Assignment 7:
Dotplots and within genome recombination

Your name:
Your email address:

Once you are done with the exercise, email your answers as an attachment to gogarten@uconn.edu and daniel.s.phillips@uconn.edu

Preparation

From lab6 you should have a folder with Aeromonas genomes in the lab6 folder on Xanadu. This folder should contain 3-4 Aeromonas chromosomes.
Another folder should contain 3-4 chromosomes from a genus of your choice, downloaded from the NCBI.
All of these sequence file should have a name with the **.fasta extension**.
Each of these folders should contain the three scripts from the class6 zipped materials file.

You also will need the table with the dnaA locations in the genome – copy it from the class 6 notes.

	Genome Name	Location of dnaA	Strand + or -
1			
2			
3			
4			

For the genomes downloaded from the NCBI, the strand bias plots would help in the discussion of the recombination events.

If you have not yet found an easy way to look at postscript files (they have the extension .ps), **here** is a collection of approaches that work well.

The following assumes that you have established **ssh** (via terminal) and **Filezilla** connections to Xanadu. If you need instruction on how to do so, they are **here**.

Move to the lab6 directory created last week. Check that the Aeromonas, and the genome of your choice sub-directories each contain 3-4 genomes with the .fasta extension.

Creating a mummer/nucmer plot

Mummer as an easy alternative to gene plots and to pairwise blastn searches. The result is something like a dot matrix comparison. nucmer (=NUCleotide mumMER) is part of the mummer package. It also handles input files with multiple contigs rather well. (e.g., if one of your genomes contains several chromosomes or plasmids, or if the genome is not closed.)

The difference to a gene plot is that in this case the search is done on the nucleotide level, and that the program keeps track of the + and the - strand. As genome wide synteny plots works only between closely related organisms, nucmer is sufficient and faster than a gene plot.
Mummer is installed on the cluster. Do not do the commands given in green, they are just the commands that the script is executing.

In case your ssh connection expired, reconnect (instructions are **here**):

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

```
cd lab6,
cd Aeromonas
```

To do everything from the command-line you would do the following:

- **module load MUMmer**/4.0.2
- **module load gnuplot**
- **module load perl**
- **nucmer file1.fna file2.fna**
- For example:
  nucmer A_salmonicida_S68_chr.fna A_veronii_HM21_chr.fna --maxmatch

To plot the matches, we use a script that comes with mummer and that creates a file to be send to gnuplot. At times, a few of the options in mummerplot do not work on the cluster, resulting in an error message. Nevertheless, mummerplot (with the postscript option) generates a file that is being understood by gnuplot as installed on the cluster.

To run mummer plot:
mummerplot --postscript --prefix give_it_a_name out.delta
- out.delta is the default output file from nucmer

You **might get an error message** - ignore it :) If you do, send the gnuplot-file to gnuplot from the commandline:
gnuplot give_it_a_name.gp

If you do not get an error message, the mummer plot will be in give_it_a_name.ps

If you want to compare many genomes, you can use the mummer2.pl script that compares all .fna files in the directory with one another. If you have many sequences, this takes time, you could use it take a break or help your neighbor. The script is in the lab7 archive (see above). Transfer it into the mummer directory on the cluster and run it from the command line by typing
`perl mummer2.pl`

This is rather tedious. Simple scripts can automate this ☺.

Open the script mummer2.pl in a text editor and try to follow what it will be doing.

As we have several genomes to analyze, we will submit the job to the queue.

The shell_for_nucmer.sh script works from your student account.
Open it in a text editor and read through it. Note that it loads the modules.

The mummer2.pl script will calculate all possible comparisons between two genomes, therefore, we will run it twice for the two sets of three to four genomes each.

On the command-line, make sure you are in the Aeromonas directory, into which you had copied the 4 Aeromonas genomes.

To submit the shell script to the queue type

```
sbatch shell_for_nucmer.sh
```

To check if it is running (R) or waiting in the queue (PD) you can check the queue using
`qstat -u YourUserName` or
`squeue -u mcb3421usrxx`

Once done, you should have all possible pairwise plots in the directory. Again, these are in .ps format.

Transfer them to your laptop, open them in a viewer (see **here** for instructions), discuss what this means with respect to the number of recombination events separating the genomes. Note, the first genome listed in the file name is the genome along the X-axis.

Which genomes did you analyze?

Did the genomes use homologous starting points?

If the answer to the above is no, does this make sense in terms of the origin of replication determined in the cumulative strand bias exercise?

How many recombination events did you find (one, two, or many).

REPEAT THIS FOR THE 3 to 4 GENOMES FROM THE SAME GENUS THAT YOU DOWNLOADED FROM THE NCBI.

(Once you downloaded the plots, check on the command line that your jobs are no longer sitting in the queue:

`squeue -u mcb3421usrxx`

If there are jobs running use **`scancel jobID#`** to kill the job.

Which genomes did you analyze?

Did the genomes use homologous starting points?

This will likely be more difficult than for the Aeromonas genomes, most of which were selected to have the ORI close to the beginning of the sequence.

If the answer to the above is no, does the plot make sense in terms of the origin of replication determined in the cumulative strand bias exercise and the dnaA location?

Try to cut and paste the image so that the origins of replication at 0/0 of the plot

How many recombination events did you find (one, two, or many).

Are there parts of the genome where the collinearity of the genomes is worse than in other parts? If yes, where are these located relative to the origin and terminus of replication?

Creating dotplots via gepard

Gepard is similar to mummer, but is allows to change parameters using a GUI. In previous years the older version of Gepard was more compatible with the macs in the computer lab.

If you want to install it on your own computer, instructions are **here**, this also includes a thorough description of the different features that are part of Gepard. If you use MacOSx getting JAVA scripts to work can be complicated. **Here** is a description of what worked for me.

Gepard is particularly useful to compare divergent protein sequences, and to determine the exact location of insertions. (If you do the cure project, do a comparison of the intein containing and intein free homolog, take pictures with the hair-cross located over the beginning and the end of the insertion.)

We will use the following:

Sce_VMA.fa (the vacuolar ATPase catalytic subunit from yeast),
SceHO.fa (the mating type switching HO endonuclease from yeast),
vma1Neurospora.fa (the vacuolar ATPase catalytic subunit from *Neurospora crassa*)

Note: Sometimes GEPARD zooms into a small window. Clicking the zoom-out icon repeatedly helps. You also can update the dotplot, or restart the program.

Save the files to your computer, start GEPARD (if you have, use 1.30) and first compare Sce_VMA.fa against vma1Neurospora.fa (they are both from fungi, but the second one does not contain the intein. In Gepard, under advanced mode, select Auto parameters and Auto matrix. Then update the dotplot. Under display, move the lever for the lower color limit. So that you can clearly see the diagonal. Then click on the diagonal at the end of the extein. Use the arrow keys to move the crosshair so that in the alignment window below, the two exteins are aligned.

With which sequence does the intein begin? (give the first 5 aa)

Where is the last residue of the N-extein located?

Position the crosshair at the end of the intein.
With which sequence does the intein end? (give the last 5 aa)

Where is the first residue of the C-extein located?

Replace the vma1Neurospora.fa sequence with the one from SceHO.fa.

Using the auto parameter settings calculate the dotplot.

At which position in the yeast vma-1 protein does the sequence begin to show similarity to the HO endonuclease?

Which intein domain (selfsplicing or endonuclease) is located here?

Turn off the autoparams (under the plot menu) and recalculate the plot with different word sizes. (under autoparams it is 25), use 10, 50 and 100. Explore the lower color limit.

Which plot looks nicest?

Repetitive proteins in Dotlet

We will use proteolipids as examples (these consist of hydrophobic hairpin loops (alpha helices) embedded in the membrane. They are the subunits of ATPsynthases/ATPases (V-, F- , and A-ATPases) that move the proton (or Na+) accross the membrane.

Compare the *Methanocaldococcus* protein **(WP_010869717.1**) against itself. Do you see any repetitive units? How many?

Does the choice of scoring matrix (in the plot menu) make a difference?

Compare the *Methanocaldococcus jannaschii* proteolipid against the one from *Methanopyrus kandleri* **GI 19887539**

How many repeat units does the *Methanopyrus kandleri* proteolipid have?

Compare the from *Methanopyrus kandleri* proteolipid against itself.

Which settings (matrix and word length) give the nicest plot convincing the viewer that the protein consists of repetitive units?

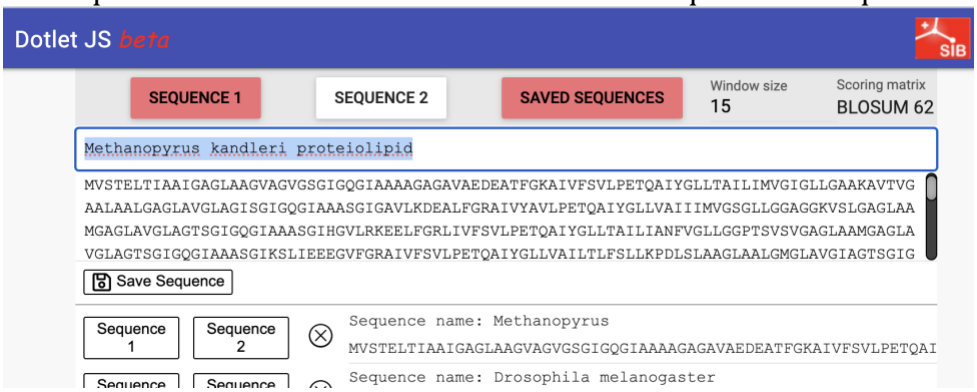
Dotlet:

An alternative to Gepard is dotlet. The original dotlet was superior, but has become incompatible with modern browsers. The JS version at https://dotlet.vital-it.ch is limited (no longer compares DNA and encoded proteins), but has some features which are nice.

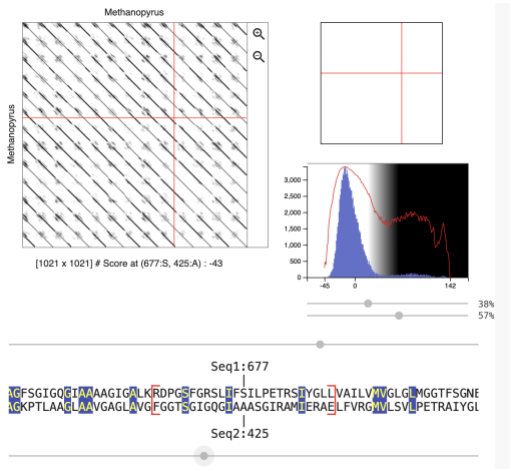
Open the dotlet page. One peculiarity is that dotlet does not recognize fasta formatted sequence files. You need to copy paste **only the sequence** and enter the name.

Use a sequence or sequence pair that only resulted in marginal diagonals in gepard. (Sce_VMA.fa and SceHO.fa; inteins from different hostgenes; the *Methanopyrus* proteolipid **GI 19887539**). After entering the sequence and a name, **save the sequence**.

Then select the sequences you want to compare as sequences 1 and 2 in the window below the sequence window. You can select the same sequence as sequence 1 and sequence 2



As with Gepard, you have a crosshair, and below the dotplot an alignment that is centered around the windows under the crosshair. Again, you can move the crosshair with the arrow-keys, and you also can use the two grey circles above and below the alignment to rapidly move the crosshair. On top you can select the window size (again, a larger window will give nicer diagonals, but also creates more noise in the form of short diagonals) and the scoring matrix.



On the right is a histogram giving the distribution of scores of all the comparisons of windows (in one sequence) to all the windows in the other sequence. The red line is the same data but with a logarithmic axis. The latter allows one to see the windows that have a much higher match than the average windows. The lever below and above the histogram allow you to select the coloring scheme.

Move the crosshair to one of the regions of low complexity (the grayish areas that reflect windows in one sequence matching many windows in the other sequence). These are more visible with a smaller window size.

Which amino acids give rise to the matches in these regions of low complexity? Do these have any particular characteristics?