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 organism, and drugs used against that target may be successfully used for that organism.

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

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, evaluated 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 [5].

Using **jackhmmer** to provide similarity reports and sequence alignments, a pipeline imported scores showing sequence similarity. The target scores were loaded into a PostgreSQL database that also contains the ChEMBL data [6].

Using a paralog search pipeline, the author searched the ChEMBL 25 database, screening targets in it against the SARS-CoV-2 genome and found a high scoring target that has three known drugs [4, 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. 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.

7. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B *et al*: **ChEMBL: a large-scale bioactivity database for drug discovery**. *Nucleic Acids Res* 2012, **40**(Database issue):D1100-1107.

# 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.**

1. Student in Brandeis Univeristy GPS program

   \* Corresponding author

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