Finding Paralog Targets for Neglected Diseases

# 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 (MyCheEMBL, EnSEMBL, NCBI) and open source software to find measures of sequence conservation 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.

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

# Introduction