One Target and two drugs for SARS-CoV-2

By

Jeremy Singer

# Abstract

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

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 two known drugs [4-6]. The target that was found is an RNA polymerase from the viral genome used by the virus to reproduce.

These drugs were validated using docking simulations that showed high binding affinity to the RNA polymerase target.

# Introduction

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 [5, 6]. Drug targets tend to be proteins that are important enough to the organism to which they belong that they tend to be conserved [7].

Using **jackhmmer** to provide similarity reports, a pipeline imported scores showing sequence similarity. The target scores were loaded into a PostgreSQL database that also contains the ChEMBL data [8].

The data were analyzed, producing a report showing the target that had high similarity to the SARS-CoV-2 genome, and 3 drugs associated with that target.

The drugs were validated by simulated docking with the target by submitting queries to SwissDock, a web based docking server [9]. Two of these drugs had high affinity.

# Materials and Methods

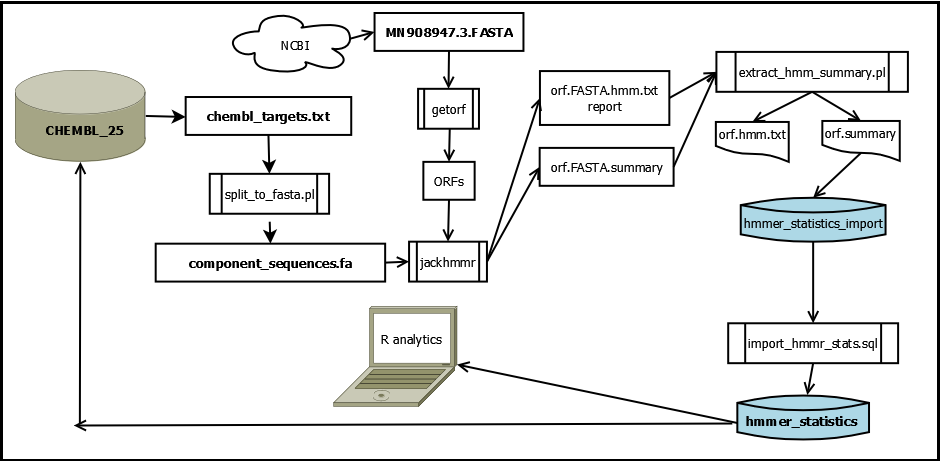


Figure 1: SARS-CoV-2 Genome analysis workflow

The nucleotide genome of SARS-CoV-2 was downloaded as MN908947.3.FASTA [10].

ORFs were translated using EMBOSS tools [11]. 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 (See 6.1. chembl\_25\_targets.sql) as file **chembl\_targets.txt**..

These targets are converted by a Perl script (See 6.2. split\_to\_fasta.pl) ) creating file **component\_sequences.fa**.

**Jackhmmer** was used to create reports and summaries of similarities with targets [8].

[osboxes@osboxes ~/genomes/MN908947.3 ] jackhmmer --domtblout orf.summary -o orf.hmm.txt mn908947.orf ~/hmmer\_targets/component\_sequences.fa

A Perl script (See 6.3. extract\_hmm\_summary.pl) extracts the similarity data from the **jackhmmer** report file, **orf.hmm.txt** and writes it to file **hmm\_stats.txt.**

From psql, 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;**

(See appendix for 6.5. import\_hmmer\_statistics.sql).

49 ORFs had enough similarity to targets to participate in our analysis.

This histogram shows the distribution of scores:

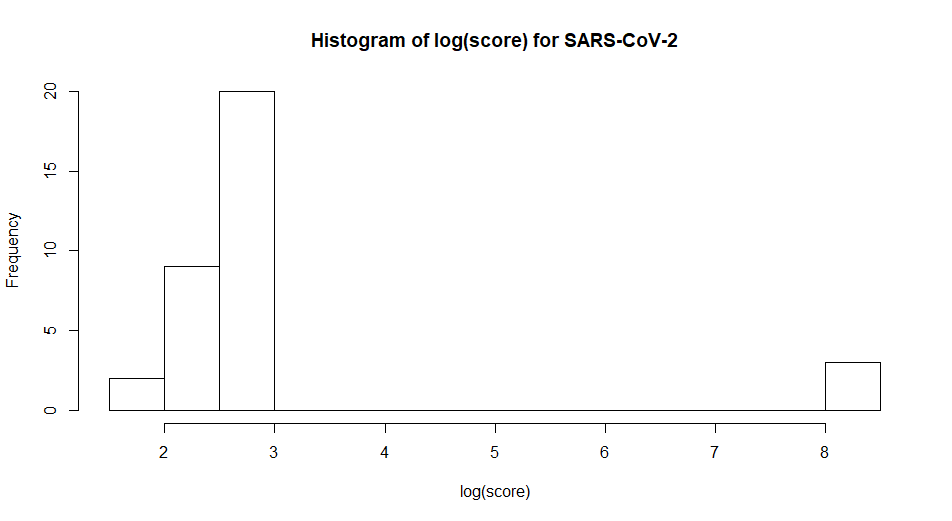


Figure 2: Target similarity score distribution of ORFs from SARS-CoV-2 genome.

Kmeans was used to identify the high scoring threshold.

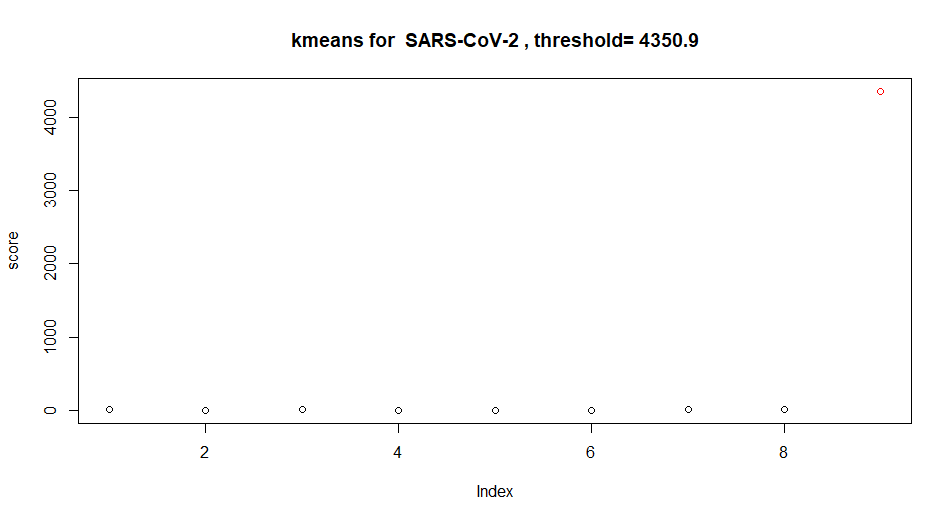


Figure 3: Kmeans showing high scoring group in red.

The threshold was used to select a report from the database identifying the target having high similarity to the SARS-CoV-2 genome, and drugs associated with it.

# Results and discussion

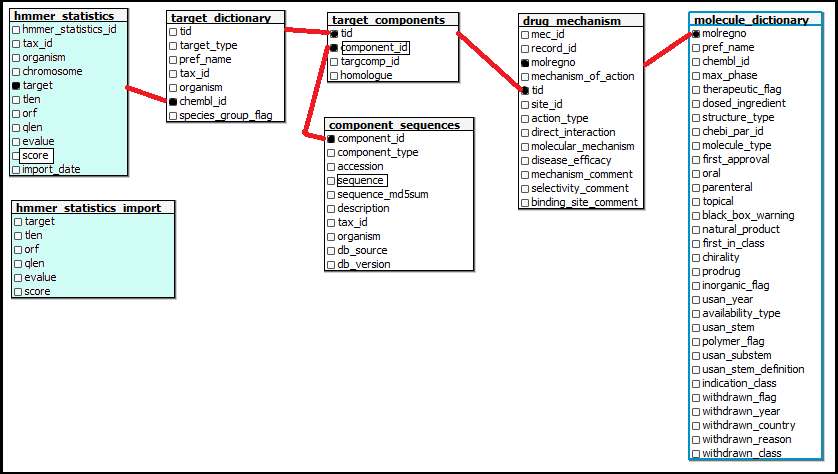


Figure : Mapping targets and drugs with sequence similarity to pathogen genomes

Tables highlighted in blue were added by the author for this project.

The ChEMBL 25 database contains a curated set of targets and molecules with relations to those targets. A table providing similarity scores between ChEMBL target sequences and pathogen ORF sequences from pathogen genomes enables mapping of targets and drugs to the pathogen.

Chembl\_id fields are unique public ids for ChEMBL entities, including targets and drugs.

**Jackhmmer** created report files (See 7.1. **orf.hmm.tx**tand 7.2. **orf.summary**) containing comparison summaries and sequence alignments between the SARS-CoV-2 genome ORFs and the ChEMBL 25 targets.

One target and three drugs were found.

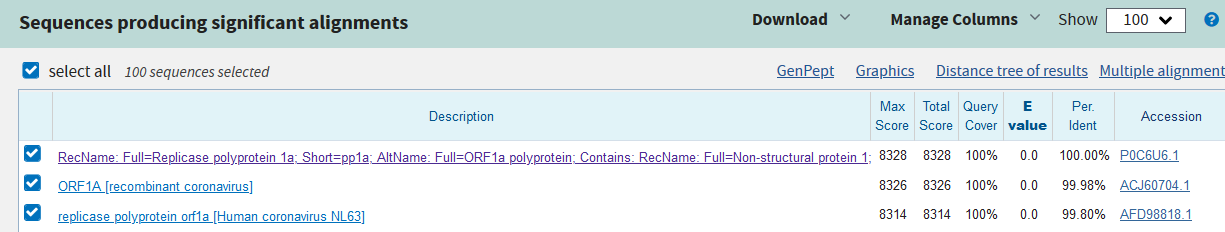
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **score** | **original tax id** | **orig\_organism** | **target\_type** | **target\_name** | **target\_chembl\_id** | **drug\_name** | **drug\_chembl\_id** |
| 4350.9 | 1773 | Mycobacterium tuberculosis | PROTEIN NUCLEIC-ACID 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 |

In the ChEMBL database, a target may have multiple sequences. CHEMBL2363965 has 59 sequences. The high scoring sequence we chose has component\_id 8515.

This high scoring match has 100% identity between the SARS-CoV-2 genome ORF that we translated and the **chembl\_25** database target that it matched. In addition, BLASTP searching the Tuberculosis genome finds no match. This is surprisising for two reasons:

1. The data in the **chembl\_25** database is from the year 2017, before SARS-CoV-2 was known.
2. It is included as part of a Tuberculosis target and matches no sequence in the Tuberculosis genome.

NCBI BLASTP finds a 100% match using a query excluding SARS-CoV-2:



[12, 13]

This protein is highly conserved in Corona viruses.

It may be that as part of treatment of Tuberculosis, a concurrent infection by a Carona virus also was treated by the drugs associated with this target.

The distance tree shows relations of this protein:

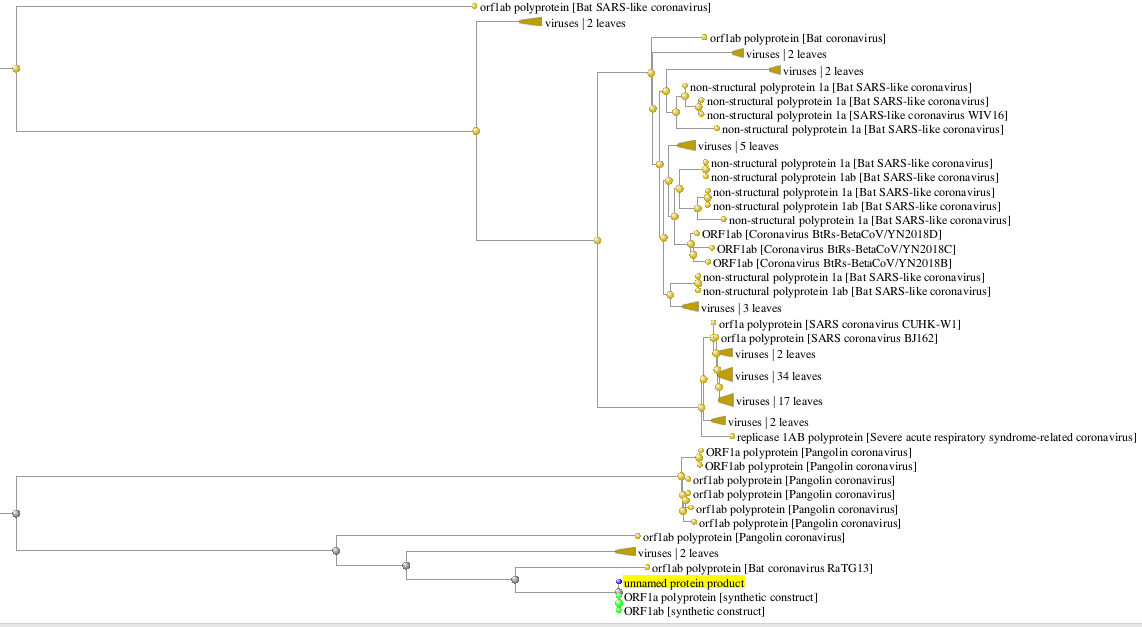
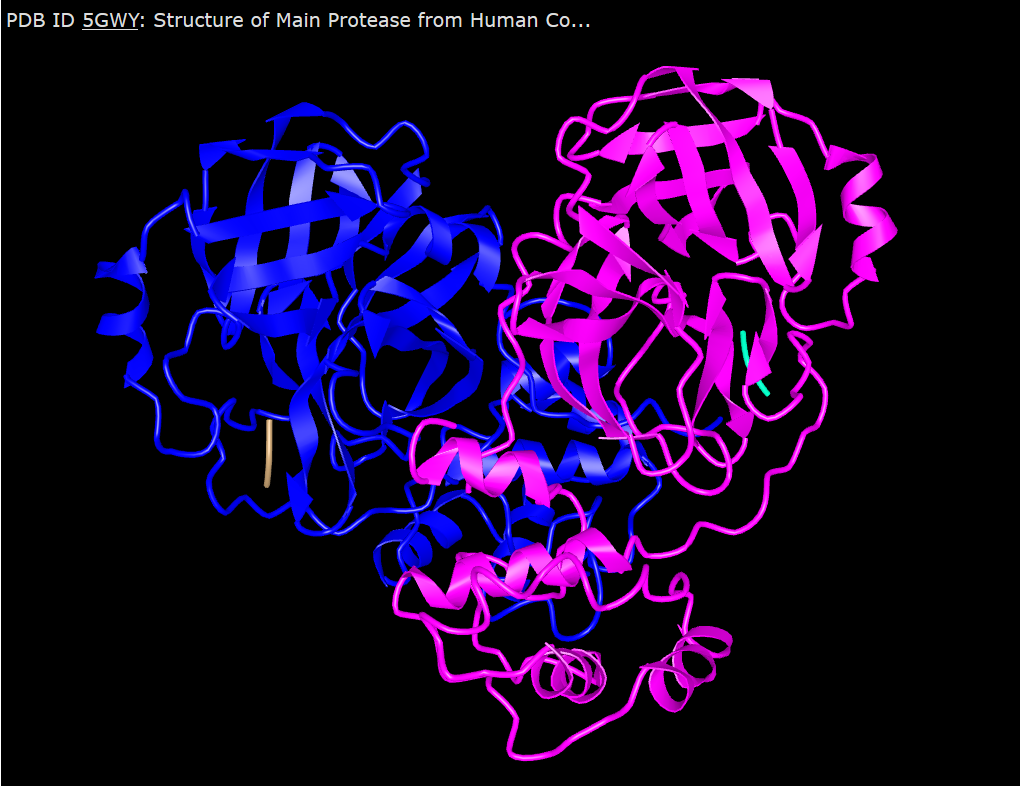


Figure 5: Distance tree for orf1a [12]

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:



[14, 15]

“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 [16].”

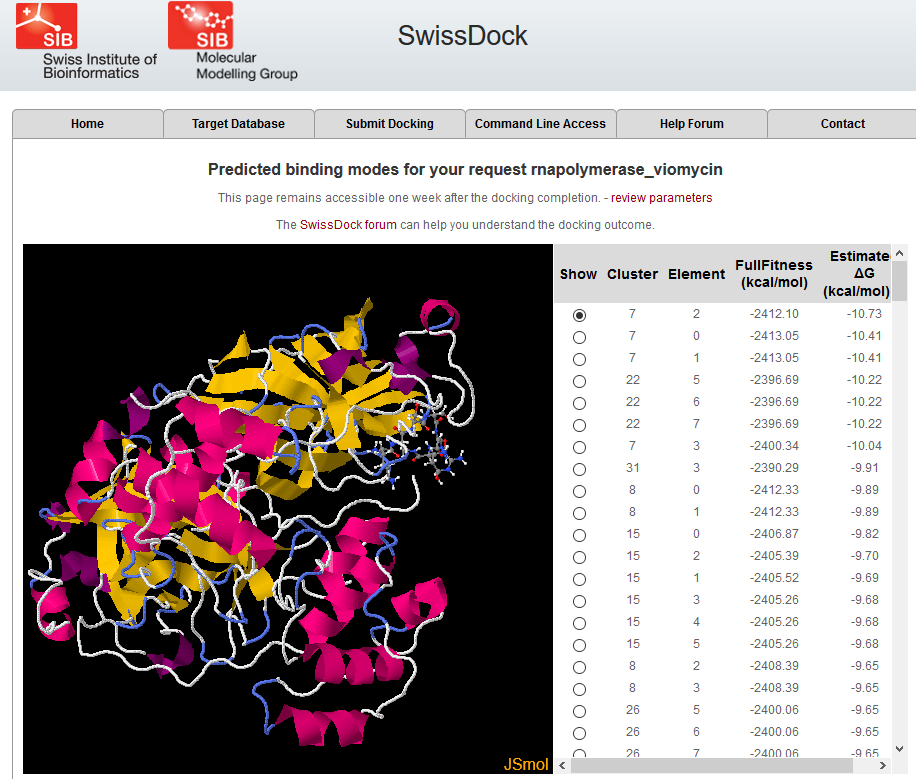
## Validating molecules with docking simulations

Docking was simulated for the three drugs found using SwissDock [9].

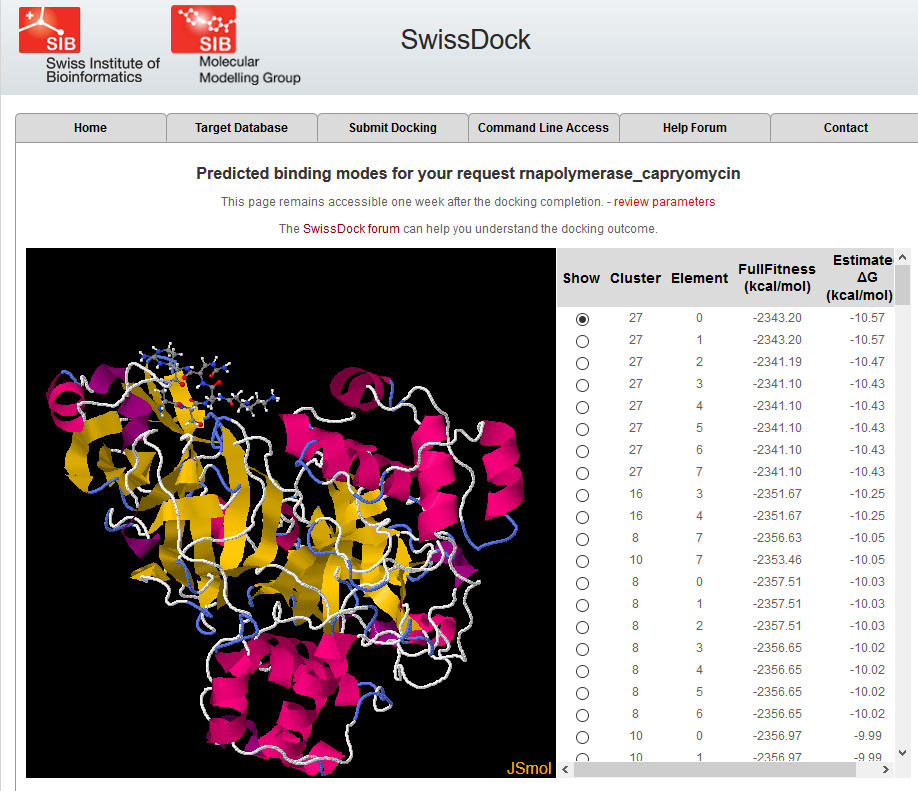
Pyrazinamide had low docking affinity. This drug was not validated, and its docking information is omitted here.

However, Viomycin sulphate, with a docking affinity of 10.73 and Capryomycin sulphate, with a docking affinity of 10.57 were validated as promissing candidates. Results are shown below.

<http://www.swissdock.ch/docking/view/swissdockd_VHEkkw_H9B11C1W2BWGOB0D7H1Y>



<http://www.swissdock.ch/docking/view/swissdockd_LUH7ua_6VCTP9T7OL7HP95N5S60>



# Conclusions

Paralog searching the CHEMBL\_25 database with ORFs from the SARS-CoV-2 genome has found a target and two promising drugs that have already been used in the treatment of Tuberculosis.

Success with these drugs in Tuberculosis treatment may have been due to their effect in combination with other drugs to quell concurrent Corona virus infection.

Computational docking studies have found good docking affinity for two of the three drugs identified: capryomycin sulfphate, and viomycin sulphate.

# Appendix

## chembl\_25\_targets.sql

Run this script at the command line of psql connected as chembl\_25 in the blast\_targets directory.

This script creates **chembl\_targets.txt** file.

\copy (select td.chembl\_id, cs.sequence from target\_dictionary td join target\_components tc on td.tid = tc.tid join component\_sequences cs on tc.targcomp\_id=cs.component\_id) to chembl\_targets.txt

## split\_to\_fasta.pl

Run this script from the *bash* command line in the **blast\_targets** directory: **perl split\_to\_fasta.pl**

This script creates the **component\_sequences.fa**file.

#######################################

# split\_to\_fasta.pl

# input recs: <key><delim><sequence>

# output : rec1 = ><key>

# rec2 = <sequence>

#######################################

my $infile = 'chembl\_targets.txt';

my $outfile = 'component\_sequences.fa';

my $delim = '\t';

open(IN, $infile) or die("Unable to open $infile\n");

my @lines = <IN>;

close(IN);

open(OUT,">",$outfile) or die ("Unable to open $outfile\n");

foreach my $line(@lines)

{

my @rec = split($delim,$line);

if (scalar(@rec) > 1)

{

print OUT ">$rec[0]\n";

print OUT "$rec[1]\n";

}

}

close(OUT);

exit(0);

## extract\_hmm\_summary.pl

(in the ~/genomes directory.)

#!/bin/perl

use Switch;

if (scalar(@ARGV) < 1) {die "No filename passed.\n";}

my $text\_fn = $ARGV[0];

my $summary\_fn;

# print $text\_fn,"\n";

$summary\_fn = $text\_fn;

$summary\_fn =~ s/.hmm.txt/.summary/;

# print $summary\_fn,"\n";

my @lines;

open($IN, "<", $summary\_fn ) or die "Can't open $summary\_fn\n";

@lines = <$IN>;

close($IN);

# print "Lines: ",scalar(@lines), "\n";

my %target;

foreach my $line(@lines){

if ( $line =~ m/^(CHEMBL\S+)\s+(\S+)\s+(\S+)\s+(\S+)\s+(\S+)\s+(\S+)\s+(\S+)\s+(\S+)\s+/ ) {

if ( ! exists $target{$1} ) { # prevent duplicate line for a target match

print $1,"\t", $3, "\t", $4,"\t",$6, "\t", $7, "\t", $8, "\n";

$target{$1} = 1;

}

}

}

## import\_hmmer\_statistics.sql

Import this script from the psql command line as user *chembl\_25*.

This script is in **~/genomes**.

truncate table hmmer\_statistics\_import;

\copy hmmer\_statistics\_import from 'hmm\_stats.txt' delimiter E'\t' CSV HEADER

insert into hmmer\_statistics

( target, tlen, orf, qlen, evalue, score)

select target, tlen, orf, qlen, evalue, score

from hmmer\_statistics\_import;

# Supplements

This paper is a sub-project of a github library made for SARS-CoV-2 under the paralog\_targets project:

<https://github.com/jeremy-b-singer/paralog_targets.git> (clone URL)

This paper can be found at:

<https://github.com/jeremy-b-singer/paralog_targets/tree/master/SARS-CoV-2>

Supplements can be found at:

<https://github.com/jeremy-b-singer/paralog_targets/tree/edits/SARS-CoV-2/supplements>

References here are to files under the **supplements** directory of this repository.

Files:

## orf.hmm.txt

Contains alignments and scores from **jackhmmer**.

## orf.summary

Consolidated statistics file created by **jackhmmer** at the same time as alignments.

## **import\_hmmer\_statistics.sql**

# 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. Singer JB: **Finding Paralog Targets for Neglected Diseases**. In*.* Edited by paralog\_targets.pdf; 2020: Methods for finding targets and drugs for a diisease caused by a pathogen searching the ChEMBL database for existing targets with protein sequences similar to ones in the pathogen genome.

5. 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.

6. 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.

7. 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.

8. 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.

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

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

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

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

13. 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.

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

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

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