Assignment7: GC Skew, and mummer plots

Your name:
Your email address:

In the first part of today's exercise, we will analyze a couple of bacterial (and archaeal) chromosomes.

A) obtaining the 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.

Go to the NCBI's current genome list.

Warm up exercise: Click on the "Prokaryotes" tab. Then click on "Filters" (at the top right). Place a check mark behind archaea, and under Assembly level select "Complete". The number in parenthesis following complete, should give the number of completely sequenced archaeal genomes, and these should be listed in the table below. (answer the questions in red)

How many complete archaeal genomes are available at the NCBI:

On the NCBI table you can click in the header to sort the data on this column.

Sort by Release Date.
How many **complete** archaeal genomes have been published so far this year?

Which genome has the most genes and the fewest genes?
Organism with most genes _ number of genes
Organism with fewest genes _ number of genes

Which genomes are the largest and smallest respectively (include the size in Megabases?)

The highest GC content (give name and GC content)?

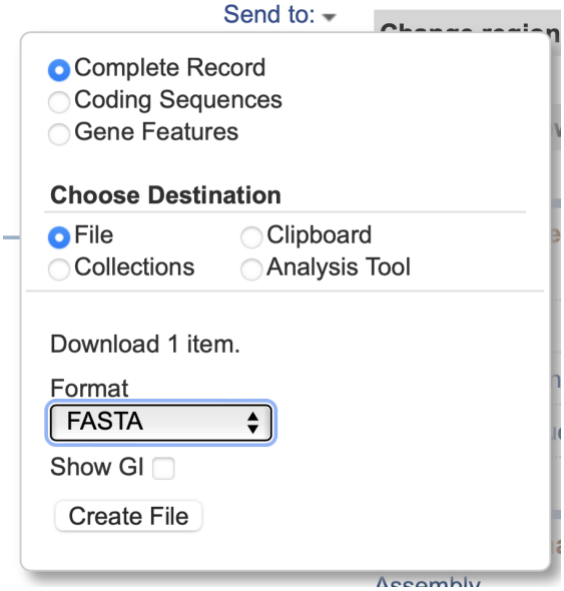
The highest AT content (give name and GC content)?

A1) For today's first exercise you will need a number of completely sequenced chromosomes in fasta format. Bacteria usually work nicer than archaea.

Borrelia is an interesting choice, because it has linear chromosomes. Aeromonas species work great, as well as Termotoga and relatives (Thermotogales). As the GC Bias analysis is pretty much scripted (keep you fingers grossed) do not hesitate to download more than a a handful of genomes (**make sure that you include a few genomes from the same genus, and maybe a few from the same species**).

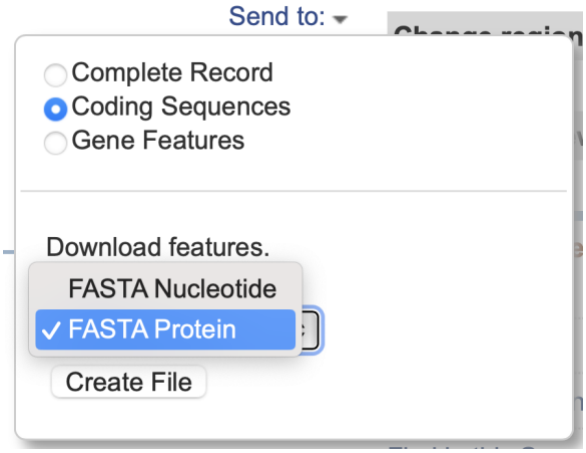
The easiest is to search for the organism of interest, and then, in the replicon column shift or right click on the link for the chromosome that starts with NZ (if there is none, the others should work as well).

In the page that opens up, click on send to (on the upper right) and then select “complete record”, “file” and “fasta”, then “create file”



This should save the nucleotide sequence into your download folder.

Repeat this for the encoded protein sequenceS:



RENAME the downloaded files so that you know which is which. The strand bias script expect the chromosome sequence in a single fasta file with the **extension .fna**. Rename the sequences using the fna extension. Use the extension .faa for the protein sequences.

Put the sequences into a folder on your laptop that is easy to locate.

Check a few of the files in a text editor to make sure they correspond to expectation. (e.g., the chromosomes.fna files should contain a single annotation line at the beginning of the file)

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

A2) For the second exercise we need a couple of genomes from closely related organisms.

Which chromosomes of closely related organisms did you download, how did you name them?

A couple of files that we will use today are in a zip archive **here.** Download them and put them into the same directory as the sequence files.

The remainder of today's work we will take place on the cluster. Some of the exercises you also could do under MacOSX (CG Skew analyses), but the cluster works better ☺

You will need to establish a Filezilla connection to the cluster:

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.

1. Cumulative Strand Bias

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 ".fna" that are in the same directory as the script. These files each should only contain a single chromosome. A few chromosome sequences are already in the Lab7_FileArchive (see above).

Using filezilla, create a directory called lab7. Inside this directory create a directory called Skew.
Copy the **...chromosome.fna** files and the **GCskew_1000_all.pl** into this directory.

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

```
ssh mcb3421usrXX@xanadu-submit-ext.cam.uchc.edu
srun --pty -p mcbstudent --qos=mcbstudent --mem=2G bash

module load perl
module load gnuplot

cd lab7
cd Skew
perl GCskew_1000_all.pl
```

The script works on all *.fna sequences in the directory and creates one table for each ending in ...mytable.txt .

You can look at each of these tables in MS Excel, and turn the first five columns into a scatter plot.

However, if you have many sequences to analyze, and if you do not have an excel program on your computer, this is rather tedious. Therefore, the revised **`GCskew_1000_all.pl`** script takes the tables and uses gnuplot to create scatterplots. Currently, the script produces postscript files. On older MacOSX versions, you can open the ps files in PREVIEW. Under newer version of MacOSX this no longer works. GIMP is a free software, that imports *.ps files and turns them into .png files. The scatter plots are in the files that have the following names:
myplot_your_name_for_the_chromosome.fna.my_table.txt.ps

Transfer the scatter plots to your computer using filezilla and open them.

Note: in filezilla, you need to click on the refresh button to see the files that the script has generated.

Put your answers into the table below,

The turning points usually are the origin and the terminus of replication.
Do you see two turning points for the cumulative strand biases? (Remember that in circular genomes the beginning of the plot is next to the end and might represent a turning point.)

Open the corresponding .faa file for the genome you are analyzing in a texteditor and search for "dnaA" in the annotation lines. In bacteria *dnaA* is often/usually the first gene next to the origin of replication.
Does the location of dnaA correspond to one of the turning points

Which of the bias measures has the largest amplitude?

Name of genome	Turning point at beginning	Two turning points	Largest Amplitude	dnaA at the beginning of the listed genome sequence?	dnaA encoded next to turning point

Other comments:

2. 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 plasmids or additional chromosomes, 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. |Mummer is installed on the cluster.

The following assumes that you established a filezilla and an ssh connection to xanadu (see above). [note: establish the ssh connection first, to be sure your password has not expired!]

Using filezilla or ssh create a new subdirectory in lab7. Call it mummer.
(e.g., **ssh** `mcb3421usrXX@xanadu-submit-ext.cam.uchc.edu,`
`cd lab7,mkdir mummer, cd mummer`
Or right click on the lab7 folder in mummer and select create a new directory and enter it.)

Use filezilla to move the genomes of the closely related genomes you want to analyze into this directory.

To do everything from the command-line:

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

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
```

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
```

If you have more than 4 sequences, you should use a batch file and submit it to the queue. The shell_for_nucmer.sh script works from your student account.

```
sbatch shell_for_nucmer.sh
```

To check if it is running (R) or waiting in the queue (PD) you can check the queue using

```
squeue
```

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).

Do not forget to email the completed form!