ligand-binding pocket and we will choose the highest scoring ones. Assuming that proteins with ligand-binding sites will most likely have molecules that can bind to those sites and inhibit their functions, selecting these proteins will allow us to make a more informed decision on what targets we will manually curate rather than just selecting some targets randomly.

After selecting the top 5 proteins ranked by P2Rank, that showed no hits in humans, we proceeded to manually curate them, looking for known inhibitors or drugs that target these proteins in BRENDA and DrugBank.

Results and Discussion

First, we loaded the model with cobrapy and ran wild-type simulation to ensure that the results matched the expected, as accessible in the pipeline that can be found together with the model on GitHub. Indeed we observed a similar biomass value of 0.01654 as shown in Figure 1.



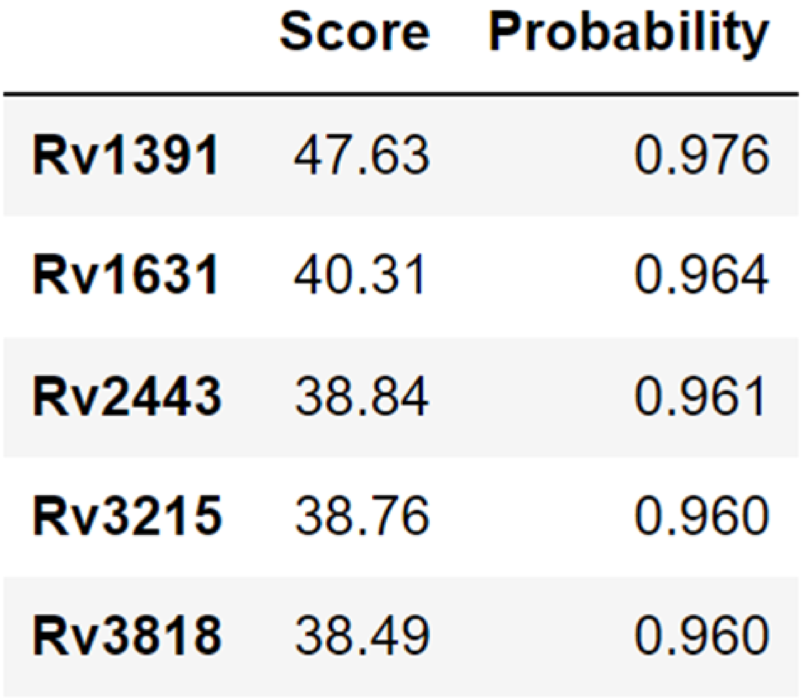
```
Objective
=========
1.0 biomass = 0.016564701900078704
```

Figure 1- Wild-type simulation results of biomass for iEK1011 using cobrapy in default environment conditions.

The identification of the essential genes and the conversion of genes to proteins was easily and quickly performed. The protein BLAST and Omega Fold steps did take a significant amount of time to run but also yielded the desired results. The structures from OmegaFold were used in P2Rank and we got the output for each identified protein. By combining the data from all outputs and selecting only the ones that had no hits in the BLAST against the human proteome we were able to select the highest scoring proteins from P2Rank that have no matches in humans. After getting the top 5 proteins we repeated the BLAST manually on another database, the non redundant protein sequences in NCBI, to be sure the results were the same - when dealing with drug targeting we really don't want to accidentally target a relevant human protein. Indeed there was a hit that did not appear in the ref seq database so we discarded that gene. The final top 5 are shown in Figure 2. Having identified these targets we ran simulations of knockouts on these genes to confirm that indeed they would cause the biomass function to be zero. Running this multiple times, the values oscillated slightly but were always zero or very close to zero so we can assume that these knockouts do reduce the biomass function to zero, effectively killing the organism. One example of the runs is shown in Figure 3.

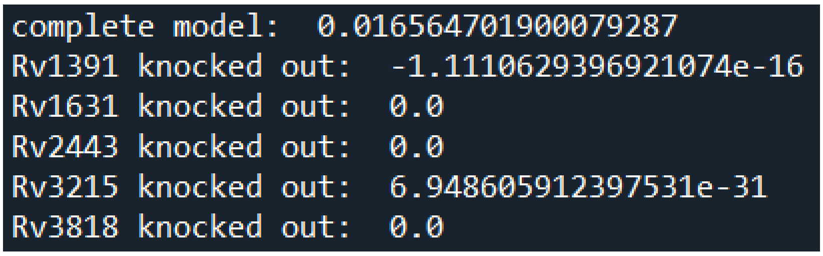
Taking these top proteins, we proceeded to manually look in BRENDA and Drug Bank for known inhibitors and drugs that could target these proteins. Rv1391 codes for Phosphopantothenoylcysteine Decarboxylase that is involved in Coenzyme A metabolism. 4'-phosphopantothenol has been identified as an inhibitor of this enzyme. Indeed, this relation has already been shown [5] and in fact the Coenzyme A synthesis is a known target for antimicrobial drugs [6]. When looking at these facts, it makes sense that the second protein identified, Dephospho-CoA kinase, coded by Rv1631, is another protein from this metabolism. CTP is a known inhibitor of this protein [7] and by searching in the Drug Bank, using the sequence of the protein, a match was found for Pretomanid. In fact, Pretomanid is already used in the treatment of drug-resistant pulmonary tuberculosis, particularly against *Mycobacterium tuberculosis*.

The third protein that was identified is a malonyl-CoA:AcpM transacylase involved in *M.tuberculosis* lipid metabolism. In BRENDA some known inhibitors were identified but they were all bivalent cations identified when the protein was first described [8] and not relevant substances. In Drug Bank the closest match was an inhibitor for Malonyl-CoA-acyl carrier protein transacylase. While this protein did not show a hit in the BLAST, this protein identified in Drug Bank is a mitochondrial protein so using this inhibitor would not be wise.



	Score	Probability
Rv1391	47.63	0.976
Rv1631	40.31	0.964
Rv2443	38.84	0.961
Rv3215	38.76	0.960
Rv3818	38.49	0.960

Figure 2- Highest scoring proteins according to P2Rank that have no hits in the BLAST against the human proteome. The codes represent the gene codes in the model.



```
complete model:  0.016564701900079287
Rv1391 knocked out:  -1.1110629396921074e-16
Rv1631 knocked out:  0.0
Rv2443 knocked out:  0.0
Rv3215 knocked out:  6.948605912397531e-31
Rv3818 knocked out:  0.0
```

Figure 3 - Values for biomass function in simulations of the wild-type and each mutant with knock-out of the identified genes.

Rv3215 codes for an Isochorismate synthase that plays a role in the synthesis of small molecules that act in iron mobilization on *Mycobacterium*. A study on inhibition of these processes [9] found several inhibitors, namely benzimidazole-2-thione and osetalmivir, that also matched in our search at Drug Bank.

Lastly, Rv3818 codes for a UDP-MurNAc hydroxylase that plays a role in cell wall metabolism, a typical target for antibiotics. While no annotation was found on BRENDA, the search in Drug Bank showed a match for kanamycin, a well-known antibiotic.

Overall, we can see that these results make sense, 4 out of the top 5 ranked targets having already identified inhibitors with pharmacological relevance. While no novel results were found, the ability of this pipeline to identify these targets shows the relevance of its use. If we were looking at a less well know organism probably no previously identified drugs would be found but the fact that this pipeline was able to select well known drug targets shows it's potential so select putative targets in those less well studied organisms where further work could be performed, either computationally via ligand interaction simulations or experimentally to find drugs that could help fight the organism in question.

Conclusion

This pipeline proved to be efficient to identify drug targets in a microorganism of interest. While the tools used are not too computationally heavy, we did have some bottlenecks in the running time, namely while running the protein BLAST and Omega Fold. Nonetheless we still think this is a relevant pipeline for identification of drug targets given its ability to identify relevant targets on *Mycobacterium tuberculosis*. Problems in reproducing our work on other organisms may arise if the gene codes used in the models don't match the codes of any proteome in databases, which would pose a problem to pass from the genes to the protein sequences in FASTA. On a note about possible improvements on the pipeline, perhaps the BLAST could be run for multiple databases, allowing to bypass situations like we had where one of the initial top 5 had a hit on humans in another database.

Code

All code and files used are available at https://github.com/fredaguas7/SystemsBiologyProject.

Bibliography

[1] Singh, Ramandeep (2020). *Drug Discovery Targeting Drug-Resistant Bacteria || Drugs against Mycobacterium tuberculosis. , (), 139–170.* doi:10.1016/B978-0-12-818480-6.00006-0

[2] https://www.who.int/publications/i/item/9789240061729 (Consultado a 24/04/2023)

[3] Allué-Guardia A, García JI and Torrelles JB (2021) Evolution of Drug-Resistant Mycobacterium tuberculosis Strains and Their Adaptation to the Human Lung Environment. Front. Microbiol. 12:612675. doi: 10.3389/fmicb.2021.612675

[4] Goossens, S., Sampson, S. L., & Van Rie, A. (2020). Mechanisms of Drug-Induced Tolerance in Mycobacterium tuberculosis. Clinical Microbiology Reviews, 34(1). https://doi.org/10.1128/cmr.00141-20

[5] Kumar, P., M. Chhibber, and A. Surolia, How pantothenol intervenes in Coenzyme-A biosynthesis of Mycobacterium tuberculosis. Biochemical and Biophysical Research Communications, 2007. 361(4): p. 903-909.

[6] Christina Spry, Kiaran Kirk, Kevin J. Saliba, Coenzyme A biosynthesis: an antimicrobial drug target, FEMS Microbiology Reviews, Volume 32, Issue 1, January 2008, Pages 56–106, https://doi.org/10.1111/j.1574-6976.2007.00093.x

[7] Walia, G.; Surolia, A. Insights into the regulatory characteristics of the mycobacterial dephosphocoenzyme A kinase: implications for the universal CoA biosynthesis pathway (2011), PLoS ONE, 6, e21390.

[8] Yi-Shu Huang, Jing Ge, Hong-Mei Zhang, Jian-Qiang Lei, Xue-Lian Zhang, Hong-Hai Wang, Purification and characterization of the Mycobacterium tuberculosis FabD2, a novel malonyl-CoA:AcpM transacylase of fatty acid synthase,Protein Expression and Purification,Volume 45, Issue 2,2006,Pages 393-399,https://doi.org/10.1016/j.pep.2005.07.003.

[9] Liu Z, Liu F, Aldrich CC. Stereocontrolled Synthesis of a Potential Transition-State Inhibitor of the Salicylate Synthase MbtI from Mycobacterium tuberculosis. The Journal of Organic Chemistry. 2015 Jul;80(13):6545-6552. DOI: 10.1021/acs.joc.5b00455. PMID: 26035083; PMCID: PMC4667787.