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 to treat SARS-CoV-2.

# 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.

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 [4]. 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 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 for that new organism.

Genbank provides a nucleotide sequence database containing the genomes of many organisms, including the SARS-CoV-2 virus, which was downloaded (MN908947.3.FASTA).

EMBOSS tools provide the **getorf** utility that finds the genes in the SARS-CoV-2 genome. 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.**

**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 data that had been downloaded.

Once the summary information is in the database, we used statistical methods such as kmeans to choose a similarity threshold to select the most promising targets, and their drugs.

Swissdock, an opensource docking server, was used to evaluate the identified drugs for docking affinity to validate the results.

# Materials and methods

(Fig 2 .) Target and drug analytical workflow.

The genome of SARS-CoV-2 (The COVID-19 virus) was downloaded from Genbank via NCBI’s website as MN908947.3.FASTA. [5].

ORFs were translated using EMBOSS tools [6]. The orfs were translated rather than using curated protein sequences because the translation tool provides the amino acid FAST 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 MN908947.3.orf, which contains all the ORFs found for the .FASTA file.

Commands run in R Studio quantify how many ORFs are contained:

|  |
| --- |
| > 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. |

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 [7].

# Results and discussion

# 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. 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.

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

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

7. 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.

# 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.

1. Student in Brandeis Univeristy GPS program

   \* Corresponding author

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