A target and two drugs for SARS-CoV-2 found by paralog search

Author: Jeremy B. Singer[[1]](#footnote-1)\*

# Abstract

Using a paralog search pipeline, the author searched the ChEMBL 25 database, screening targets in it against the SARS-CoV-2 genome finding a target that has an identical target sequence. The the target, RNA polymerase, was found to have 100% identity with a gene in the viral genome of SARS-CoV-2.

Three known drugs in the ChEMBL 25 database are associated with the target that was identified. Two of those drugs showed high binding affinity in docking simulations, validating them as promising drug candidates for laboratory testing against SARS-CoV-2 infection.

# Introduction

(Fig 1.)

An analysis pipeline searches the **chembl\_25** version of ChEMBL’s database for targets and drugs using viral genomic information from Genbank.

SARS-CoV-2, also known as COVID-19, is a virus that causes flu like symptoms including respiratory distress, in many cases requiring respirators to maintain oxygenation in patients. It is highly contagious, and is currently causing pandemic infection, with a fatality rate estimated between 2% and 3% [1]. Persons over 60 have may have much higher fatality rates [2].

A target repurposing strategy might provide drugs more quickly and cheaply than creating new drugs and finding new targets [3]. This strategy could produce treatments to ameliorate the disease until a vaccine becomes available, or in addition to vaccines.

**DISCLAIMER:** The workflow described in this paper was initially conceived as a way to rapidly and cheaply identify targets and drugs for neglected tropical diseases [3, 4]. **Results from this workflow are not definitive, clinicly approved theraputic options.**  **These methods are intended to identify targets and drugs for further study.** While some dosage and toxicity data are available for the drugs in question from their original usage, the the therapeutic effect of these drugs against SARS-CoV-2 has not been studied.

The workflow uses targets and drugs in databases that are downloadable from the internet, genomic data (in the form of amino acid sequences, (also available from internet sources) and a software stack that can run on a PC, or on a VM hosted on a PC. Specifically, the data and software stack can run on a PC running Windows 10, Centos 7 Linux running in a Virtualbox VM, and a PostgreSQL database running under Centos 7 [5-9]. All software needed are available for free, and run on a PC with at least a 4 core 64 bit Intel compatible CPU, and having 8 GB of ram.

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 [10]. Drug targets tend to be proteins that are important enough to the organism to which they belong that they tend to be conserved [11]. If we can find a protein sequence in a disease organism that is sufficiently similar to a known target, the protein may be a promising target in that novel organism, and drugs used against that target in the original organism may be successfully used against the new pathogen.

Genbank provides a nucleotide sequence database containing genomes of many organisms, including the SARS-CoV-2 virus [12].

EMBOSS tools provide the **getorf** utility that finds the genes in the SARS-CoV-2 genome[13]. ORFs (Open Reading Frames) are amino acid sequences, including protein coding genes, that we wish to screen for sequence similarity to targets from ChEMBL.

An ETL (Extract Translate and Load) process downloads the drug target sequences into a file that can be queried with sequence similarity searching software such as **jackhmmer [14].**

**jackhmmer** produced similarity reports and scoring summaries for the ORFs. ETL scripts parsed these results, and uploaded them to supplementary tables in the PostgreSQL database that contained the **chembl\_25** database schema that had been downloaded.

Once the summary information is in the database, statistical methods such as kmeans established a similarity threshold to select the most promising targets and their drugs [15-17].

Swissdock, an opensource docking server, was used to evaluate the identified drugs for docking affinity to the target(s) found to validate whether these drugs might be effective [18].

# Materials and methods

(Fig 2 .) Target and drug analytical workflow.

Before executing the workflow, VirtualBox and the Centos 7 Linux image are installed and executed on the work PC. An empty PostgreSQL database is installed, and a single user named “user” created [5-7, 9].

The PostgreSQL dump archive of the ChEMBL version 25 database is downloaded, decompressed, and restored in the Centos 7 VM.

EMBOSS tools are installed [13].

The genome of SARS-CoV-2 (The COVID-19 virus) was downloaded from Genbank via NCBI’s website as MN908947.3.FASTA. [12].

ORFs were translated using EMBOSS tools [13]. The orfs were translated rather than using curated protein sequences because the translation tool provides the amino acid FASTA sequences in a form convenient for querying. While many ORFs may not be actual proteins, the query that uses these ORFs will eliminate them from consideration.

This command creates a file containing all the open reading frames (ORFs) found and translates the nucleotide sequences into amino acid sequences:

**getorf MN908947.3.FASTA**

This creates file **mn908947.orf**, which contains all the ORFs found for the .FASTA file.

The set of target sequences comes from the **ChEMBL\_25** PostgreSQL database and was downloaded by a *psql* script **chembl\_25\_targets.sql** as file **chembl\_targets.txt**.

These targets are converted by a Perl script (**split\_to\_fasta.pl**) creating file **component\_sequences.fa**.

Using **jackhmmer** to provide similarity reports and sequence alignments, a pipeline imported scores showing sequence and structural similarity [14]. This process created files **orf.hmm.txt** and **orf.summary**.

The Perl script **extract\_hmm\_summary.pl** reads the **orf.hmm.txt** file and creates the file **hmm\_stats.tx**t.

From the psql database query prompt, the data were imported into the **chembl\_25** database:

[postgres@osboxes /home/osboxes/genomes] **psql -U postgres -d chembl\_25**

psql (9.2.24)

Type "help" for help.

chembl\_25=# **\i import\_hmmer\_statistics.sql**

TRUNCATE TABLE

INSERT 0 49

chembl\_25=# **update hmmer\_statistics set tax\_id=2697049, organism='SARS-CoV-2’ where tax\_id is null;**

The UPDATE statement is necessary to differentiate the statistics from those of other uploaded genomes.

Kmeans analysis provided a similarity threshold for selecting targets interestingly similar to the target database using a user-defined R function (**organism\_hmmer\_threshold.R**) [15] .

get\_kmeans\_threshold(conn,2697049)

Using the returned threshold value, targets and drugs were retrieved using query **target\_SARS-COV-2\_drugs.sql** creating **target\_SARS-CoV-2\_drugs.txt**.

SWISSDOCK, a protein/ligand docking simulation website, was used to validate the drug candidates [18].

# Results and discussion

## The nature of targets

In the context of parasitic disease organisms, the “targetness” of a protein has to do with how indispensable its function is to the organism in question, since we are trying to kill the organism, or impair its success [11]. We are interested in protein targets that are highly conserved, because this indicates that the protein, in its conserved form, is necessary for the success of the disease organism [19]. We do not know specifically whether it is necessary for its infectious ability, its metabolic role, ability to transcribe DNA, translate proteins, or participate in the structure or outer integument of the organism. In addition, it will not be known whether the binding properties of the protein to any particular ligand has been preserved, even if the target is still useful as a target.

This description of the nature of protein targets suggests that paralogous proteins in our organism of interest could also be targets, if they are sufficiently similar to existing targets. The closeness of the match will suggest that the function of the protein has been conserved between the previously identified target organism and our organism of interest. Those sequences in the pathogen organism which are most necessary for its survival are also least likely to change, as mutation would tend to impair functions necessary for survival [19]. At the same time, we are searching exactly for those critically necessary proteins as targets for drugs that can impair them.

This approach is completely data driven, mechanism-agnostic method. The only principle followed is that sequence and structural conservation are directly related to survival of the organism of interest.

Uncertainty about whether existing drugs will effectively bind or interfere with the target proteins we identify is somewhat compensated for by the improvements of convenience and cost due to availability of the existing drugs, understanding of their dosage, and safety from existing studies [20, 21]. To find likely targets in the genome, we need to measure similarity between ORFs from its genome and our target database. When we have computed these similarities, we need to choose threshold criteria for filtering the most promising candidates.

## Exploiting similarity with a curated target and drug database

The original idea for the workflow discussed in this paper came from observing that the ChEMBL database structure supports relating drug targets, drug molecules, and target component sequences. The database structure might provide an easy way to find targets and drugs for other pathogens if we could find proteins in their genomes that were similar enough. What was missing was a table with similarity results for those pathogens, and criteria for filtering the results.

(Fig 3.)

Commands run in R Studio quantify how many ORFs are produced in results from **getorf**:

|  |
| --- |
| > aa=read.table(file="mn908947.orf",header = FALSE, sep='~', stringsAsFactors = FALSE)  > aa=aa[!is.na(aa[,1]),] # filter out NA  > aa=data.frame(lines=aa, stringsAsFactors = FALSE)  > orf\_headers=aa[substr(aa[,1],1,1)=='>' ,]  > length(orf\_headers)  [1] 1572 |
|  |
| |  | | --- | |  |   1572 ORFs were found. |

Although annotated ORFs for SARS-CoV-2 can be found elsewhere, this analysis relies only on the original nucleotide genome and the **chembl\_25** database curated by ChEMBL [10]. The number of ORFs may include pseudogenes and other non-protein coding genes. Those ORFs were filtered out because they did not have sufficient similarity to any target sequences to be included in the results when used to search the target database using **jackhmmer** [14].

**Jackhmmer** scores similarity between amino acid sequences by aligning query and target sequences [14]. In addition, **jackhmmer** uses hidden Markoff models (HMM) that assess patterns by looking for larger domains [14].

(Fig 4.)

The summary for a query may hit multiple targets. Each target record is repeated for each domain that **jackhmmer** matches. For this study, we are only using similarity across the whole protein as a measure of conservation. The Perl script (**extract\_hmm\_summary.pl**) de-duplicates results, reporting only these global measures of similarity. The *score* statistic is a number that additively reflects the similarity of sequences and domains in the protein.

49 ORFs had enough similarity to targets to participate in our analysis, and work loaded into the **hmmer\_statistics** table in the **chembl\_25** database.

This histogram shows the distribution of scores:

(Fig 5.)

Kmeans analysis shows that there is an outlier having a much higher similarity score. The threshold, 4350.9, was used to filter in the most similar targets and their drugs (See query **target\_SARS-COV-2\_drugs.sql**, results **target\_SARS-CoV-2\_drugs.txt)**.

Table : target and drugs retrieved.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **score** | **original tax id** | **orig\_organism** | **target\_type** | **target\_name** | **target\_chembl\_id** | **drug\_name** | **drug\_chembl\_id** |
| 4350.9 | 1773 | Mycobacterium tuberculosis | PROTEIN NUCLEIC-ACID COMPLEX | 70S ribosome | CHEMBL2363965 | VIOMYCIN SULFATE | CHEMBL3989823 |
| 4350.9 | 1773 | Mycobacterium tuberculosis | PROTEIN NUCLEIC-ACID COMPLEX | 70S ribosome | CHEMBL2363965 | CAPREOMYCIN SULFATE | CHEMBL2218913 |
| 4350.9 | 1773 | Mycobacterium tuberculosis | PROTEIN NUCLEIC-ACID COMPLEX | 70S ribosome | CHEMBL2363965 | PYRAZINAMIDE | CHEMBL614 |

# Conclusions

# References

1. **Early Release - Case-Fatality Risk Estimates for COVID-19 Calculated by Using a Lag Time for Fatality - Volume 26, Number 6—June 2020 - Emerging Infectious Diseases journal - CDC**. 2020.

2. **Coronavirus Age, Sex, Demographics (COVID-19) - Worldometer**. 2020.

3. Pollastri MP, Campbell RK: **Target repurposing for neglected diseases**. *Future Med Chem* 2011, **3**(10):1307-1315.

4. **WHO | World Health Organization**. *WHO* 2019.

5. **PostgreSQL: The world's most advanced open source database** [<https://www.postgresql.org/>]

6. **PostgreSQL: Downloads** [<https://www.postgresql.org/download/>]

7. **The CentOS Project** [<https://www.centos.org/>]

8. **Explore Windows 10 OS, Computers, Apps, & More | Microsoft** [<https://www.microsoft.com/en-us/windows>]

9. **VirtualBox** [<https://www.virtualbox.org/>]

10. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E *et al*: **The ChEMBL database in 2017**. *Nucleic Acids Res* 2017, **45**(Database issue):D945-954.

11. Lv W, Xu Y, Guo Y, Yu Z, Feng G, Liu P, Luan M, Zhu H, Liu G, Zhang M *et al*: **The drug target genes show higher evolutionary conservation than non-target genes**. In: *Oncotarget.* vol. 7; 2016: 4961-4971.

12. **Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, co - Nucleotide - NCBI** [<https://www.ncbi.nlm.nih.gov/pubmed/>]

13. **EMBOSS: The European Molecular Biology Open Software Suite (2000)** [<http://emboss.sourceforge.net>]

14. Wheeler TJ, HHMI Janelia Farm Research Campus A, VA 20147, USA, Eddy SR, HHMI Janelia Farm Research Campus A, VA 20147, USA: **nhmmer: DNA homology search with profile HMMs**. *Bioinformatics* 2013, **29**(19):2487-2489.

15. **forgy: Initialization of cluster prototypes using Forgy's algorithm in inaparc: Initialization Algorithms for Partitioning Cluster Analysis**. 2020.

16. Hartigan JA, Wong MA: **Algorithm AS 136: A K-Means Clustering Algorithm**. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 1979, **28**(1):100-108.

17. Sheridan RP, Shpungin J: **Calculating similarities between biological activities in the MDL Drug Data Report database**. *J Chem Inf Comput Sci* 2004, **44**(2):727-740.

18. **SwissDock - The online docking web server of the Swiss Institute of Bioinformatics - Home** [<http://www.swissdock.ch/>]

19. Alberts B: **Molecular Biology of the Cell**, 6 edn. Kindle Edition. : W. W. Norton & Company.

20. Dimasi J: **Innovation in the pharmaceutical industry: New estimates of R&D costs**. *Journal of Health Economics* 2016, **47**(May 2016):3.

21. Janes J, Young ME, Chen E, Rogers NH, Burgstaller-Muehlbacher S, Hughes LD, Love MS, Hull MV, Kuhen KL, Woods AK *et al*: **The ReFRAME library as a comprehensive drug repurposing library and its application to the treatment of cryptosporidiosis**. 2018.

# Supporting information

## **S1 Fig. An analysis pipeline searches the chembl\_25 version of ChEMBL’s database for targets and drugs using viral genomic information from Genbank.**

## **S2 Fig. Target and drug analytical workflow.**

## MN908947.3.FASTA.

The nucleotide FASTA formatted genome sequence of SARS-CoV-2.

## chembl\_25\_targets.sql

psql script to download chembl\_25 target sequences.

## chembl\_targets.txt

Downloaded target sequences.

## split\_to\_fasta.pl

Converrts a text file containing sequences to .FASTA format file.

## component\_sequences.fa

Searchable target component file.

## mn908947.orf

ORFs from SARS-CoV-2 genome.

## orf.hmm.txt

A report containing scores and alignments between the ORFs and targets.

ORFs having insufficient similarity have a record that says “[No hits detected that satisfy reporting thresholds].”

## orf.summary

A tab delimited file containing records with statistics of significant matches for ORFs and targets.

## extract\_hmm\_summary.pl

Extracts summary statistics from **orf.hmm.txt** and writes them to file hmm\_stats.txt.

## hmm\_stats.txt

Uploadable statistics file.

## import\_hmmer\_statistics.sql

Uploads statistics file to a work table, and then to **hmm\_statistics** table.

## **S3 Fig. Query joins connecting targets, sequences, and drugs. Tables with blue backgrounds are supplementary tables populated by this workflow.**

## **S4 Fig. ORF FASTA summary from jackhmmer**

## **S5 Fig. Target similarity score distribution of ORFs from SARS-CoV-2 genome to ChEMBL targets.**

## Organism\_hmmer\_threshold.R

This R function stratifies hmmer\_statistics scores for an organism (specified by tax\_id) into as many clusters as specified (the default is 2) and returns the lower bound of the highest cluster. The function selects only those targets that have drugs.

Database connection information was hard coded to the values required by the machine used for this investigation.

## **S6 Fig. Kmeans selection threshold for SARS-CoV-2. Triangular point indicates similarity threshold for best target selection.**

## target\_SARS-COV-2\_drugs.sql

## target\_SARS-CoV-2\_drugs.txt

## CHEMBL2363965\_targets\_drugs.xlsx

Spreadsheet is Table 1.

1. Student in Brandeis Univeristy GPS program

   \* Corresponding author

   E-mail: [jbsing@brandeis.edu](mailto:jbsing@brandeis.edu) (JBS) [↑](#footnote-ref-1)