December 30th, 2017

Ruth Nussinov, Ph.D.

Jason Papin, Ph.D.

Philip E. Bourne, Ph.D.

Editors-in-Chief

PLOS Computational Biology

Re: Submission of Manuscript Entitled, "Discovering new targets and drugs for neglected diseases by paralog matching"

Dear Drs. Nussinov, Papin, and Bourne:

I am pleased to submit the manuscript entitled "Discovering new targets and drugs for neglected diseases by paralog matching" for your consideration for publication in *PLOS Computational Biology*.

Current methods of repurposing drugs and targets use approaches that are ligand centric or mechanism centric for the target infectious disease organism; the approach in this paper is genome-centric, using the genes of the organism itself as keys to reveal their “target-ness” by similarity to proteins in the targets database.

The manuscript describes a novel pipeline to discover disease targets for neglected disease pathogens by computing similarity between a targets database to all the genes of an organism of interest. Analysis of the data using R determines which targets may be repurposed by establishing statistical criteria that identify a similarity scoring threshold that distinguishes between functionally related genes and all other genes in the data set. This data-directed method may also highlight the presence of processes in the pathogens of interest that were not previously suspected – the genome itself will lead us there.

Results in this paper validate a method for discovering cross species targets by identifying 29 distinct drugs for malaria (53, counting different formulations). This list of drugs includes many already known to be effective, as well as identifying new candidate drugs. 592 new targets are identified for further study.

In addition, I downloaded five other neglected disease pathogens (*Trypanosoma Brucei, Trypanosoma Cruzi, Leishmania Major, Chlamidia trachomatis, and Toxoplasmosis Gondii)* and ran them through the same pipeline, identifying potential targets and drugs.

The methods described in the paper use free, open source software as well as *Extract Transform and Load* (ETL) methods written by the author using Perl, bash, and R to computationally discover new information using publicly available data from *myChEMBL, NCBI, and EnsemblGenomes*. It is this *computational* content which makes it suitable for inclusion in this journal.

I have provided an example workbook demonstrating the techniques used, so that others can easily reproduce these results or make new discoveries using other genomes or target databases.

The **blast\_statistics.txt** values provided in the supporting data can be used to easily explore annotations for the drugs and targets discovered in this paper in combination with the existing *myChEMBL* PostgreSQL database.

I hope this will inspire others by showing that one can discover new and valuable knowledge with a very small budget and some effort. These methods of repurposing drugs and targets can derive new value from old data.

The results presented in this manuscript have not been published previously nor are they being considered by any other journal. The author has declared no conflicts of interest.

Please let me know if you have any questions or need any additional information. Thank you very much for your consideration.

Sincerely yours,

Jeremy Singer

B.A., Brandeis University 1975

Currently studying for a Master of Science in Bioinformatics at Brandeis University’s GPS program.

Senior Database Engineer at LifeImage.

e-mail: jsinger@rcn.com