

Setting up the program

1) Download and install BLAST from:

<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/LATEST/>

2) Download and install Apache Tomcat from:

<http://tomcat.apache.org/download-70.cgi>

3) Download and install Perl (version perl v5.10.0) from:

<http://www.activestate.com/activeperl/>

4) Download and install Python (version 2.5) from:

<http://www.python.org/download/releases/2.5.4/>

5) Download and install Cygwin from:

<http://www.cygwin.com/>

6) In cygwin, run the installer with “gcc” and “make” checked off (see cygwin instructions for details)

7) Download and install mcl from:

<http://www.micans.org/mcl/src/>

Install as it says in the readme in what you unpack, then, from within cygwin, follow the normal instructions, *except* when you run the configure script. When you run the configure script, you need to run it as

```
./configure --enable-blast
```

The executables the program needs will be in the \cygwin\usr\local\bin directory.

8) Put the files:

```
clusterscript.py  
cluster.config
```

And the FASTA input file at the location "C:/Users/Erasmus".

Specifically, you can download clusterscript.py and cluster.config and place them in a folder in your C:/Users folder named 'Erasmus' along with your input fasta file.

Please do not change the name of the .py or .config file.

9) Place the Erasmus.war file in the webapps folder of Tomcat.

10) Unzip the libraries.zip folder and place the lib files in the lib folder of tomcat.

At this point, you should start the webpage by opening a web browser, and going to:

<http://localhost:8080/Erasmus/index.html>

11) Open the cluster.config file with a text editor such as notepad. Set it up the cluster.config file like, for example:

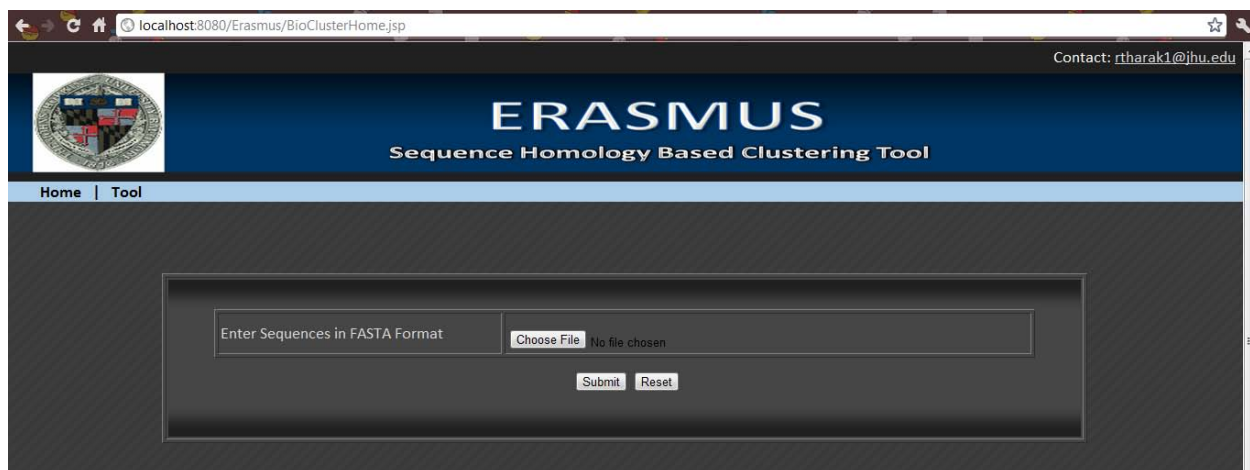
```
FORMATDB_PATH=C:/blast/bin
BLAST_PATH=C:/blast/bin
NUMBER_PROCESSORS=1
MCXDEBLAST_PATH=C:/cygwin/usr/local/bin
MCXASSEMBLE_PATH=C:/cygwin/usr/local/bin
MCL_PATH=C:/cygwin/usr/local/bin
CLMFORMAT_PATH=C:/cygwin/usr/local/bin
```

You might have to change the paths depending on where your executables are. Specifically, find each of the executables mentioned above, formatdb.exe, blastall.exe, mcxdeblast, mcl.exe, and clmformat.exe, and make sure the pathnames to each of them match the pathnames after the equal signs above.

And the program will be ready to run.

Using the program

The program takes your FASTA formatted lists of proteins and runs them through a clustering algorithm. To use it, click the 'tool' link next to the 'home' link. This will take you to a page with a file upload screen, that looks like this:



Here, submit choose your Fasta file and click 'submit.' Once you do this, the program will run and cluster your proteins. Once the proteins are clustered, you will see a screen that looks like this:



ERASMUS

Sequence Homology Based Clustering Tool

Home | Tool

Inflation Factor Level : 150.cl

Total No. of Proteins : 27

[Cluster Further](#) [Download to XLS](#)

NO.	ACCESSION	DESCRIPTION	CLUSTER(S)
1	XP_787833.1	PREDICTED: similar to brain-specific homeobox protein [Strongylocentrotus purpuratus]	1 +
2	EDL87672.1	similar to 25 kDa brain-specific protein (p25-alpha) (predicted), isoform CRA_a [Rattus norvegicus]	7 +
3	XP_001507186.1	PREDICTED: similar to brain-specific transmembrane protein BTCL2 [Ornithorhynchus anatinus]	2 +
4	XP_346073.1	PREDICTED: similar to Homeobox protein GBX-2 (Gastrulation and brain-specific homeobox protein 2) (Homeobox protein STRA7) [Rattus norvegicus]	9 +
5	XP_002128350.1	PREDICTED: similar to Brorin precursor (Brain-specific chordin-like protein) (von Willebrand factor C domain-containing protein 2) [Ciona intestinalis]	1 +
6	EDL87673.1	similar to 25 kDa brain-specific protein (p25-alpha) (predicted), isoform CRA_b [Rattus norvegicus]	0
7	CAB62569.1	brain-specific synapse associated protein, Bassoon [Canis familiaris]	0
8	XP_002124578.1	PREDICTED: similar to brain-specific angiogenesis inhibitor 3, partial [Ciona intestinalis]	86 +

The first column is a numbered list of the total number of protein groups (i.e. “clusters”) you have ended up with. The second column is the accession number of the “center” protein for each group. The algorithm automatically chooses the best candidate for the “representative” protein of the cluster. To see how it does this, see the paper on the TribeMCL algorithm, “An efficient algorithm for large-scale detection of protein families.”

The third column gives you the description of the representative protein. The fourth column gives you how many proteins are in each cluster, and the plus sign next to it will show what is in the cluster.

The algorithm can make the clusters larger or smaller depending on what parameters are used. Because in some biologies, you need to have larger clusters, you can choose to “cluster further” by clicking the button at the top. The algorithm will then start to include proteins in clusters that are less related. You can keep doing this until you are satisfied that the clusters are large enough. Be sure to check what is in the cluster every time, to see what is in them.

Once you have decided that the protein list is sufficiently clustered, you can export to excel by clicking “download to XLS.”