Finding Paralog Targets for Neglected Diseases

By

Jeremy Singer

# Abstract

This paper describes a method that can be used to discover and repurpose existing drugs and drug targets by discovering cross species genomic sequence similarities. It uses public domain databases (ChEMBL[1], EupathDB[2]) and open source software to find measures of sequence similarity with existing targets.

This method can be applied to pathogens with at least a medium sized genome (several thousand genes.) *Neglected tropical diseases* caused by pathogenic protists are good subjects for this approach because many have genomes of sufficient size and because many have genomic features in common with organisms for which there are known targets.

The genome of the apicomplexan parasite *Plasmodium falciparum*, which is responsible for the most virulent form of malaria, was chosen to validate a method that identifies paralogs to existing disease targets because it has known cross-species targets.

ChEMBL[1] provides a PostgreSQL database that contains a list of thousands of targets and target protein sequences as well as ligands for those targets. Using this database and open source software, this paper identified *508* distinct approved drugs with their associated targets validating this approach.

6 other pathogens (Plasmodium vivax, Cryptosporidium parvum, Toxoplasma gondii, Trypanosoma brucei, Trypanosoma cruzi, Leishmania major*)* were also downloaded and run through the same pipeline, identifying potential targets and drugs.

# Introduction

Neglected tropical diseases are those diseases that affect tropical areas underserved for health care due to the poverty of those areas. These diseases affect over a billion people in over 149 countries, and damage the economies of these areas at a cost of many billions of dollars[3, p. 1].

Repurposing drugs and generating leads for finding new drugs by repurposing targets could be a cost -effective way for combating these diseases. Finding new targets can be difficult, as it requires understanding many specific details for each pathogen. A systematic method of discovering new targets that does not require this specific understanding can reduce the cost and effort of finding these targets.

This paper describes a method for drug repurposing and target repurposing based on discovering similarities between existing targets and pathogen genomes.

ChEMBL[1] 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 in that organism.

The analysis pipeline uses **BLASTP**[5]or **jackhmmer** [6] to produce similarity reports, parse the results, and upload to supplementary tables in the PostgreSQL database.

This analysis pipeline was first applied to the genome of *Plasmodium falciparum* using both BLASTP[5] and **jackhmmer**[6] to generate similarity statistics, and custom scripts included in the Appendix. The scores returned from these two different programs were compared to evaluate which could provide better discrimination criteria of useful targets and drugs.

*Plasmodium falciparum* was chosen for this evaluation because it is the most significant of these neglected diseases. In 2018, there were over 228 million cases of malaria worldwide, causing over 408 thousand deaths.[7] Emerging drug resistance to existing drugs such as choloroquin and sulfadoxine-pyrimethamine, as well as quinine increase demand for new drugs that are more effective.[8],[9]

In addition to *P. falciparum*, we processed the following additional pathogens using **jackhmmer**[6] to measure similarity statistics to targets in the CHEMBL\_25[1] database and loaded these statistics into supplementary tables in the CHEMBL\_25[1] database that we created for that purpose.

Queries using the existing ChEMBL\_25 database[1], in combination with these similarity statistics were used to identify candidate targets and drugs for each of these pathogens based on similarity thresholds computed using kmeans[10] clustering.

# Materials and Methods



Figure : Processing overview

PlasmoDB[11] and Chembl provide the data that are analyzed in this process flow.

Amino acid sequences of putative *Open Reading Frames* (ORFs) for *Plasmodium falciparum 3D7* were downloaded from *PlasmoDB.org* as file**PlasmoDB-46\_Pfalciparum3D7\_ORFs\_AA.fasta**.[12],[11]

The FASTA formatted dataset consists of all ORFs in a single file. Each ORF consists of a header line followed by a number of lines containing multiple characters of single letter codes representing an amino acid.

Header lines are formatted according to two different patterns. The first pattern encodes the ORF id that is comprised of the organism code, chromosome, and identifier. The second pattern contains a type identifier that identifies the record as belonging to the mitochondrion, and contains a unique identifier for the ORF id. A script fans out the ORF records into individual files in a directory structure having a separate subdirectory structure for each chromosome. (See script 7.2.1. fan\_out\_fasta.R ).



Figure : ORF header structure determines fan out destination

The set of target sequences comes from the ChEMBL\_25[1] PostgreSQL database and downloaded by a *psql* script (See 7.1.1. chembl\_25\_targets.sql) as file **chembl\_targets.txt**.

These targets are converted by a Perl script (See 7.1.2. split\_to\_fasta.pl) to FASTA formatted sequences (See 7.1.2. split\_to\_fasta.pl) creating file **component\_sequences.fa**.

## Gathering BLASTP statistics

The **makeblastdb** utility converts the FASTA formatted file of targets to a **BLASTP** searchable database. The command is run in director **~/blast\_targets** :

**Makeblastdb -in component\_squences.fa -out chembl\_25\_targets**

A script applies the **BLASTP** utility to each of the ORFs for the *P. falciparum* organism against the BLAST database (see 7.2.2. do\_all\_blast.sh). The script calls blast with these parameters:

**blastp -db ~/blast\_targets/chembl\_25\_targets -query $orf -num\_alignments 10 -out**

The $orf parameter is replaced in turn by each ORF in the genome to query the target database. The ${orf}.blastp.txt parameter specifies where the output of each query will go.

num\_alignments parameter specifies the maximum number of alignments that will be performed for each query. The rest of the parameters are defaulted for the program (Protein-Protein BLAST 2.10.0+).

Bash script ( 7.2.5. do\_all\_blast\_stats.sh) applied Perl script (7.2.3. extract\_header.pl) to parse each BLAST report into a *.stats* file for each ORF.

Bash script ( 7.2.4. make\_blast\_statistics.sh ) consolidated all the <ORF>.*blastp.txt.stats* files into a tab delimited text file, **blast\_statistics.txt**.

The **blast\_statistics** table, which was created previously, (7.2.6. create\_blast\_statistics\_tbl.sql), was populated by SQL script (7.2.7. import\_p\_falciparum.sql).

## Gathering HMMER statistics

Bash script **do\_all\_jackhmmer** applied the **jackhammer**[6] utility to each of the ORFs of the *P. falciparum* organism, generating a report and a summary file for each (see 7.3.1. do\_all\_jackhmmer.sh).

Bashscript ( 7.3.3. do\_all\_hmmer\_stats.sh) applied a Perl script ( 7.3.2. extract\_hmm\_summary.pl) that extracted the **jackhmmer** statistics from the reports and summaries and produced a consolidated tab delimited file for all ORFs called **hmm\_stats.txt**.

Import SQL script (7.3.5. import\_hmmer\_statistics.sql) imported these statistics into previously created tables (see 7.3.4. create\_hmmer\_stats\_tbls.sql).

Consolidated statistics records having the same ORF/target were downloaded using this join:

\copy ( select h.orf

, b.target,b.score as blast\_score

, h.score as hmmer\_score

, b.expect as blast\_expect, h.evalue

FROM blast\_statistics b

join hmmer\_statistics h

on b.orf\_id = h.orf and b.target = h.target)

to C:\Users\Jeremy-satellite\Documents\RBIF120\consolidated\_stats.txt CSV delimiter ' ‘

# Discussion and Results

The “targetness” of a protein has to do with how indispensable its function is to the disease organism in question. 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. 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, we do not know 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. At the same time, we are searching exactly for those critically necessary proteins as targets for drugs that can impair them.

Our uncertainty about whether existing drugs will effectively bind or interfere with the target proteins we identify are somewhat compensated for by the improvements of convenience and cost from the availability of the existing drugs and understanding of their dosage and safety from existing studies.

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.

Both **BLASTP**[5]and **jackhmmer**[6] score similarity between amino acid sequences by aligning query and target sequences. Their approach to scoring, however, differs. **BLASTP**[5]detects similarities by scoring the likelihood of successive letters of the amino acid being the same in query and target sequences in comparison to chance. By contrast, **jackhmmer**[6] uses hidden Markoff models (HMM) that assess patterns by looking for larger domains. We might therefore expect that this approach might be more discriminating, as it takes into account similarities at the level of protein domains. We examined the differences in results between **BLASTP**[5] and **jackhmmer**[6] to assess which approach would measure similarity in a way that would best identify the conservation we are looking for. We hypothesize that these similarity scores can be stratified into a group whose similarity is due only to evolutionary relationship, and a second group that is much more similar due to conserved similarity due to necessity to maintain conserved function.



Figure : Understanding BLAST statistics

**BLASTP**[5] produces alignment reports containing the similarity statistics we are looking for (See 6.2.2. do\_all\_blast.sh). We extracted these in an ETL process and imported them into the **CHEMBL\_25** database. (See 6.2.3. extract\_header.pl, 6.2.4. make\_blast\_statistics.sh )

Figure : ORF.FASTA.summary from jackhmmer

The summary for a query may hit multiple targets. Each target record is repeated for each domain that **jackhmmer**[6] matches. We are only interested in the total score for the full sequence. The Perl script (7.3.2. extract\_hmm\_summary.pl) creates a single record per ORF/target by de-duping these values. Score is used in a similar way to BLAST, but is computed differently.

## Choosing appropriate metrics for selecting target candidates

**BLASTP**[5] and **jackhmmer**[6] also compute *expected* values, which can be very small numbers. In contrast, the *scores* values are always integers that are easily comparable.

Other statistics computed by BLASTP[5] are *identities*, which are the percentage of exact matches, and *positives*, which are inexact matches that conserve function because the amino acids involved in the comparison function in a compatible manner to each other in the protein.



Figure : Pairwise comparison of BLAST metrics

An advantage of the *score* statistic is that it is additive. While both statistics reflect the cumulative values for matches, products of probability scores that accumulate in the *expect* statistic can lose resolution during computation. Although *expect* should follow (inversely) to score, the comparison here is uninformative because of loss of resolution at low values. *Positives* trend like *identities* but show *greater than or equal*  relationship to them. *Score* also trends with *positives,*  showing greater detail.

Differences between the comparison of *scores* by *identities* and *expect* by *identities* varies substantially due to the imprecision caused by computations using *expect.*

## Comparison of BLASTP and jackhmmer scores

The default inclusion threshold for **jackhmmer**[6] is much more stringent than the default threshold for **BLASTP**[5]. The **blast\_statistics** table imported 562,039 records, where the **hmmer\_statistics** table contains only 127,306 records.

A query shows that there are 20,178 **blast\_statistics** records that join with the **hmmer\_statistics** records with the same OFR/target. A tab delimited file named **consolidated\_stats.txt** was downloaded with these records (see 6.4.1. consolidated\_orf\_target.sql).



Figure : BLASTP vs HMM scoring comparison

The figure shows that the two kinds of scores generally trend the same, but there are many excursions for *hmm\_score* that are much greater than the *blast score* would lead one to expect. (See 6.4.2. compare\_scores.R) Thismay be attributed to **jackhmmers’**s [6] increased sensitivity to structure over **BLASTP**[5].

The scale of *hmmer\_score*to *blast\_score* is **0.5523**.We can use this later to estimate the appropriate selection threshold for *hmmer* statistics.

While the **jackhmmer**[6]statistics may provide us with a more accurate scoring of the similarity of parasite ORFs with our target universe, we can use the broader **BLASTP**[5] statistics to provide statistical guidelines for selecting a significance threshold. By pairing these two sets of statistics, *blast statistics* can help us inform useful methods for selection from *jackhmmer statistics*.

## How departures from normality can reveal potential targets



Figure : Histogram of log(score) for P. falciparum 3D7

Histogram of log BLASTP[5] scores, showing superimposed normal plot in red.

Vertical blue lines show thresholds that are 1, 2, 3, 4, and 5 median absolute deviations from the median.

Normal curve was calculated using **R dnorm** function[13], which expects a standard deviation parameter, corrected here by a factor using **mad** function. Log scale was chosen to improve display of the curve.

Choosing a match metric and threshold is a conundrum. What determines whether a match is sufficiently close to suggest that it is likely to be a target?

While a very close match of the *positives* or *identities* (100% or 99%) would indicate that a match identifies a target by tautology, lower scores are hard to justify. A small sequence with a fairly high percentage may still not be as likely a match as a much larger string with a lower *positives* score and a much higher *score* statistic, as the cumulative score takes into account the greater difficulty of achieving a high score for the larger string.

The distribution of log(*score)* for all the matches in *P. falciparum*  appears to be somewhat normal.

There are several reasons why we should expect these results.

1. The distribution is not symmetric around the mean because BLASTP discards matches where the score is too low, and only returns up to the ten highest scores per query.
2. Normally distributed matches are what we would expect due to random mutation for non-conserved sequences, and account for the normal appearance of the graph.
3. Due to conservation of essential peptides, there may be more high scoring matches than predicted by normally distributed processes. We can look for these likely target candidates in the area of the distribution which normally should be approaching zero asymptotically.
4. In addition, the target universe may be “lumpy” in the sense that there may be more than one centroid that has a family of similarities in the *P. falciparum* genome. This “lumpiness” is not important for purposes of finding unusual target similarity, as it is buried in a normal cohort.

It is interesting to note that the smallest *score* value for a comparison that has a 100% positives value is 21.6, which is close to the median of 23.9 ( or log values of 3.07 and 3.17 respectively). Searching the upper tail means missing some targets that are closer to the median.

Median and Median Absolute Deviation were used as measures of centrality and dispersion rather than mean and standard deviation because these measures are more appropriate considering the asymmetry and contamination of the distribution (in the sense that the distribution contains more than a single normal distribution plus other non-normally distributed data.)



Figure :QQ Plot for all P. falciparum BLASTP log(scores)

The red line shows expected normality. This plot shows departure from normality at the right (upper) portion of the graph.



Figure : normal subset of log BLAST stats for P. falciparum

We chose a threshold of 2 *maximum absolute deviations* above the median as a cutoff below which the similarity results are *normal* and therefore *insignificant.* The normal score threshold was 27.16, and the log score threshold was 3.3073. We expect that scores above that will be more than randomly significant.

Applying the Q-Q plot to the *hmmer\_score* values does not provide as neat an answer:



Figure : HMMER scores are less well behaved

The Q-Q plot shows a pronounced *jink* in the middle of the graph. This suggests that the HMMER scores may be easier to stratify into different similarity classes.

## Using kmeans to find a significant cluster

Another approach to find a useful threshold is **kmeans** clustering[14],[10],[15]. We are looking for a region that is less dense, with higher scores indicating unusually high similarity. The **kmeans**[10] functions uses a machine learning algorithm that classifies data by minimizing within-cluster sum of squares distance. We used it to classify our data into two clusters.



Figure : kmeans clustering of hmmer\_score

The purple line shows the threshold separating less similar insignificant similarity (in red) from highly similar ORFs in black. The similarity threshold found by **kmeans**[10] is more stringent than the one we found using *median* and *maximum absolute deviation*.

(See 6.4.3. Score normality and kmeans analysis)

1708 scores out of 20178 belong to the significant cluster.

Note that the significant cluster is much more dispersed than the insignificant cluster. The significance threshold should show a reasonable amount of separation in these clusters.

## Downloading drugs and targets

Having determined a significance threshold, we are ready to select drugs and targets from the *CHEMBL\_25* database. The query specifies tax\_id to find targets for the associated organism.

The score in this query was determined from kmeans, as described (See 6.4.3. Score normality and kmeans analysis).

select distinct h.target, score, td.tax\_id as original\_tax\_id, td.organism as orig\_organism,td.pref\_name, md.pref\_name, md.chembl\_id

from hmmer\_statistics h

join target\_dictionary td

on h.target = td.chembl\_id

join drug\_mechanism dm

ON dm.tid = td.tid

join molecule\_dictionary md

ON dm.molregno = md.molregno

WHERE h.tax\_id = 36329

and score >= 385.3

order by score desc;

This query was reformatted as a **psql** command that downloads a tab delimited text file (See 6.5. Download P. falciparum drugs and targets).

The tab delimited file was imported into excel, providing a filterable document. This is useful, if we wish to exclude certain original organisms (9606 is human.) While the query downloads the data in descending score order, a user can sort this document by other columns, such as the drug *pref\_name*.

**P\_falciparum\_hmmer\_drugs.xslx** is included separately in the supplements.

At least one line for each target/drug with score from **jackhmmer** with document references, in pref\_name / journal year order. (See 6.6. Download P. falciparum drugs and targets, with annotations)

File **P\_falciparum\_hmmer\_drugs\_annotated.xslx** is included separately in the supplements.

## Validating drugs found

848 drugs/targets were found based on similarity criterion identified by **kmeans**[10] threshold of 385.3.

Of the drugs found, 2 are already known in the *chembl\_25* database for P. falciparum:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **avg\_score** | **original\_tax\_id** | **orig\_organism** | **pref\_name** | **chembl\_id** | **mechanism\_of\_action** | **max\_phase** | **first\_approval** |
| 661.15 | 5833 | Plasmodium falciparum | SULFACYTINE | CHEMBL1201056 | Dihydropteroate synthetase inhibitor | 4 | 1975 |
| 661.15 | 5833 | Plasmodium falciparum | SULFADOXINE | CHEMBL1539 | Dihydropteroate synthetase inhibitor | 4 | 1981 |

Of the remaining drugs found, 550 have gone through phase 4 clinical trials.

506 of those drugs have first approval dates.

Azithromycin also is in clinical trials for treatment of *P. falciparum* malaria.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **avg\_score** | **original\_tax\_id** | **orig\_organism** | **pref\_name** | **chembl\_id** | **mechanism\_of\_action** | **max\_phase** | **first\_approval** |
| 1063.857143 | 2 | Bacteria | AZITHROMYCIN | CHEMBL529 | Bacterial 70S ribosome inhibitor | 4 | 1991 |

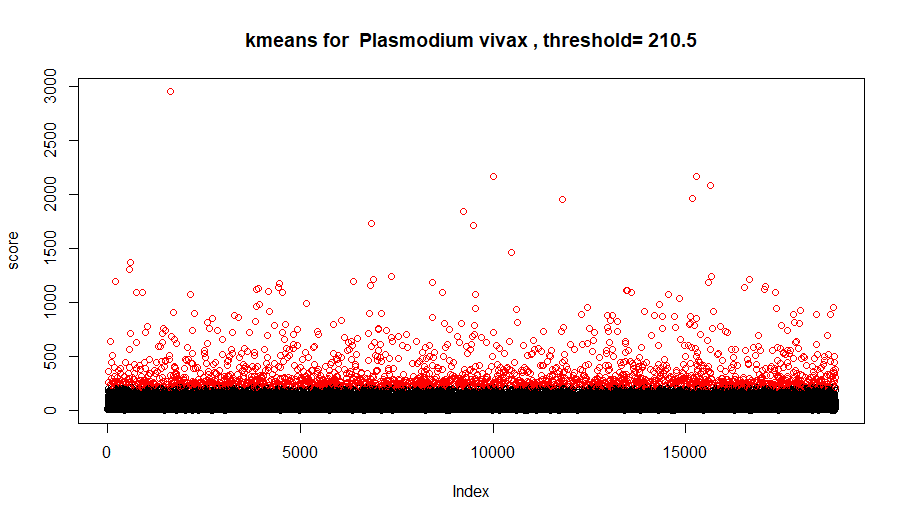
102 drugs are anti-bacterial. Of these, 41 are Bacterial 70S ribosome inhibitors; the rest are Bacterial penicillin-binding protein inhibitors.

## Drugs for other organisms

### Plasmodium Vivax

Search was based on **PlasmoDB-46\_Pfalciparum3D7\_ORFs\_AA.fasta**.[16]

Kmeans threshold computed using (6.7. organism\_hmmer\_threshold.R). tax\_id = 5855.



Drugs were downloaded by (6.8. P\_vivax\_jackhmmer\_drugs.sql ).

Spreadsheets of results are contained in the supplements as **P\_vivax\_hmmer\_drugs.xlsx** and **P\_vivax\_hmmer\_drugs\_annotated.xlsx**.

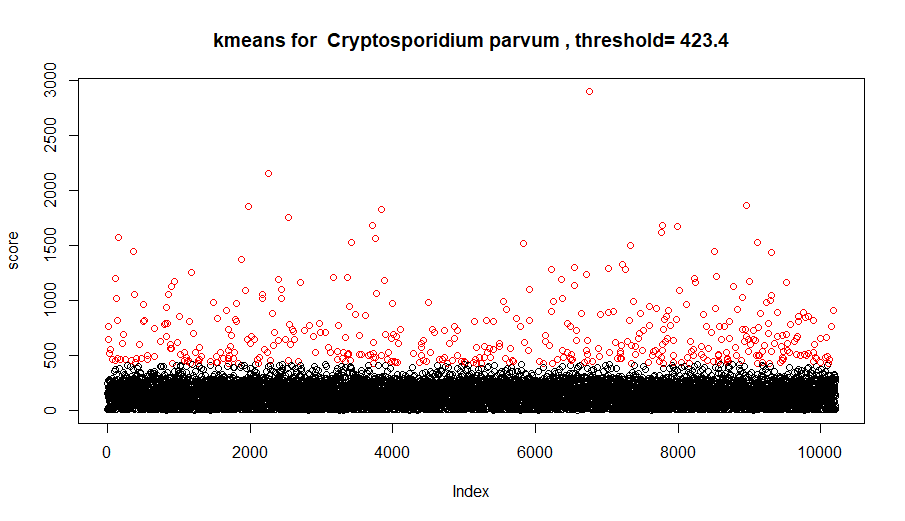
1234 putative drugs/targets were identified.

### Cryptosporidium parvum

Search was based on **CryptoDB-46\_CparvumIOWA-ATCC\_ORFs\_AA.fasta.**[17]

Kmeans threshold computed using (6.7. organism\_hmmer\_threshold.R). tax\_id = 5807.

Threshold is 423.4.



Spreadsheets of results are contained in the supplements as **C\_parvum\_hmmer\_drugs.xlsx** an**d C\_parvum\_hmmer\_drugs\_annotated.xlsx**.

### Trypanosoma cruzi Brazil A4

Search was based on **TriTrypDB-46\_TcruziBrazilA4\_ORFs\_AA.fasta.**[18]

This file has 1,707,427 consisting of 528,196 ORFs.

The **Contig** directory created by **fan\_out\_fasta\_tryp.R** contains too many files to be handled by bash **ls**  command:

[osboxes@osboxes ~/genomes/TcBrA4 ] ls -l Contig/\*

-bash: /usr/bin/ls: Argument list too long

Using Perl, we find that the number of files in the **Contig** directory:

main::(-e:1): 4

DB<1> **@files=glob("\*.\*");**

DB<2> **print(scalar(@files));**

108582

Two bash scripts were rewritten as Perl scripts to overcome these limitations (See 6.3.2. do\_all\_jackhmmer.pl, 6.3.5. do\_all\_hmmer\_stats.pl).

These were applied against organism directory TcBrA4:

**perl do\_all\_jackhmmer.pl TcBrA4/**

and

**perl do\_all\_hmmer\_statistics.pl TcBrA4/**

After these scripts completed:

**sudo su postgres**

**psql -U postgres -d chembl\_25**

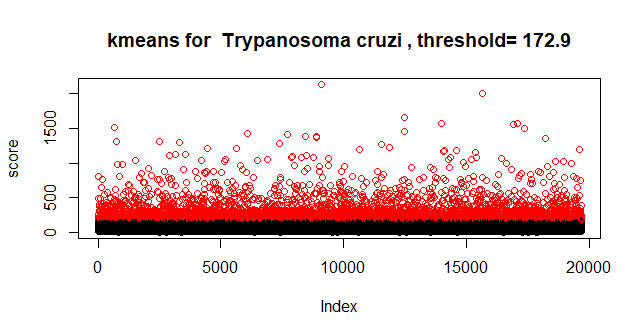
**\i import\_hmmer\_statistics.sql**

TRUNCATE TABLE

INSERT 0 144447

chembl\_25=#

update hmmer\_statistics set tax\_id = 5693, organism = 'Trypanosoma cruzi' where tax\_id is null;



Spreadsheets of results are contained in the supplements as **T\_cruzi\_hmmer\_drugs.xlsx** and **T\_cruzi\_hmmer\_drugs\_annotated.xlsx**.

### T. brucei gambiense DAL972

Search was based on **TriTrypDB-46\_TbruceigambienseDAL972\_ORFs\_AA.fasta.**[19]**,**[18]

**perl do\_all\_jackhmmer.pl Tbg972/**

**perl do\_all\_hmmer\_stats.pl Tbg972/**

[osboxes@osboxes ~/genomes ] **sudo su postgres**

[sudo] password for osboxes:

[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**

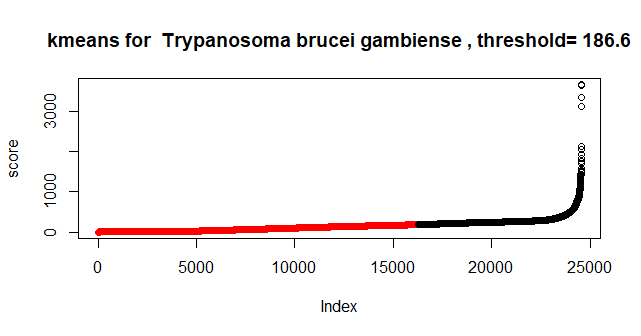
TRUNCATE TABLE

INSERT 0 179183

chembl\_25=# update hmmer\_statistics set tax\_id = 31285,organism = 'Trypanosoma brucei gambiense' where tax\_id is null;

UPDATE 179183

chembl\_25=#



This kmeans threshold does not seem to be partitioning the data in a convincing way.

# References

[1] A. Gaulton *et al.*, “ChEMBL: a large-scale bioactivity database for drug discovery,” *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D1100–D1107, Jan. 2012, doi: 10.1093/nar/gkr777.

[2] C. Aurrecoechea *et al.*, “EuPathDB: the eukaryotic pathogen genomics database resource,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D581–D591, 04 2017, doi: 10.1093/nar/gkw1105.

[3] “WHO | World Health Organization,” *WHO*. [Online]. Available: http://www.who.int/neglected\_diseases/diseases/en/. [Accessed: 12-Feb-2020].

[4] A. Gaulton *et al.*, “The ChEMBL database in 2017,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D945–D954, Jan. 2017, doi: 10.1093/nar/gkw1074.

[5] S. F. Altschul *et al.*, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997, doi: 10.1093/nar/25.17.3389.

[6] T. J. Wheeler and S. R. Eddy, “nhmmer: DNA homology search with profile HMMs,” *Bioinformatics*, vol. 29, no. 19, pp. 2487–2489, Oct. 2013, doi: 10.1093/bioinformatics/btt403.

[7] “World malaria report 2019.” [Online]. Available: https://www.who.int/news-room/feature-stories/detail/world-malaria-report-2019. [Accessed: 12-Feb-2020].

[8] M. A. Khan, R. A. Smego, S. T. Razi, and M. A. Beg, “Emerging drug--resistance and guidelines for treatment of malaria,” *J. Coll. Physicians Surg.--Pak. JCPSP*, vol. 14, no. 5, pp. 319–324, May 2004, doi: 05.2004/JCPSP.319324.

[9] “WHO | Responding to antimalarial drug resistance,” *WHO*. [Online]. Available: http://www.who.int/malaria/areas/drug\_resistance/overview/en/. [Accessed: 12-Feb-2020].

[10] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A K-Means Clustering Algorithm,” *J. R. Stat. Soc. Ser. C Appl. Stat.*, vol. 28, no. 1, pp. 100–108, 1979, doi: 10.2307/2346830.

[11] “PlasmoDB Download Files.” [Online]. Available: https://plasmodb.org/common/downloads/Current\_Release/Pfalciparum3D7/fasta/data/. [Accessed: 12-Feb-2020].

[12] C. Aurrecoechea *et al.*, “PlasmoDB: a functional genomic database for malaria parasites,” *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D539-543, Jan. 2009, doi: 10.1093/nar/gkn814.

[13] “Lognormal function | R Documentation.” [Online]. Available: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/Lognormal. [Accessed: 10-Mar-2020].

[14] J. Macqueen, “SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS,” *Multivar. Obs.*, p. 17.

[15] “forgy: Initialization of cluster prototypes using Forgy’s algorithm in inaparc: Initialization Algorithms for Partitioning Cluster Analysis.” [Online]. Available: https://rdrr.io/cran/inaparc/man/forgy.html. [Accessed: 03-Mar-2020].

[16] J. M. Carlton *et al.*, “Comparative genomics of the neglected human malaria parasite Plasmodium vivax,” *Nature*, vol. 455, no. 7214, pp. 757–763, Oct. 2008, doi: 10.1038/nature07327.

[17] M. Heiges *et al.*, “CryptoDB: a Cryptosporidium bioinformatics resource update,” *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D419-422, Jan. 2006, doi: 10.1093/nar/gkj078.

[18] M. Aslett *et al.*, “TriTrypDB: a functional genomic resource for the Trypanosomatidae,” *Nucleic Acids Res.*, vol. 38, no. suppl\_1, pp. D457–D462, Jan. 2010, doi: 10.1093/nar/gkp851.

[19] “Data Set GeneDB: The Sanger Institute Pathogen Genomics Database.” [Online]. Available: https://tritrypdb.org/tritrypdb/app/record/dataset/DS\_365c388131. [Accessed: 10-Mar-2020].

# Appendix: Scripts

These scripts can be found in the **supplements** directory for this **git** repository.

## BLAST Targets

### chembl\_25\_targets.sql

Run this script at the command line of psql attached 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 which can be found in the **supplements**/**blast\_targets** directory.

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

# 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);

## Process FASTA

### fan\_out\_fasta.R

This script is run within RStudio to fan out the single FASTA file from *Plasmodb* into separate directories by organism and chromosome. Each ORF is separated out for ease of obtaining BLAST and HMMER reports for each.

# Fan out AA\_fasta file from plasmodbc

# based on the structure of Plasmodium AA\_orf files.

# FASTA headers come in two varieties:

# 1. >Pf3D7\_01\_v3-1-60871-61059 | organism=Plasmodium\_falciparum\_3D7 | location=Pf3D7\_01\_v3:60871-61059(+) | length=63 | sequence\_SO=chromosome

# ^unique ORF identifier---^ <other stuff> <sequence\_SO=<ORF type> i.e. chromosome, apicoplast, mitochondrial

# ^head indicator

# ^organism

# ^chromosome id

# ^orf\_name

# 2. >Pf\_M76611-5-344-75 | organism=Plasmodium\_falciparum\_3D7 | location=Pf\_M76611:75-344(-) | length=90 | sequence\_SO=mitochondrial\_chromosome

# ^orfname---------^ <other stuff> sequence\_SO=mitochondrial\_chromosome

# parsing strategy is: for non-mitochondrial, parse out chromosome\_name, orf\_name.

# For mitochondrial, orfname is one piece.

library(stringr)

setwd('~/genomes')

aa\_file=file.choose()

aa=read.table(file=aa\_file,header = FALSE, sep='~', stringsAsFactors = FALSE)

aa=aa[!is.na(aa[,1]),] # filter out NA

firstrec=aa[1] # scalar

aa=data.frame(lines=aa, stringsAsFactors = FALSE)

parsed=strsplit(firstrec,'\_')

organism\_pref=substring(parsed[[1]][1],2)

# make a directory for this organism

system(paste('mkdir',organism\_pref))

orf\_headers=aa[substr(aa[,1],1,1)=='>' ,]

mi\_headers=orf\_headers[grep('sequence\_SO=mitochondrial\_chromosome',orf\_headers)]

chrom\_headers = setdiff(orf\_headers, mi\_headers)

parsed=strsplit(chrom\_headers,'\_')

chromosomes=unique(sapply(parsed,function(p){p[2]}))

# make a directory for each chromosome

for (chromosome in chromosomes){

dirname=paste(organism\_pref,chromosome,sep='/')

system(paste('mkdir',dirname))

}

dirname=paste(organism\_pref,'mitochondrion', sep='/')

system(paste('mkdir',dirname))

orf.df = data.frame(line='')

orf\_name=''

orf.df=data.frame(line='')

for(orf\_line in aa[,1]){

print(paste('orf\_line: ',orf\_line))

if (substr(orf\_line,1,1)=='>'){

print('FASTA header line')

if ( is.na(orf\_name) || nchar(orf\_name) > 0){

orf\_name=paste0(orf\_name,'.FASTA')

print("write statement")

write\_dir\_name = paste(organism\_pref,chromosome, orf\_name,sep='/')

write.table(orf.df, file=write\_dir\_name,row.names = FALSE,col.names = FALSE, quote=FALSE)

}

orf.df = data.frame(line=orf\_line)

if ( length(grep('mitochondrial',orf\_line)) > 0){

chromosome='mitochondrion'

print(paste("Chromosome:", chromosome))

parsed=unlist(strsplit(orf\_line,' '))

orf\_name=substr(parsed[1],2,nchar(parsed[1]) -1)

} else {

parsed=unlist(strsplit(orf\_line,' '))

parsed=unlist(strsplit(parsed[1],'\_'))

chromosome=parsed[2]

orf\_name=parsed[3]

}

print(paste('chromosome:',chromosome,', orf\_name:', orf\_name))

} else {

print('rbind FASTA sequence')

orf\_line.df=data.frame(line=orf\_line)

orf.df = rbind(orf.df, orf\_line.df);

}

}

if (is.na(orf\_name) || nchar(orf\_name) > 0){

print("write statement")

write\_dir\_name = paste(organism\_pref,chromosome, orf\_name,sep='/')

write.table(orf.df, file=write\_dir\_name,row.names = FALSE,col.names = FALSE,quote=FALSE)

}

### do\_all\_blast.sh

Run this script in the genome directory.

Specify the *Organism\_dir* on the command line.

#!/bin/bash

if [ -z $1 ]

then

while [ -z $org\_dir ]

do

read -p "Organism directory: " -a org\_dir

done

else

org\_dir=$1

fi

echo $org\_dir

for chrom\_dir in $( ls -d $org\_dir\*/ );do

for orf in $( ls $chrom\_dir\*.FASTA);do

echo "BLASTP " $orf

blastp -db ~/blast\_targets/chembl\_25\_targets -query $orf -num\_alignments 10 -out ${orf}.blastp.txt

done

done

### extract\_header.pl

This Perl script extracts statistics from BLAST reports.

use Switch 'fallthough';

my @lines = <STDIN>;

my $phase = 0;

my @rec = {};

my $rec\_string;

my %recs = ();

my $query;

foreach my $line(@lines)

{

switch($phase){

case 0 {

if ( $line =~ m/Query=\s\*(\S+)/)

{

$query = $1;

}

if ( $line =~ m/\>\s\*(\S+)/) # orf id

{

$phase = 1;

$rec[0] = $1;

}

}

case 1 {

if ( $line =~ m/Length\=(\S+)/ ){

$rec[scalar(@rec)] = $1;

$phase = 2;

}

}

case 2 {

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

$rec[scalar(@rec)] = $1;

$line =~ m/Expect\s\=\s(\S+),/;

$rec[scalar(@rec)] = $1;

$phase = 3;

}

}

case 3 {

if ( $line =~ m/Identities\s\=\s\S+\s\((\S+)\%/){

$rec[scalar(@rec)] = $1;

$line =~ m/Positives\s\=\s\S+\s\((\S+)\%/;

$rec[scalar(@rec)] = $1;

$line =~ m/Gaps\s\=\s\S+\s\((\S+)\%\)/;

$rec[scalar(@rec)] = $1;

$rec\_string = join("\t",@rec);

$recs{$rec\_string} = 1;

$phase = 0;

@rec = {};

}

}

}

}

foreach my $record(keys %recs){

print "$query\t$record\n";

}

### make\_blast\_statistics.sh

Create the blast\_statistics file by concatenating all the \*.blast.stat files.

#!/bin/bash

if [ -z $1 ]

then

while [ -z $org\_dir ]

do

read -p "Organism directory: " -a org\_dir

done

else

org\_dir=$1

fi

echo $org\_dir

echo "orf\_id target query\_length score expect identities positives gaps" > blast\_statistics.txt

for chrom\_dir in $( ls -d $org\_dir\*/ );do

cat $( ls $chrom\_dir\*.blastp.txt.stats) >> blast\_statistics.txt

done

### do\_all\_blast\_stats.sh

Apply the Perl script (**extract\_header.pl**) that extracts statistics to all the BLAST reports.

#!/bin/bash

if [ -z $1 ]

then

while [ -z $org\_dir ]

do

read -p "Organism directory: " -a org\_dir

done

else

org\_dir=$1

fi

echo $org\_dir

for chrom\_dir in $( ls -d $org\_dir\*/ );do

for orf in $( ls $chrom\_dir\*.blastp.txt);do

echo "BLAST stats " $orf

perl ~/genomes/extract\_header.pl < ${orf} > ${orf}.stats

done

done

### create\_blast\_statistics\_tbl.sql

Enter this at the psql command line:

CREATE TABLE blast\_statistics

(

sk\_blast\_statistics SERIAL -- synthetic primary key

, tax\_id bigint -- NCBI taxonomy id of target

, organism character varying(100) -- convenience name of organism

, chromosome character varying(50)

, orf\_id character varying(50)

, target character varying(50) -- typically, chembl\_id

, query\_length int

, score numeric

, expect numeric

, identities numeric

, positives numeric

, gaps numeric

, import\_date timestamp not null default clock\_timestamp()

);

CREATE TABLE blast\_statistics\_import

(

orf\_id character varying(50)

, target character varying(50)

, query\_length int

, score numeric

, expect numeric

, identities numeric

, positives numeric

, gaps numeric

);

### import\_p\_falciparum.sql

(run this at psql prompt logged in as chembl\_25:

truncate table blast\_statistics\_import;

\copy blast\_statistics\_import from 'blast\_statistics.txt' delimiter E'\t' CSV HEADER

insert into blast\_statistics

( tax\_id, organism, orf\_id, target, query\_length, score, expect, identities, positives, gaps)

SELECT 36329 -- tax\_id

, 'Plasmodium falciparum 3D7'

, orf\_id

, target

, query\_length

, score

, expect

, identities

, positives

, gaps

FROM blast\_statistics\_import;

## HMM targets

### do\_all\_jackhmmer.sh (deprecated)

#!/bin/bash

if [ -z $1 ]

then

while [ -z $org\_dir ]

do

read -p "Organism directory: " -a org\_dir

done

else

org\_dir=$1

fi

echo $org\_dir

for chrom\_dir in $( ls -d $org\_dir\*/ );do

for orf in $( ls $chrom\_dir\*.FASTA);do

echo "jackhmmer " $orf

jackhmmer --domtblout $orf.summary -o $orf.hmm.txt $orf ~/hmmer\_targets/component\_sequences.fa

done

done

### do\_all\_jackhmmer.pl

(replaces do\_all\_jackhmmer.sh)

#!/usr/bin/perl

# do\_all\_jackhmmer.pl

# Applies jackhmmer to .FASTA files in chromosome directories under <organism directory>

#

if (scalar(@ARGV) < 1) { die ("Specify organism directory\n"); }

my $org\_dir = pop(@ARGV);

if ( !( -e $org\_dir and -d $org\_dir ) ) {

die "$org\_dir is not a directory\n";

}

my @chrom\_dirs = glob("$org\_dir\*");

foreach my $chrom (@chrom\_dirs) {

my @fastas = glob("$chrom/\*.FASTA");

foreach my $fasta (@fastas) {

if ( !( -e "$fasta.summary" ) ) {

print "jackhmmer $fasta\n";

system("jackhmmer --domtblout $fasta.summary -o $fasta.hmm.txt $fasta ~/hmmer\_targets/component\_sequences.fa");

}

}

}

### 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;

}

}

}

### do\_all\_hmmer\_stats.sh (deprecated)

Run this script in the **~/genomes** directory to create **hmm\_stats.txt** file which gathers all the generated stats.

#!/bin/bash

if [ -z $1 ]

then

while [ -z $org\_dir ]

do

read -p "Organism directory: " -a org\_dir

done

else

org\_dir=$1

fi

echo $org\_dir

echo "target tlen orf qlen evalue score" > hmm\_stats.txt

for chrom\_dir in $( ls -d $org\_dir\*/ );do

for orf in $( grep -L "\[No hits" $chrom\_dir\*hmm.txt ); do

perl ~/genomes/extract\_hmm\_summary.pl $orf >> hmm\_stats.txt

done

done

### do\_all\_hmmer\_stats.pl

(Replaces do\_all\_hmmer\_stats.sh)

#!/usr/bin/perl

# do\_all\_hmmer\_stats.pl

# Gathers jackhmmer stats in chromosome directories under <organism directory>

#

if (scalar(@ARGV) < 1) { die ("Specify organism directory\n"); }

my $org\_dir = pop(@ARGV);

if ( !( -e $org\_dir and -d $org\_dir ) ) {

die "$org\_dir is not a directory\n";

}

my $hdr = "target\ttlen\torf\tqlen\tevalue\tscore\n";

open(OUT, '>','hmm\_stats.txt');

print OUT $hdr;

close(OUT);

my @chrom\_dirs = glob("$org\_dir\*");

foreach my $chrom (@chrom\_dirs) {

my @fastas = glob("$chrom/\*.FASTA");

foreach my $fasta (@fastas) {

print( "Extract hmm stats for $fasta\n");

system("perl ~/genomes/extract\_hmm\_summary.pl $fasta.summary >> hmm\_stats.txt");

}

}

### create\_hmmer\_stats\_tbls.sql

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

CREATE TABLE hmmer\_statistics

(

hmmer\_statistics\_id SERIAL

, tax\_id numeric

, organism character varying(100)

, chromosome character varying(50)

, target character varying(50)

, tlen int

, orf character varying(50)

, qlen int

, evalue numeric

, score numeric

, import\_date timestamp not null default clock\_timestamp()

);

CREATE TABLE hmmer\_statistics\_import

(

target character varying(30)

, tlen int

, orf character varying(30)

, qlen int

, evalue numeric

, score numeric

);

### 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;

## Consolidated statistics analysis

### consolidated\_orf\_target.sql

\copy ( select h.orf

, b.target,b.score as blast\_score

, h.score as hmmer\_score

, b.expect as blast\_expect, h.evalue

FROM blast\_statistics b

join hmmer\_statistics h

on b.orf\_id = h.orf and b.target = h.target)

to C:\Users\Jeremy-satellite\Documents\RBIF120\consolidated\_stats.txt CSV delimiter ' ‘

### compare\_scores.R

# Scores comparison

consolidated\_stats=read.csv(file = "consolidated\_stats.txt", sep='\t', stringsAsFactors = FALSE)

attach(consolidated\_stats)

plot(blast\_score,hmmer\_score,main='Comparison of BLASTP vs HMM scores for P. falciparum with targets')

abline(a=0,b=median(hmmer\_score/blast\_score)+mad(hmmer\_score/blast\_score),col='red')

detach()

### Score normality and kmeans analysis

# score normality

stats=read.csv(file='../process\_plasmodium/blast\_statistics.txt', sep="\t", stringsAsFactors = FALSE)

organism="P. falciparum 3D7"

attach(stats)

qqnorm(log(score),main='Q-Q Plot for BLASTP log(scores)')

qqline(log(score),col='red',lwd=3)

detach()

norm\_log\_thresh=median(log(stats$score))+2\*mad(log(stats$score))

norm\_thresh=median(stats$score)+2\*mad(stats$score)

print(paste('norm\_log\_thresh:',norm\_log\_thresh))

print(paste('norm\_thresh:', norm\_thresh))

norm=stats[log(stats$score) < norm\_log\_thresh,]

qqnorm(log(norm$score), main='Q-Q Plot for log normal scores')

qqline(log(norm$score), lwd=3, col='red')

highly\_similar=stats[stats$score > norm\_thresh,]

consolidated\_stats=read.csv(file = "consolidated\_stats.txt", sep='\t', stringsAsFactors = FALSE)

attach(consolidated\_stats)

qqnorm(log(hmmer\_score),main='Q-Q Plot for HMMER log(scores)')

qqline(log(hmmer\_score),lwd=3,col='red')

# kmeans analysis

kh=kmeans(hmmer\_score,2,nstart=25)

k\_threshold = min(hmmer\_score[kh$cluster==1])

plot(hmmer\_score,col=kh$cluster,main='2 kmeans clusters for hmmer\_score',

sub=paste('Significance threshold:',k\_threshold))

print(paste('Kmeans threshold for significance:', k\_threshold))

abline(h=k\_threshold,lwd=3, col='purple')

detach()

## Download P. falciparum drugs and targets

/\*\* this is the copy directive used by the **psql** command line to download \*\*/

\copy (select avg(score) as avg\_score, td.tax\_id as original\_tax\_id, td.organism as orig\_organism, md.pref\_name, md.chembl\_id, md.max\_phase, md.first\_approval from hmmer\_statistics h join target\_dictionary td on h.target = td.chembl\_id join drug\_mechanism dm ON dm.tid = td.tid join molecule\_dictionary md ON dm.molregno = md.molregno WHERE h.tax\_id = 36329 and score > 385.3 group by td.tax\_id, td.organism, md.pref\_name, md.chembl\_id, md.max\_phase, md.first\_approval order by pref\_name ) to 'C:/Users/Jeremy-satellite/Documents/RBIF120/paralog\_targets/supplements/targets/P\_falciparum\_hmmer\_drugs.txt' CSV HEADER delimiter ' '

## Download P. falciparum drugs and targets, with annotations

\copy (select avg(score) as avg\_score, td.tax\_id as original\_tax\_id, td.organism as orig\_organism, md.pref\_name, md.chembl\_id,journal, year, volume, issue, first\_page, last\_page,pubmed\_id, doi, title, authors from hmmer\_statistics h join target\_dictionary td on h.target = td.chembl\_id join drug\_mechanism dm ON dm.tid = td.tid join molecule\_dictionary md ON dm.molregno = md.molregno JOIN compound\_records cr ON md.molregno = cr.molregno join docs on cr.doc\_id = docs.doc\_id WHERE h.tax\_id = 36329 and score > 385.3 group by td.tax\_id, td.organism, md.pref\_name, md.chembl\_id ,journal, year, volume, issue, first\_page, last\_page, pubmed\_id, doi, title, authors order by md.pref\_name, year, volume, issue) to 'C:/Users/Jeremy-satellite/Documents/RBIF120/paralog\_targets/supplements/targets/p\_falciparum\_hmmer\_drugs\_annotated.txt' CSV HEADER delimiter ' '

## organism\_hmmer\_threshold.R

# organism\_hmmer\_threshold.R

# computes kmeans based threshold for selecting targets.

library(RPostgres)

db\_name='chembl\_25'

user\_name = 'postgres'

host='192.168.1.180'

port=5432

conn = dbConnect(drv=RPostgres::Postgres(),

dbname=db\_name,

user=user\_name,

host=host,

port=port)

get\_kmeans\_threshold<-function(conn, tax\_id){

q\_tax\_org = paste0('SELECT distinct organism ',

'FROM target\_dictionary ',

'where tax\_id=',

tax\_id)

q\_org\_score = paste0(

'select distinct score, orf, target

from hmmer\_statistics h

join target\_dictionary td

on h.target = td.chembl\_id

join drug\_mechanism dm

ON dm.tid = td.tid

join molecule\_dictionary md

ON dm.molregno = md.molregno

WHERE h.tax\_id ='

, tax\_id

)

org=dbGetQuery(conn,q\_tax\_org)

org\_score=dbGetQuery(conn, q\_org\_score)

organism=org$organism[1]

attach(org\_score)

kmo=kmeans(score,2)

if (kmo$centers[1] < kmo$centers[2]){

thresh=min(score[kmo$cluster==2])

} else {

thresh=min(score[kmo$cluster==1])

}

plot(score,col=kmo$cluster, main=paste('kmeans for ',organism, ', threshold=',thresh))

detach()

return(thresh)

}

#dbDisconnect(conn)

## P\_vivax\_jackhmmer\_drugs.sql

\copy (select avg(score) as avg\_score, td.tax\_id as original\_tax\_id, td.organism as orig\_organism, md.pref\_name, md.chembl\_id, dm.mechanism\_of\_action, md.max\_phase, md.first\_approval from hmmer\_statistics h join target\_dictionary td on h.target = td.chembl\_id join drug\_mechanism dm ON dm.tid = td.tid join molecule\_dictionary md ON dm.molregno = md.molregno WHERE h.tax\_id = 5855 and score > 210.5 group by td.tax\_id, td.organism, md.pref\_name, md.chembl\_id, dm.mechanism\_of\_action, md.max\_phase, md.first\_approval order by pref\_name ) to 'C:/Users/Jeremy-satellite/Documents/RBIF120/paralog\_targets/supplements/targets/P\_vivax\_hmmer\_drugs.txt' CSV HEADER delimiter ' '