Assignment 8: Blast, dotplots, and mummer to compare genomes

Your name:
Your email address:

In the first part of today's exercise, we will analyze a couple of phage genomes. These are available in today's zip archive here. The file we will analyze contains 11 genomes of phages that infect *Paenibacillus* sp.. It is a multiple fasta formatted file.

1) **Comparing phage genomes**

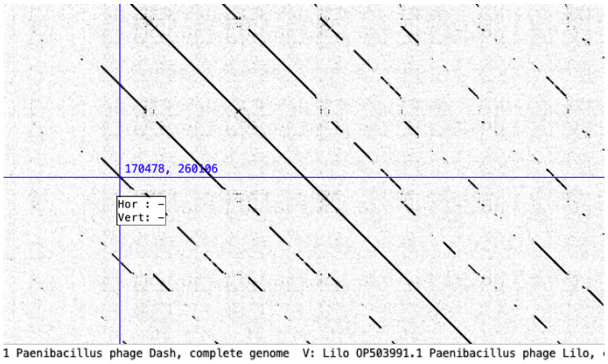
The genomes are in the file **all_ordered_renamed4.fasta**. Check the file format using a text editor.

1A) GENOME PAIR RAPID DOTTER (GEPARD)

If Gepard is installed on your computer, open it and load the file **all_ordered_renamed4.fasta** for both sequence 1 and 2.

If gepard is not installed, follow the instruction in this document (from the archive) to install it *Dotplot_Learning Activity Instructions-7146.docx*.
Briefly, go to https://cube.univie.ac.at/gepard ; scroll to the bottom of the page, download gepard via the gepard1.3 link. Move the folder to the application folder. Open it and click on gepard.jar. You need to go to settings and allow the program to run.

Click in Advanced Mode, in the plot menu, place a Check in "Auto params", then click create dotplot. Check the dotplot. If you hover with the pointer over the plot, it lists at the bottom to which genomes the windows under the pointer belong. If you click, the cross hair will move to that location (and the sequences corresponding to the two windows under the cross should be listed at the bottom, this works in version 1.4, but often fails in version 2.1).
In the example the cross hair is over the comparison between the Dash and Lilo genomes.



```
1 Paenibacillus phage Dash, complete genome  V: Lilo OP503991.1 Paenibacillus phage Lilo, 
```

Which phages have similar genomes?   (You can use the number 1st, 2nd ... to denote the genomes).

In the plot menu, uncheck the auto params and explore the impact of different word lengths -- you need to click on update plot every time. (8 , 10, 15, 30, 80, >100 are good choices. -- Depending on the power of your computer values below 8 might take too long to calculate.) In case the program zooms in onto a small section of the plot, click on the magnifying glass (with a -) repeatedly to zoom out until you see the complete plot.

Which word length choices highlight the similarities between the first three phages?

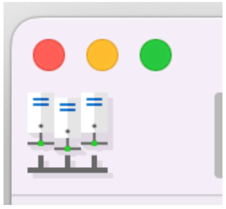
Which work to highlight the similarities between the next three (Lilo, Callan, Dash)

1B) Mummer

Open filezilla, connect to your account on xanadu.

Start Filezilla

Open the site manager (under the file menu, or the left most icon in the bar on top of the application window).



Click on the NewSite button at the bottom left of the site manager window.
On the right
under **protocol** select **SFTP**
under **host** enter **transfer.cam.uchc.edu**
under **LogOn** Type select **normal**

under **User** enter **mcb3421usrXX** (Do not use a leading 0 if your number is below 10). E.g., your username would be mcb3421usr1.)
under **Password** enter *your xanadu password*
Hit connect.

Create a lab8 directory and enter it. (Right or shift click on your home directory in filezilla, select "create directory and enter it", in the window that pops up, replace "New directory" with lab8).

Transfer the file **all_ordered_renamed4.fasta** (from the lab8archive) into the lab8 directory on xanadu.

Establish an ssh connection to your account on Xanadu:

Open an ssh terminal window (either using PuTTY or Terminal), log into Xanadu:

```
ssh mcb3421usrXX@xanadu-submit-ext.cam.uchc.edu
```

Enter your password

To run nucmer, we move to a compute node and load the necessary modules:

```
srun --pty -p mcbstudent --qos=mcbstudent --mem=2G bash
module load MUMmer/4.0.2
module load gnuplot
module load perl
```

The only different to last week's exercise is that we want all matches, not only the best ones. This is accomplished using the `--maxmatch` option in nucmer (for more info try `nucmer -h`)

The following command runs nucmer comparing the multiple fasta file against itself:

```
nucmer --maxmatch all_ordered_renamed4.fasta all_ordered_renamed4.fasta
```

To get the plot we again use mummerplot:

```
mummerplot --postscript --prefix maxmatch_renamed4 out.delta
```

Use filezilla to transfer the postscript file (remember of refresh the directory listing in filezilla) to your laptop and open it in preview or GIMP.

How does the image compare to the gepard analysis?

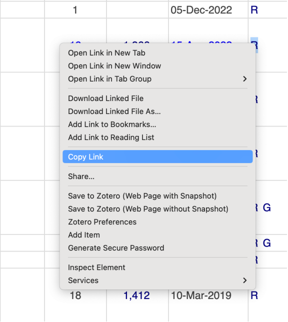
Which do you like better? Are there any instances where the genomes are not colinear? (inversions/recombinations).

If you have time, calculate a dotplot in Gepard for one of the intein-free and intein invaded homologs identified in lab 6.

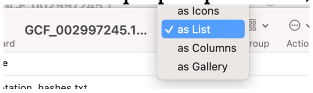
2. Comparing divergent genomes using blastp from the command line

For today's first exercise you will the protein sequences from two completely sequenced genomes as multiple fasta formatted files. One bacterium, one archaeon or one Eukaryote should work nicely. The lab8 zip archive includes a few examples, but please feel free to go after an organism that you are interested in.

Go to the NCBI's genome browser at https://www.ncbi.nlm.nih.gov/genome/browse/ ; select prokaryotes, find the completely sequenced organism (same as last week). Once you have located the organism of interest, move to the last (most right) column of the table. Right (or shift, or command)click on the "R" and copy the link.



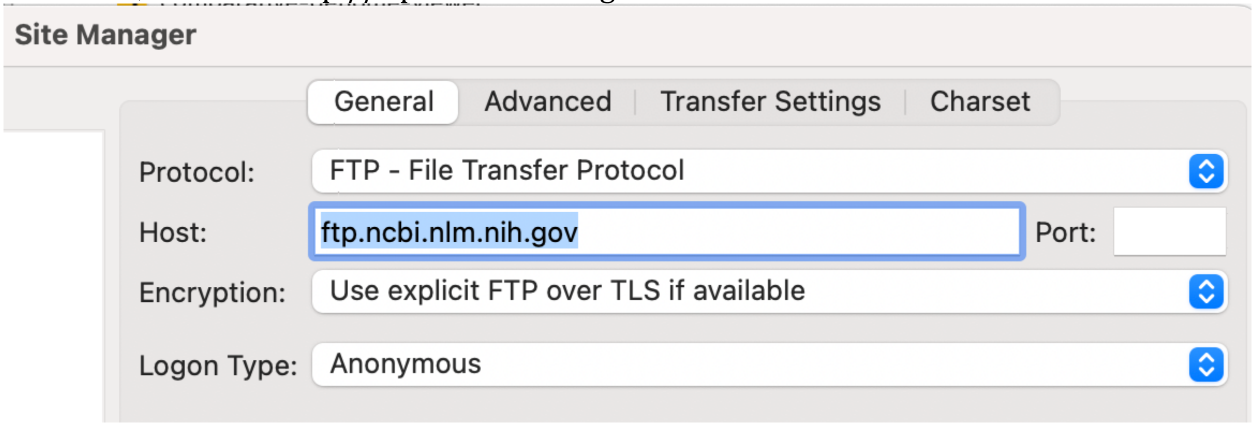
Open finder. In the menu on top select the "Go" pulldown menu, select "Go To Server". In the window that pops up, paste the address you copied from the "R" link. Click connect. In the next pop-up menu, select guest. In the folder that opens, select list view



We want to download the file ending in *_protein.faa.gz .
Double clicking downloads the file and usually unpacks it. (If not we can gunzip it on the cluster). Rename the file into something recognizable and meaningful. Remember no spaces in the name!

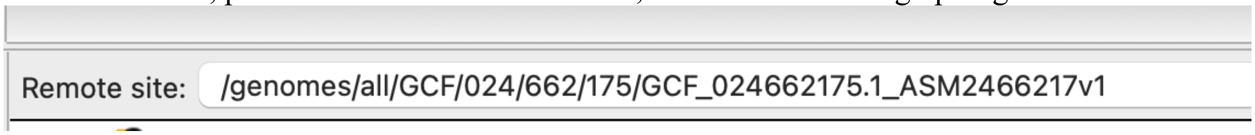
Repeat for another genome.

Alternative:
In filezilla, connect to ftp://ftp.ncbi.nlm.nih.gov



Protocol: FTP - File Transfer Protocol
Host: ftp.ncbi.nlm.nih.gov
Port:
Encryption: Use explicit FTP over TLS if available
Logon Type: Anonymous

Once connected, past the link into the remote site, and delete everthing upto /genomes.



Remote site: /genomes/all/GCF/024/662/175/GCF_024662175.1_ASM2466217v1

This opens a file with all data available at NCBI on the genome.

Use filezilla to transfer the .faa files into your account on Xanadu.

In case your ssh section was terminated, log in again:

Open an ssh terminal window (either using PuTTY or Terminal), log into Xanadu:
ssh mcb3421usrXX@xanadu-submit-ext.cam.uchc.edu
Enter your password

`cd` into the directory into which you had transferred the genomes
move to a compute node:

```
srun --pty -p mcbstudent --qos=mcbstudent --mem=2G bash
```

Check the beginning of your *.faa files:

```
head *.faa
```

Check the available modules – in case of blast you need to use the version number.

```
module avail
```

Load the modules for blast and perl:

```
module load blast/2.13.0
module load perl
```

Pick one of your “genomes” to turn into a searchable databank (in this example Borreliella_burgdorferi_protein.faa):

```
makeblastdb -in Borreliella_burgdorferi_protein.faa -dbtype prot -parse_seqids
```

This should be pretty fast. And report how many sequences were added to the databank.
To search the databank we will use one "genome" as "query", the the databank created above as databank.
In the example the databank is Borreliella_burgdorferi_protein.faa and the genome from yeast is the query:

```
blastp -query Haloferax_volcanii_DS2_protein.faa -db
Borreliella_burgdorferi_protein.faa -out blast_all.txt -outfmt 6
-evalue 10
or
blastp -query Saccharomyces_cerevisiae_S288C_protein.faa -db
Borreliella_burgdorferi_protein.faa -out blast_all.txt -outfmt 6
-evalue 10
or
blastp -query Homo_sapiens_protein.faa -db
Borreliella_burgdorferi_protein.faa -out blast_all.txt -outfmt 6
-evalue 1
```

(the first is fast, the 2nd will take 5 minutes, the 3rd
Depending on the size of the databank and the size of the query this might take a few minutes to finish.
You can check the progress in filezilla watching the size increase in the output file.

-outfmt 6 specifies a tabular output format. (We use an e-value of 10 to also obtain some insignificant hits.)

The blast program will now take every sequence in query file and do a blast search against the database (i.e., this performs a couple of thousand blast searches).
This will take a few minutes. While waiting check in what other students are doing :) Or check this listing of options to use with the blast command.

We are also interested in the best blast hit for each query. The script blastTopHit.pl goes to thought he output table and for every query only prints out the first entry.
(It is possible to restrict the blast search to return only one hit, but it is alleged that this only return the first significant hit encountered, not the best one).
To get the perl script you can directly copy the file from the above link:

```
curl -O  https://j.p.gogarten.uconn.edu/mcb3421_2021/labs/blastTopHit.pl
```

To run the perl script, execute:

```
perl blastTopHit.pl blast_all.txt > blast_top.txt
```

Check the files blast_all.txt and blast_top.txt using more or head:

```
head blast_all.txt blast_top.txt
```

Here is a description of the columns in the outfmt -6 option. A text file containing the header is here.

Use the cat command to add these to the blast output files:

```
cat header.txt blast_top.txt > blast.top_h.txt
```

Check the files with header blast_all_h.txt and blast_top_h.txt using more or head:

```
head blast.top_h.txt
```

Go back to **Filezilla**. Navigate to your lab8 directory, and transfer the blast.top_h.txt file to your Desktop. Load it into Excel. On my computer I can right click on the file and select open with and then Excel. On the Macs in the lab this will be slightly more complicated, as office is not installed, thus you will need to either use office 365 (in a browser) or first install MS office in your account.

Click in the top left field of the table, then convert the spreadsheet into a table with headers.

What is the accession ID pair of the most conserved protein between your two genomes (sort the table first on bitscores (higher to lower) , **then** on e-values (lower to higher)?

To learn what the most similar sequence(s) encode, we can copy the Accession numbers for the most similar matches (use the database IDs in the second column), paste them into a text editor (I use BBedit), replace the new line symbols "\n" with "," and copy the comma separated list into the following command:

```
blastdbcmd -entry IDnumber1,IDnumber2,IDnumber3 -db YourDatabaseName
```

for example (all on one line):

```
blastdbcmd -entry WP_002662068.1,WP_106011477.1,WP_002557027.1,WP_002557108.1,WP_106013134.1,WP_002656192.1,WP_106011525.1,WP_002665762.1,WP_002657046.1,WP_106013121.1,WP_002661984.1,WP_010883927.1,WP_002657221.1,WP_002661428.1,WP_106011512.1,WP_002657221.1,WP_106011512.1,WP_002661428.1,WP_002656799.1,WP_002661757.1,WP_106011512.1,WP_002660520.1,WP_002661852.1,WP_002656738.1,WP_002660520.1,WP_002660666.1,WP_106011471.1 -db Borreliella_burgdorferi_protein.faa
```

This should return the fasta formated sequences of the matches.

What is given as function in the fasta annotation lines?

In Excel, select the third column from the left (this is the percent identity for each BLASTp match), and Insert -- Statistic Chart (all-blue column chart icon in the Charts section) --

Histogram
(Aside: I had problems in editing the histograms this creates on a Mac. It worked much better, when I selected tools, analysis tool pack (you might need to install this first), Histogram. I added a column somewhere that gave the bins from 0 to 100% in 5% increments)

(More about histograms in Excel, including the formula used for default binning.)

You get a histogram of all percent identities. Repeat the histogram for E-values ≤ 10^{-3}, and E-values >.1

Do you observe a smooth distribution of % identity values (a distribution with two peaks would be noteworthy)?

What can explain the difference (or lack thereof) in the histograms for significant and insignificant hits?

Can you explain, why % identity is not a good criterion to distinguish significant from insignificant hits?

For the following questions, the BLAST Command Line Applications User Manual may be helpful → link

How can you get information on the possible parameters in blastp using the command line? (try blastp -h first)

How can you set the wordsize to 2 ?

How can you filter the query sequence for regions with low complexity?

Do not forget to email the completed form!