Finding Paralog Targets for Neglected Diseases

By

Jeremy Singer

# Abstract

This paper describes a method that can be used to discover and repurpose existing drugs and drug targets by discovering cross species genomic sequence similarities. It uses public domain databases (ChEMBL, EnSEMBL, NCBI) and open source software to find measures of sequence similarity with existing targets.

This method can be applied to pathogens with at least a medium sized genome (several thousand genes.) *Neglected tropical diseases* caused by pathogenic protists are good subjects for this approach because many have genomes of sufficient size and because many have genomic features in common with organisms for which there are known targets.

The genome of the apicomplexan parasite *Plasmodium falciparum*, which is responsible for the most virulent form of malaria, was chosen to validate a method that identifies paralogs to existing disease targets because it has known cross-species targets.

ChEMBL provides a PostgreSQL database that contains a list of thousands of targets and target protein sequences as well as ligands for those targets. Using this database and open source software, this paper identified 29 distinct drugs and 592 targets validating this approach.

Six other pathogens (*Trypanosoma Brucei, Trypanosoma Cruzi, Leishmania Major,*  *Chlamidia trachomatis, Toxoplasma Gondii, and entamoeba )* were also downloaded and run through the same pipeline, identifying potential targets and drugs.

# Introduction

*Neglected Tropical Diseases* are those diseases that affect tropical areas underserved for health care due to the poverty of those areas. These diseases affect over a billion people, and damage the economies of these areas at a cost of many billions of dollars.

Repurposing drugs and generating leads for finding new drugs by repurposing targets could be a cost -effective way for combating these diseases.

ChEMBL provides a downloadable database that includes drug targets and drug information for those targets, as well as amino acid sequences of the protein targets. Drug targets tend to be proteins that are important enough to the organism to which they belong that they tend to be conserved. If we can find a protein sequence in a disease organism that is sufficiently similar to a know target, the protein may be a promising target in that organism, and drugs used against that target may be successful.

Two methods were used to screen proteins from *plasmodium falciparum*, the pathogen that causes the most virulent form of malaria to validate this approach, and the similarity statistics from these screens were uploaded to a new table. We used BLASTP and HMMER to generate these statistics.

Using these statistics, we used analytical methods to identify promising targets and drugs.

In addition to *p. falciparum*, we processed the following additional pathogens using [*preferred method]*:

[*pathogen list].* The statistics were loaded into supplementary tables in the PostgreSQL database.

Queries using the existing ChEMBL database, in combination with these similarity statistics were used to identify targets and drugs for each of these pathogens.

# Materials and Methods

# Results

# Discussion

# References

1. Neglected tropical diseases, <https://www.who.int/neglected_diseases/diseases/en/>
2. REPURPOSING STRATEGIES FOR TROPICAL DISEASE DRUG DISCOVERY <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4853260/>
3. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach AR. The ChEMBL database in 2017. Nucleic Acids Res. 2017 45(D1):D945-D954. DOI: 10.1093/nar/gkw1074