Biol5705
Module: Gene Sequence Analysis
Assignment 3

Run Blast locally

Download Blast software: ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/

Windows users choose version with extension:
- win32.exe
or
- win64.exe

Mac users choose version with extension:
-. dmg

Install the Blast software.

Download the genomes and query sequence we are going to use:
http://morganlangille.com/teaching/biol5705/assignment_3_genomes.zip

The Blast software is run from the command line. This means you will have to open a terminal/shell window. If on Windows this can be done by searching for the "cmd.exe" from the start menu. If on Mac "Utilities/Terminal"

You will need to change into the directory where you stored the sequence files you just downloaded.
On Windows type:
chdir C:\Users\your_username\location_of_directory\

On Mac (or linux):
cd /home/your_username/location_of_directory

Now we are going to make a BLAST database using the proteins from 5 bacterial genomes. (Note these are the same genomes you used in your Genome Alignment from last assignment, but the data is in simpler FASTA file format instead of genbank).

To make the BLAST database type:
```
makeblastdb  -parse_seqids -in 'LESB58.faa PA14.faa PA7.faa PAO1.faa PADK2.faa' -out my_first_db
```

I will explain this command more now.
'makeblastdb' is the program that we are executing. All of the things after it with a '-' are called options:

'-in' is the name of file(s) that contains all of the sequences to be included in the database (Note this must be given)

'-out' is the name that will be given to our new BLAST database

'-parse_seqids' is an option that if given tells the program to parse NCBI sequence ids from the definition lines in the FASTA format. (Note this is NOT mandatory, and if you had sequences without NCBI ids you would not use this)

We will now run our BLAST using 'query.fna' as a single query sequence (Note this file could contain many sequences).

Lets run blastp in the simplest of cases:
```
blastp -db my_first_db -query query.fna -out simple_results.txt
```

This will run the 'blastp' program. You could just as simply use blastn or tblastx if that was appropriate.

Options:
'-db' is the name of the database. Here we specify the database we just made.

'-query' is the name of the file containing our query sequence(s).

'-out' is the name of the file to store our blast results in (this could be anything)

Take a look at the Blast results by opening the 'simple_results.txt' file.

Let's run the Blast again but using a few extra options

```
blastp -db  my_first_db -query query.fna -out pretty_results.html -evalue 0.0001 -html -parse_deflines
```

Extra options:
'-evalue' allows you to set an evalue cutoff instead of using default 10. Here we specify it to be 0.0001.

'-html' specifies that the output should be written in html. Which means that we can use a web browser to look at the BLAST output and that it will create useful links within the blast results.

'-parse_deflines' is similar to the '-parse_seqid' option from when we created the blast database using 'makeblastdb'. It parses the sequence ids knowing that they are from NCBI. This is useful as our output will contain links to the actual sequences now at NCBI.
Note also that we changed the name of the output to 'pretty_output.html'

For future reference you can get a list of all options for every blast command. By typing:

blast_program -h (this will only give a very brief listing of the options)
or
blast_program -help (this gives descriptions for each option)

so to get a full descriptive list of options for blastn:
blastn -help

Now we are done BLASTing and we are going to retrieve some sequences from our blast report and put them into a new file called 'my_homologs.fna'

You should notice that the 3 top blast hits all have 100% identity and are from the same organism.

1) Are these orthologs or paralogs?.

Choose any 1 of these genes that are exactly the same for including in our new file 'my_homologs.fa'. (note you can use a .txt extension instead if this is easier for you).

There are 2 ways to get the gene sequence you want.

Method 1)
Using the sequence id in the Blast report and the organism information, open the corresponding genome file and search for the sequence in the genome file.
Copy and paste the sequence (including the definition line ('>blah')) into your 'my_homologs.fa' file.

OR
Method 2)
Click on the gene link in the Blast report that will take you to the gene record at NCBI.
We want the gene in simple fasta format, so from “Display Settings:” choose 'FASTA (text)'.
Then copy and paste the gene into your 'my_homologs.fa' file

Now repeat this for all the rest of the genes from our BLAST that have greater than 50% identity. (Note this should not be that many sequences)

2) Save your “my_homologs.fa” file and submit it with your assignment answers.

Create a Multiple Sequence Alignment

Download Muscle from http://www.drive5.com/muscle/.

Use the command 'muscle' along with the the options '-in' and '-out' to define the input (my_homologs.txt) and output files (my_msa.fa). All other options can be left as default (e.g. not specified on the command line).

Edit your MSA

Install/Run JalView from: http://www.jalview.org/download.html

JalView annoyingly loads some example data when it runs. Just close the 3 windows within the JalView windows.

Load your MSA: (File-> Input Alignment)

- Colour your alignment using percent identity.
3) Find a region >15 base pairs that is conserved (identical) in all 5 sequences.

- Remove any columns that you think might not be homologous (by selecting columns and Edit->Delete).
(Note: this alignment is not that bad and I realize that this is subjective, but want to see that 1) you understand which might be “bad” columns and 2) you know how to remove columns in an alignment).

4) Tell me roughly what columns you deleted.

Export the alignment as a png image.

4) Send me the image of your edited and coloured alignment.