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 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**

The European Molecular Biology Laboratory hosts a downloadable, curated database of drugs, targets, and their component sequences in several popular database formats known as ChEMBL [1]. An analysis pipeline searches the **chembl\_25** version of ChEMBL’s database for targets and drugs using viral genomic information from Genbank as shown in Fig 1.

**Fig 1. An analysis pipeline searches the chembl\_25 version of ChEMBL's database for targets and drugs using viral genomic information from Genbank.**

Searching the **chembl\_25** database using viral genomic information provides a way to quickly identify promising targets and drugs without the preconditions of deep understanding of pathogen behavior. This method can be used to rapidly develop leads that can be investigated as treatments for newly appearing pathogens.

*SARS-CoV-2*, also known as *COVID-19*, is a novel 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, as of October 8th, 2020, causing pandemic infection, with a fatality rate estimated between 2% and 3% [2]. Persons over 60 years of age may have much higher fatality rates [3].

A target repurposing strategy might provide drugs more quickly and cheaply than creating new drugs and finding new targets [4]. This strategy could produce treatments to ameliorate disease symptoms until a vaccine becomes available, or work in conjunction with future vaccines.

The workflow described in this paper was initially conceived as a way to rapidly and cheaply identify targets and drugs for neglected tropical diseases [4, 5]. Results from this workflow are not definitive, clinically approved therapeutic options. These methods are intended to identify targets and drugs for further study.While dosage and toxicity data are available for the drugs in question pertaining to their intended usage, 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 Virtual Machine (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 [6-10]. All necessary software is available for free, and runs on a PC with at least a 4-core 64-bit Intel compatible CPU with a minimum 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 [11]. Drug targets tend to be proteins that are important enough to the organism to which they belong that they tend to be conserved [12]. 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 [13].

European Molecular Biology Open Software Suite (EMBOSS) tools provide the **getorf** utility that finds the genes in the *SARS-CoV-2* genome[14] and extracts it into Open Reading Frames (ORFs). ORFs are amino acid sequences, including protein coding genes, that we wish to screen for sequence similarity to targets from ChEMBL.

An Extract Translate and Load (ETL) process downloads the drug target sequences into a file so that it that can be queried with sequence similarity searching software such as **jackhmmer [15]. 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 previously-downloaded **chembl\_25** database schema.

Once the summary information was loaded in the database, statistical methods such as kmeans established a similarity threshold to select the most promising targets and their drugs [16-18]. Swissdock, an open source 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 [19].

# **Materials and methods**

Figure 2 (Below) details the author’s workflow methodology.

**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 [6-8, 10].

The PostgreSQL dump archive of the ChEMBL version 25 database is downloaded, decompressed, and restored in the Centos 7 VM. EMBOSS tools are installed [14]. The genome of *SARS-CoV-2* (The COVID-19 virus) was downloaded from Genbank via NCBI’s website as MN908947.3.FASTA. [13].

ORFs were translated using EMBOSS tools [14]. 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 eliminated them from consideration.

This command created a file containing all the ORFs found and translated 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 [15]. This process created files **orf.hmm.txt** and **orf.summary**.

The Perl script **extract\_hmm\_summary.pl** read the **orf.hmm.txt** file and created the file **hmm\_stats.txt**. From the psql database query prompt, the data were imported into the **chembl\_25** database. Fig 3, below shows commands:

**Fig 3. PostgreSQL commands to import data into the chembl\_25 database.**

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**) [16] .

**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 [19].

# **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 [12]. 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 [20]. 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 [20]. At the same time, we are searching exactly for those critically necessary proteins as targets for drugs that can impair them.

This approach is a 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 is somewhat compensated for by the easy accessibility of the existing drugs, and further bolstered by our understanding of their dosage and safety statistics from existing studies [21, 22]. 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 between a curated target and drug database with a pathogen protein**

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. Figure 4 shows how sequence similarity to pathogen sequences provided a way to relate them to known targets and drugs.

**Fig 4. Query joins connecting targets, sequences, and drugs.**

Tables with blue backgrounds are supplementary tables populated by this workflow. Commands run in R Studio quantify how many ORFs are produced in results from **getorf** in Fig 5, below:

**Fig 5. Commands to quantify number of ORFs.**

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 [11]. 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** [15].

**Jackhmmer** scored similarity between amino acid sequences by aligning query and target sequences [15]. In addition, **jackhmmer** uses hidden Markoff models (HMM) that assess patterns by looking for larger domains [15]. (See Fig 6 for jackhmmer summary.)

**Fig 6. ORF FASTA summary from jackhmmer**

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 were loaded into the **hmmer\_statistics** table in the **chembl\_25** database. Figure 7 shows a histogram of the distribution of scores:

**Fig 7. Target similarity score distribution of ORFs from SARS-CoV-2 genome to ChEMBL targets.**

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)**. Figure 8 shows kmeans distribution of scores with two clusters.

**Fig 8. Kmeans selection threshold for SARS-CoV-2.**

Triangular points indicate similarity threshold for best target selection. Table 1 shows the results of this query.

**Table 1. Target and drugs retrieved.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Score** | **Original Tax\_id** | **Original\_ organism** | **Target** | **ORF** | **Pref\_name** | **Chembl\_id** |
| 9567.8 | 9606 | Homo sapiens | CHEMBL3987 | MN908947.3\_281 |  |  |
| 9562.6 | 9606 | Homo sapiens | CHEMBL5542 | MN908947.3\_281 |  |  |
| 4350.9 | 1773 | Mycobacterium tuberculosis | CHEMBL2363965 | MN908947.3\_281 | VIOMYCIN SULFATE | CHEMBL3989823 |
| 4350.9 | 1773 | Mycobacterium tuberculosis | CHEMBL2363965 | MN908947.3\_281 | CAPREOMYCIN SULFATE | CHEMBL2218913 |
| 4350.9 | 1773 | Mycobacterium tuberculosis | CHEMBL2363965 | MN908947.3\_281 | PYRAZINAMIDE | CHEMBL614 |

These three drugs include two relatives of Streptomycin, and Pyrazinamide. ChEMBL describes these as ribosome inhibitors (see results of query in Fig 9 below):

**Fig 9. PSQL query returning details for drugs and targets previously identified.**

But there is some reason to believe that some of these may have antiviral action too, especially Pyrazinamide [23]. Lian, et al, in their 2019 study showed that a multidrug regimen including Pyrazinamide reduced mortality due to hepatotoxicity in Tuberculosis patients from coinfection with Hepatitis B [23].

In the ChEMBL database, targets may have many component sequences, which makes our understanding of mechanisms less specific than they may seem at first. The following queries demonstrate that this is the case.

To find the number of protein targets in the **chembl\_25** database, execute the query in figure 10 below:

**Fig 10. PSQL query shows the number of protein targets in the chembl\_25 database.**

There are 899 protein targets in the **chembl\_25** database. Targets may have many components. How many targets in the **chembl\_25** database have only one component? (See Fig 11 below):

**Fig 11. chembl\_25 database target singletons and multi-component targets.**

There are 7354 singleton and 1014 targets with multiple components in the **chembl\_25** database.

Target CHEMBL2363965 has 60 component sequences. The following PSQL script finds the number of component sequences for the target (Fig 12.):

**Fig 12. Query shows number of component sequences for target CHEMBL2363965.**

The target component sequence of interest is the one that was found to have a high similarity score. To find what that is, we first find the ORF from the viral genome that had high similarity for this target (Fig 13):

**Fig 13. PSQL query to find ORF with high similarity to our target.**

The similarity score for target CHEMBL2363965 is high, and the query shown joins to drugs that are part of a multidrug regimen for Tuberculosis. Higher similarity scores were found with targets CHEMBL3987 and CHEMBL5542, which are single protein targets for which no drugs were found in this database.

Using NCBI BLASTP to query the Tuberculosis genome with ORF MN908947.3\_281 produces no results. This raises the possibility that these drugs are actually combatting a viral coinfection, and are not used here as ribosome inhibitors [24].

NCBI BLASTP finds a 100% match using a query excluding *SARS-CoV-2* using ORF MN908947.3\_281 (Fig 14):

**Fig 14.**  **NCBI BLASTP results for ORF MN908947.3\_281 excluding *SARS-CoV-2* [24, 25].**

This protein is highly conserved in Corona viruses. It may be that as part of treatment of Tuberculosis, a concurrent infection by a Corona virus also was treated by the drugs associated with this target. The distance tree shows relations of this protein (Fig 15):

**Fig 15. Distance tree for orf1a** [24]

Link from accession [P0C6U6.1](https://www.ncbi.nlm.nih.gov/protein/P0C6U6.1?report=genbank&log$=prottop&blast_rank=1&RID=DCW398Y4014) gets us to the structural information from the protein database:

**Fig 16. 5GWY structure of Main Protease from Human Corona Virus From Protein Database.** [26, 27]

The *Protein feature* function description comes from the RCSB protein database:

“The replicase polyprotein of coronaviruses is a multifunctional protein: it contains the activities necessary for the transcription of negative stranded RNA, leader RNA, sub genomic mRNAs and progeny virion RNA as well as proteinases responsible for the cleavage of the polyprotein into functional products [28].”

Interfering with the virus’ ability to replicate genomic RNA would be expected to slow or stop progression of infection by the virus, which makes this a good target.

## **Validating molecules with docking simulations**

Docking was simulated for the three drugs found using SwissDock [19]. Pyrazinamide had low docking affinity. This drug was not validated, and its docking information is omitted here. However, Viomycin sulphate, with a docking affinity of 10.73 and Capryomycin sulphate, with a docking affinity of 10.57 were validated as promising candidates. Results are shown below in Fig 17 and Fig 18.

**Fig 17. Predicted binding modes for RNA polymerase and Viomycin** [19].

**Fig 18. Predicted binding modes for RNA polymerase and Capryomycin** [19].

# **Conclusions**

Paralog searching the CHEMBL\_25 database with ORFs from the *SARS-CoV-2* genome has found targets and identified two promising drugs that are used in the treatment of Tuberculosis. The target identified was RNA polymerase, which is necessary for the replication of the viral RNA genome. The success of these drugs in treating Tuberculosis may be due to their effect on concurrent Corona virus infection.

Computational docking studies have found good docking affinity for two of the three drugs identified that bind to RNA polymerase: Capryomycin sulphate, and Viomycin sulphate. The third drug, Pyrazinamide, has been identified in another study citing its use as part of a multidrug regimen for Tuberculosis that uses it to treat viral co-infection [23]. These in-silico studies should be followed up by in vitro and in vivo studies to determine efficacy.

## **Acknowledgments**

I thank Eric Fischbach for his help editing this paper.

Thanks, also, to my professors Gretchen Ehrenkaufer and Alan Cheng for their advice and encouragement.

# References

1. 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.

2. **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.

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

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

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

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

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

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

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

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

11. 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.

12. 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.

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

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

15. 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.

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

17. 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.

18. 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.

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

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

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

22. 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.

23. Lian JH, Ping;Lu, Yingfeng;Liu, Yueying;Wang, XiaoXiao;Zhang, Yimin;Jia, Hongyu;Yang, Yida: **Prophylactic antiviral treatment reduces the incidence of liver failure among patients coinfected with Mycobacterium tuberculosis and hepatitis B virus**. *Elsevier* 2019, **270**.

24. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. In: *Nucleic Acids Res.* vol. 25; 1997: 3389-3402.

25. **NCBI Blast:(2) - CHEMBL2363965\_8515** [<https://www.ncbi.nlm.nih.gov/pubmed/>]

26. **RecName: Full=Replicase polyprotein 1a; Short=pp1a; AltName: Full=ORF1 - Protein - NCBI** [<https://www.ncbi.nlm.nih.gov/pubmed/>]

27. group NCS: **5GWY: Structure of Main Protease from Human Coronavirus NL63: Insights for Wide Spectrum Anti-Coronavirus Drug Design**. 2020.

28. **RCSB PDB - Protein Feature View - Replicase polyprotein 1ab - P0C6X9 (R1AB\_CVMA5)** [<http://www.rcsb.org/pdb/protein/P0C6X9>]

# **Supporting information**

## 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

Converts 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.

## 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.

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

Database query that retrieves target and drug data.

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

Target and drug information retrieved from database query.

target\_SARS-CoV-2\_drugs.xlsx

Spreadsheet is Table 1, generated from retrieved data.

1. DEPARTMENT, Brandeis University Graduate Professional Studies (GPS) program, Waltham, MA, United States of America

   \* Corresponding author

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