Assignment 5:

Intro to Xanadu, PRSS, the transitive property of homology, seaview

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

Your name:
Your email address:

Some cluster computer basics

Spend 5 minutes reading this introduction to the command-line interface. If you think something like this might be in your future, keep reading after class.

Find the terminal application (Terminal.app). By default this is located in the Applications folder, in the subfolder "utilities" (`/System/Applications/Utilities/Terminal.app`). The easiest might be to search for Terminal.app in the finder or in spotlight.

Double click the Terminal application icon. In the window that opens, type
`ssh mcb3421usrXX@xanadu-submit-ext.cam.uchc.edu`
where XX is a number assigned to you. If your number is below 10, do not use a leading 0. Your username would be mcb3421usrX. Write down your username into your notebook!)
It will ask you to accept a new host key, type **`yes`** <return> [<return> means hit the return key.]
Now enter the password _____. Hit return.
<**Note: the terminal will not echo the keystrokes you make to enter the password.**> (you will need to type without visual feedback).

It is important to know a little bit about the structure of the cluster. When you log into the cluster, you are at the head node. Everything you execute will run on the head node. You never-ever should run anything that needs computational resources on the head node. To run programs or execute a command such as blast, you should have moved to a compute node, or submit the command to the queue (which runs it on a compute node)!
You can read more about the xanadu cluster here.

To move from the login node to run an interactive session to a compute node, type
`srun --pty -p mcbstudent --qos=mcbstudent --mem=2G bash`
and then hit the return or enter key. This takes you to a "compute" node.
If the cluster is busy, and the queue you select is busy, the move to a compute node can take several minutes.
You can check the node you are on by typing **`hostname`** at the command-line. The head-node is hpc-ext-2.cam.uchc.edu, the compute node something like xanadu-68.
`whoami` returns your username.

Now type **`qstat -u YourUserName`**
This command shows the job-ID of your session.

Type **ls**
This gives you a directory (or folder) listing. There may be nothing to see (yet). Type **mkdir lab5**
This command stands for "make directory". Now type **ls**
again. You should see your newly-created directory. Type **pwd**
This stands for "print working directory". This forward-slash (/) separated "path" is a shorthand for the directory you are currently in.
Type **cd lab5** **(cd ..** would move you back up to your home directory)
Now type **pwd** again. You have moved into the lab5 directory.

If you need more information on a command (e.g., pwd), you can use the man command. **man pwd** displays the manual pages for the pwd command. (You can move forward using the space-bar, or the arrow keys. to leave the manual page type q.)

There are many ways to get files into your account, e.g., the file transfer application that is part of the SSH Client or Filezilla. Another possibility is the wget commant. To move this file into the directory on your cluster, you could type

```
wget
https://j.p.gogarten.uconn.edu/mcb3421_2020/labs/HaloATPases+Intein.tx
t --no-check-certificate
```

You could copy paste the above line to the terminal window, or right click "this file" and copy the link address.

Do the same for this file:

```
wget https://j.p.gogarten.uconn.edu/mcb3421_2020/labs/HaloATPases-
Intein.txt --no-check-certificate
```

Once you downloaded the files into the lab5 directory, you can display their content by typing

```
cat HaloATPases+Intein.txt
```
(the file also has several not haloarchaeal subunits)

```
cat HaloATPases-Intein.txt
```

Alternatively you could use the **more** or **less** commands that allows you to go through a file page by page (Type <space> to get to the next page, type "q" to quit).
These are typical multiple sequence fasta files from the NCBI. Take a look at the annotation lines.

Where is the species and strain given in the comment-line/annotation-line? What pattern could you use to automate retrieval of the species and strain designation?
Your answer --->

To display both files you could use the wildcard "*", which stands for any characters.
cat H* displays both files to the screen.

By default the output of cat goes to the standard output, in our case the screen. But you can redirect the output to a file using the > symbol.

```
cat H* > HaloATPases_and_homologs.txt
```

copies both files into a new file.

Use more to scroll through the file
more HaloATPases_and_homologs.txt

There is a similar command called less (strange humor, more or less) that allows you to move forward and backwards in a file:
less HaloATPases_and_homologs.txt

The unix command line comes with helpful features.

A) One is **automatic line completion**.
If you type a command and hit the <tab> key, the interface will complete the line as long as there is only one possible completion.
In our case,
less H<tab> will complete the line to **less HaloATPases**. If you enter the <tab> key again, the different possible completions will be listed, enter the next character of the file you want to see (_ or - or +) and hit the <tab> key again.

B) Another is the history. The upward arrow in the cursor field will recall the last command (and hitting it repeatedly will allow you to go back in the command history.) You can use the forward and backward arrow to move the cursor to within the command and then modify the text (this is great, if you mistyped, or were in the wrong directory). The only complication is that the mouse does not work to move the cursor location in the command line. Try it out!
You can output the complete history of the commands you submitted by typing **history** .

type **more HaloATPases and_homologs.txt**
then use the recall command and the arrow keys to correct the command.

To align the sequences in mafft, we first need to load the mafft module:
type: **module load mafft**
To execute an alignment type (or better, use copy and paste):
mafft --auto --reorder --maxiterate 1000 HaloATPases_and_homologs.txt > HaloATPases_and_homologs_MAFFT_aligned.fasta

This writes the output into a file **HaloATPases_and_homologs_MAFFT_aligned.fasta**
The **reorder** option puts out the sequences in order of similarity.

Type **exit**
You will return to the master node. Now type **qstat -u YourUserName**
You should have no jobs running. If there is a job listed, then type **qdel**
followed by a space, and then the job-ID number, followed by return or enter key.
Then type **qstat -u YourUserName** again, to confirm that there are no running jobs.

Then type **exit** to exit from the master node.

Start Filezilla

In the application window, click on the site manager icon (the button with the three

computers 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** (If your number is 1, do not use a leading 0. Your username would be mcb3421usr1.)
under **Password** enter *your xanadu password*
Hit connect.

You should see 4 windows. On top the paths to the directories on your computer (left) and on the cluster (right). On the bottom the contents of the two active directories (i.e. the directories between which the files will be transferred). You can move to a directory/ subdirectory by double clicking. The folder with two dots behind it is the parent directory. If you double click on a file in the bottom windows it is transferred (copied) between the two directories on the two computers. You also can drop and drag files in the window representing the directory on xanadu.
Right clicking offers a couple of options, the most convenient is to change permissions. You can easily give or revoke permissions to others.

Which groups and which type of permissions can you select using FileZilla?
Your answer --->

On your laptop, create a directory called lab5 (you can use the finder or filezilla.
Use filezilla to download the file **HaloATPases_and_homologs_MAFFT_aligned.fasta** into the local lab5 directory.

Open the file in seaview on your laptop (in finder right click and open with; or open seaview and then in the file menu open fasta file). Scroll to the right. Can you identify the intein?
What are the last 2 aa of the N-extein (immediately upstream of the intein?
Your answer --->

What is the first aa of the C-extein? Does the intein insertion site ring any bells?
Your answer --->

What is the first aa of the intein?
Your answer --->

What are the last four aa of the intein?
Your answer --->

How do the beginning and the end of the alignment look?
Your answer --->

Realign the sequences in SEAVIEW using muscle (a popular and efficient alignment program):

- Under the Align-menu select Alignment options, place a check into muscle.
- Then under the Align-menu select Align-all.

Do you observe notable differences compared to the MAFFT alignment? If you unsure, open the downloaded MAFFT alignment in new SEAVIEW window.
Your answer --->

1. PRSS (20 minutes)

PRSS is a program written by Bill Pearson that uses the shuffling approach to determine if two sequences are significantly similar.
Using PRSS*, determine if there is significant similarity between the proteins with gi numbers: 2506213, 2493127, 4323566, 2983405, 1303679. **To start the program click on "Shuffle Sequence".** Sample values are in the table below. Note that 2e-3 means 2 times 10^{-3}. Note, you can do many comparisons at once, using multiple gi numbers for the second sequence (here).
Each of you will pick one row and enter the **pseudo E-value for 10000 comparisons** into **the table (here).**
Note: If someone already entered a value, add your value in addition, separated by a ";"
Note: if you enter multiple gi numbers as the second sequences, the output lists them in order of significance level. Not in input order.

Group discussion of results.

OPTIONAL

For a comparison that resulted in a significant E value (strong - smaller than 1e-5), but not overwhelming (i.e. > 1e-90) repeat the analysis selecting a different substitution matrix (Blosum80, PAM250).
How does the E-value change?
Your answer --->

For a comparison that resulted in a significant E value (strong - smaller than 1e-5), but not overwhelming (i.e. > 1e-90) repeat the analysis increasing and decreasing the gap opening penalty (e.g., -1, -10, -13, -20, -100).
Which gap opening penalty gives the smallest E-value? What might be the reason for your finding?
Your answer --->

END optional

2. BLAST (10 minutes)

Repeat a few of the pairwise comparisons using Pairwise BLAST
(go here, then
select the protein BLAST.
Click the box to select align two or more sequences, which is at the bottom of the "Enter Query Sequence" box. Make sure the blastp tab in the header is selected).
You can paste the GI numbers directly into the box labeled "Enter accession number(s), gi(s), or FASTA sequence(s)" You can force the program to report insignificant alignments by increasing the expect value. To do this, click on the plus sign labeled "Algorithm parameters". A number of additional options will drop down, including the "Expect Threshold," which is set to 10 by default, but you can enter a larger number to obtain less significant matches (do this only if the default parameter does not return any result). When all of your parameters are set, click the BLAST button.

How do these E-values compare to the ones obtained using PRSS?
Your answer --->

Note: The NCBI BLAST interface adjusts the gap penalties to a new default value for each substitution matrix.

3a. Transitive homology? Part one (5 minutes)

You find that sequence A (gi|1303679) has a significant similarity to sequence B (gi|2493127) over the entire length and sequence B has significant similarity to C (gi|2983405) over the entire length, but C and A are not significantly similar (assuming an E-value <e-8).
Can you nevertheless conclude that A is homologous to C? (Two characters -- sequences, or morphological characters -- are homologous if they are derived from the same character existing in some ancient organism.)
Your answer --->

3b. Transitive homology? Part two (10 minutes)

Does the same reasoning hold for gi 6320016, gi 1303679, gi 2507047?

Why might this case be different from the previous one?
How does the output from the pairwise blast comparison help you to draw a conclusion (compare 6320016 with 1303679, and 6320016 with 2507047)? (Check both the Graphic Summary and the DotPlot.) (go to blastp at the NCBI, select align two sequences. Paste the GI numbers into the two boxes).

Your answer --->