Assignment 6:
GC Skew and other strand bias plots

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

We will try to analyze a couple of bacterial chromosomes.

Slides with examples are **here**

A) obtaining, and naming genome sequences

There are many different ways to get the genome sequences onto the cluster; however, the easiest will be to first download the files to your computer, unzip the files and then transfer them to a directory on the cluster. A compressed file with several *Aeromonas* genomes and scripts that we will use today is **here**. These are complete genomes, and the chromosome sequences start close to the dnaA gene. Download the file into a class 6 directory on your computer and unzip it.

The genome download page at NCBI currently only works on **Chrome and Firefox**. Go to the NCBI's genome page and search for a bacterial genus/species you are interested to explore (possible examples are Frankia, Nostoc, Paraclostridium, Mycobacterium, ...).

Which genus did you select? Your answer:

How many genomes are listed? Your answer:

Click on "Filters" (at the top left). Move the lever for the assembly level to "Chromosome and Complete".

How many genomes with completely sequenced chromosomes are available for your genus? Your answer:

Click on select columns. Add checkmarks to GC content, Size, Release date, Protein coding and Chromosome count. Click apply. In the column move to the right, and sort on GC-content.

What was the highest and lowest GC content in the group you selected? Give the organisms name (including the strain) and the %GC in your answer:

Then move left to the *size column*, sort.
What was the largest and smallest genomes in the group you selected? In your answer give the organisms name (including the strain) and the size in Mbases in your answer:

move left to the Release Date column.
When were the first and the most recent complete genomes published? Your answer:

Which genome has the most genes and the fewest protein coding genes? (sort on the protein coding column)
Organism with most genes _ number of genes. Your answer:
Organism with fewest genes _ number of genes. Your answer:

For today's exercise you will need a number of completely sequenced bacterial chromosomes in fasta format.

In the Genome Table (remember Safari **does not work** for the download), place a checkmark into four of the rows, then click on download, then select download package.

In the window that opens, place checkmarks into Refseq only, Genome sequence (FASTA), Annotation features (GTF), Annotation Features (GFF), and Protein (FASTA).
If you really want to use a genome for which a Refseq sequence is not available, repeat using "genbank only".
Then click Download. A compressed file should appear in your download folder. Unzip this NCBI_dataset folder.

B) The location of the dnaA gene

Open the folder, open the NCBI dataset folder, and open each of the folders (one for each genome).

In one folder at a time, open the .gtf and the .fna file in a text editor. Search the .gtf file for dnaA. If dnaA is the first gene in the table, you are lucky. Note the location and strand on which dnaA is encoded.

Then open the .fna file. In the annotation line note the species and strain name and copy it into the table below.

	Genome Name	Location of dnaA	Strand + or -
1			
2			

3			
4			

We only want to analyze the chromosome. In some instances, the .fna file includes the chromosome and several additional plasmids or mini chromosomes. In the .fna file search for ">". If there is only one annotation line, this is the chromosome. If there are multiple annotation lines, figure out which is for the chromosome, and delete the annotation lines and subsequent plasmid and mini chromosome sequences.

Save the file under a name composed as follows:
Genus_species_strain_chromosome.fasta .

Recall that unix considers spaces as separators between entries in a list. If the name is Thermotoga maritima, and you use it as an input to a program, the program will look for two files, one called Thermotoga, and the other one maritima. Therefore, **replace all spaces in the names with _** !

The strand bias script expects the chromosome sequence in a single fasta file with the **extension .fasta**.

Which chromosomes did you download, how did you name them?

C) Strand Bias Plots

In filezilla, connect to Xanadu using the connection setting you saved last week.

In case you did not save the settings, here they are again:

protocol select **SFTP**
host enter **transfer.cam.uchc.edu**
LogOn select **normal**
under **User** enter **mcb3421usrXX** (If your number is 1, do not use a leading 0. Your username would be mcb3421usr1.)
under **Password** enter *your xanadu password*

On the Xanadu site of filezilla, right click and create and enter a directory. Call is lab6.

Right click again and create a directory called Aeromonas.

You should still be in the lab6 directory on Xanadu. Right click again and create a directory using the name of the genus for which you downloaded the genomes e.g., Frankia.

Transfer the 4 genomes of your selected genus into the genus named directory. Also transfer the files in the script folder (in the downloaded Genomes_and_Scripts_for_Lab6_2024.zip file) into the same directory.

Transfer 4 genomes from the Aeromonas genome folder (in the downloaded lab6.zip file) into the Aeromonas folder on Xanadu. Also transfer the 3 scripts from the lab6.zip folder into this directory.

The script GCskew_1000_all.pl goes through all the genomes, one at a time, and counts how many more Cs than Gs (or As than Ts, or Gs +Ts more than Cs + As and Gs+As more than Cs and Ts). It counts continuously, if the C is encountered the C versus G counter goes up, and the Gs +Ts versus Cs and As goes down one.

Every 1000 nucleotides the script prints the current counts into a table. If you want more datapoints you can change the modulo command in line 54 , e.g., to
`if ($number%500==0){}` #if you want to print out the counters every 500 nucleotides

The script runs on all file that ends in ".fasta" that are in the same directory as the script. These files each should only contain a single chromosome. A few chromosome sequences from Aeromonads are already in the Genomes_and_Scripts_for_Lab6_2024.zip folder (see above).

Go to terminal and ssh to Xanadu

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

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

cd lab6

cd Aeromonas

module load perl

module load gnuplot

perl GCskew_1000_all.pl
```

This script should create a table for each of the 4 genomes, and then use gnuplot to create a scatterplot from the contents of the table. The plots are saved as postscript files *myplot_your_name_for_the_chromosome.fna.my_table.txt.ps* . Check for errors (often due to wrong file extensions, or spaces in the file name).

Transfer the four plots into the lab6 directory on your laptop, ~~and open the files in GIMP (select a resolution of 300 dpi) or inkscape. Note, if you load GIMP for the first time it takes~~

~~some time. If you do not close the program, the next time it loads the plot much faster.~~
The version of GIMP installed on the note books does not work with the ps files.

Use inkscape instead.

Repeat this for the 4 genomes that you downloaded from the NCBI.

Do the plots have more than one turning point?

Genome species and strain	Number of turning points, approx. location.	Is dnaA encoded next to a turning point?	dnaA at the beginning of the listed genome sequence?	Which strand-bias measure has the largest amplitude?
Aeromonas				
Aeromonas				
Aeromonas				
Aeromonas				

Any other comments you have on the plots:

You also could look at each of the tables "*your_name_for_the_chromosome.fna.my_table.txt*" in MS Excel, and turn the first five columns into a scatter plot. However, as you analyze many sequences, this is rather tedious.

THIS IS HOW FAR MOST STUDENTS GOT IN CLASS.

Email results to gogarten@uconn.edu and daniel.s.phillips@uconn.edu