Assignment 9:

Sequence alignment, dissecting protein sequences into domains, Neighbor Joining trees, preparing sequences to predict folding in alphafold

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:

A) Getting the sequences to calculate MAFFT and muscle alignments of intein free and intein containing phage proteins

We will work on the intein containing and intein free homologs from lab4. Open the two phams (or the one pham, if it contains both the intein containing and intein free homologs) in a text editor and copy them into a single multiple fasta file (call it **`my_phams.fasta`**).

If you did not succeed with the exercise in lab 4 use ONE of the files linked below:
The Phams homologous to Violet_10 – the intein free homolog had 1300+ sequences, I retained only those most similar to the intein containing ones:
https://j.p.gogarten.uconn.edu/mcb3421_2024/labs/lab9/Violet10_Topgun_11_combined_small_unaligned.fasta : the original Pham contains about 200 very similar terminase sequences

The Phams homologous to Racecar-193 and C3PO_75:
https://j.p.gogarten.uconn.edu/mcb3421_2024/labs/lab9/Racecar_193__C3PO_75_combinedPhams.fasta : only a few but rather divergent sequences, one with a rather short intein:

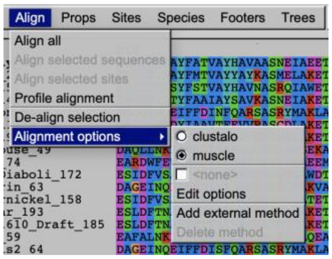
Selected homologs to the helicase in phage Alice (with intein). Only unique sequences were retained for the intein free homologs:
https://j.p.gogarten.uconn.edu/mcb3421_2024/labs/lab9/Alice_helicase_unique_unaligned.fasta

Take a note on the phage name and gene number for one of the phages with and one without the intein:
Your answer:

(Save the file under the name my_phams.fasta)

B) Aligning the sequences using muscle:

Open SEAVIEW and load the file **`my_phams.fasta`**.
Under *Align* > *Alignment options* select muscle. Then select *Align* > *Align all*



How does your alignment look?

Your answer --->

Save the file (from the file menu) as a mase formatted file (call it **`my_phams_muscle.mase`**)

C) Aligning the sequences using mafft (You only need to do this if your muscle alignment is unsatisfactory).

You have two options for this (option 2 is below):

Option1) use the xanadu cluster:

Use filezilla and transfer the **unaligned** file (my_phams.fasta) into a lab9 directory on Xanadu.

Open terminal and ssh to Xanadu:

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

Log into a compute node:

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

Load the mafft module:

```
module load mafft
```

Then execute mafft to align your sequences:

```
mafft --reorder --leavegappyregion --maxiterate 100 --localpair
my_phams.fasta > my_phams_mafft.fasta
```

Use filezilla to move the alignment to your laptop, open it in SEAVIEW and compare it to the muscle alignment.

Option 2) use the MAFFT webserver

Connect to the webserver is at https://mafft.cbrc.jp/alignment/server/index.html

Copy/Paste your unaligned sequences in fasta format into the text box, or select the file to upload.

Under advanced settings select “L-INS-I ..”

Under advanced settings select “Leave gappy regions”

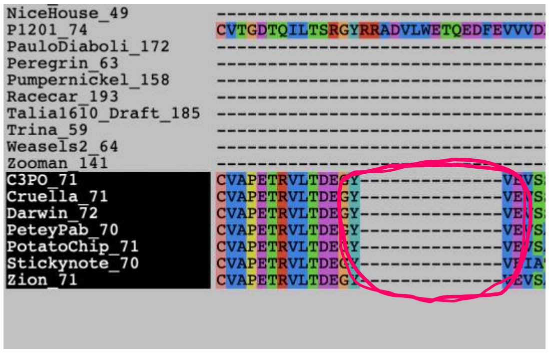
Leave everything else in the default settings and click on “Submit”

Save the file in fasta format and rename it into `my_phams_mafft_mafft.fasta`

This seems to work just as well as using the Xanadu server, which allows to run more iterations.

D) Editing the alignment by hand (You only need to do this if your best alignment is unsatisfactory).

If the intein is not cleanly aligned, you can edit the alignment. Editing means inserting or deleting gaps. As the sequences are aligned already, you usually want to insert/delete gaps in multiple sequences at once. For example, to delete the circled gap in the alignment below,



```
NiceHouse_49              --------------------------------
P1201_74                  CVTGDTQILTSRGYRRADVLWETQEDFEVVVD
PauloDiaboli_172          --------------------------------
Peregrin_63               --------------------------------
Pumpernickel_158          --------------------------------
Racecar_193               --------------------------------
Talia1610_Draft_185       --------------------------------
Trina_59                  --------------------------------
Weasels2_64               --------------------------------
Zooman_141                --------------------------------
C3PO_71                   CVAPETRVLTDEGY--------------VEVS
Cruella_71                CVAPETRVLTDEGY--------------VEVS
Darwin_72                 CVAPETRVLTDEGY--------------VEVS
PeteyPab_70               CVAPETRVLTDEGY--------------VEVS
PotatoChip_71             CVAPETRVLTDEGY--------------VEVS
Stickynote_70             CVAPETRVLTDEGY--------------VEIA
Zion_71                   CVAPETRVLTDEGY--------------VEVS
```

you would highlight the sequences on which to act on the left, in one of the sequences at the end of the gap click on the first amino acid, and then use the delete (backwards) key to remove the gaps. To insert gaps press the space bar.

Note, the C-extein now will be out of alignment, you will need to insert gaps in another part (using the spacebar), or delete gaps in all the other sequences.

Do you see clear boundaries surrounding the intein sequence? Which program/approach provided a better alignment of the inteins, which had nicer intein extein boundaries?

Your answer --->

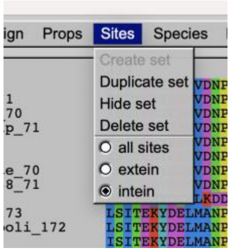
Sometimes it is impossible to get satisfying alignments. This may be due to multiple divergent inteins sitting in neighboring regions of the extein. Or this may be due to the intein found in a phage that has no related phages (i.e., phages in the same cluster). The alignment still might be good enough to determine the beginning and end of at least one intein sequence.

E) Defining sites in SEAVIEW and saving intein and extein sequences separately.

As the inteins did not necessarily evolve together with the intein, we want to analyze the intein sequences separately from the extein sequences.

Create sets of sites that correspond to the extein and the intein:
First go to Sites Pull down menu and create a set called "all sites", then duplicate this set, call it intein. Duplicate it again and call it extein.

Under sites, you now have three set:



Place the check in intein, scroll to the right, and in the row of xxx below the alignment, click on the x below the **last aa of the N-extein** (the x disappears, and the column is grayed out). Then **right click** (shift click if you use a trackpad, or double click if that is how you set up the trackpad) on any of the xxx below the N-extein, -> all the x below the N-extein should disappear.   Do the same at the end of the intein: remove the x under the first aa of the C-Extein, then right click / shift click on any of the Xs to the right. Only the X under the intein should remain.

Do the same for the extein only sites. Select "Sites >extein", remove the x under the first aa of the intein, remove the x under the last aa of the intein, right click on any of the x under the intein.

To save the intein sequences only, select Sites "intein". Highlight the sequences that have the intein, then under File > Save selection > enter a name e.g. my_phams_intein.fst > select fasta > save .

To save the extein sequences only, select Sites "extein". Highlight **all** sequences, then under File > save selection enter a name e.g. my_phams_extein.fst.

To save all the data go to Sites, select hide set, remove the highlights from the sequence names, the under File > save as > in the pull-down menu select mase > save (the file should be called my_phams_Extein_and_Intein.mase

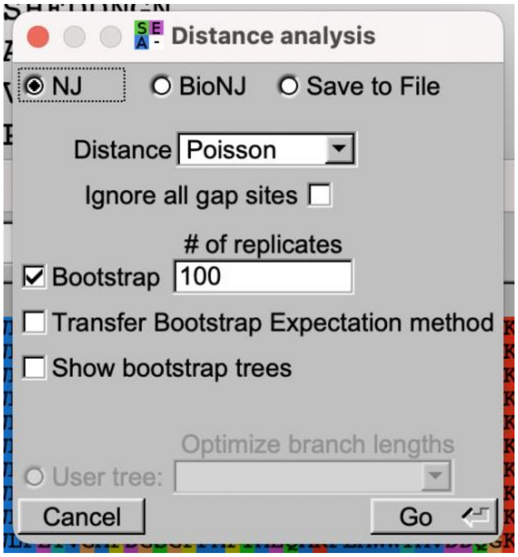
Check in a text editor that the intein and extein fasta files are complete and in fasta format (a frequent error is that the inteins file contains empty sequences for the homologs not invaded by the intein.

Save the following four sequences into separate text files (as single sequence fasta documents):

- ExteinAndIntein.fa
  From my_phams_Extein_and_Intein.fst save the aa encoded by the gene you started out with as single sequence fasta file. In this file delete all the gap characters (-)
- Extein_spliced.fa
  From the file my_phams_extein.fst save the aa of the extein of the gene you started out with. In this file delete all the gap characters (-)
- Extein_not_invaded.fa
  From the file my_phams_extein.fst save the aa of the gene you identifies as an intein free homolog. In this file delete all the gap characters (-)
- Intein.fa
  From the file my_phams_intein.fst save the aa of the intein in the gene you started out with. In this file delete all the gap characters (-)

F) Reconstructing the evolutionary history for the intein and extein sequences

Seaview provides different approaches to reconstruct the evolutionary history from sequences. (More details on Monday). The fastest is an algorithmic approach called neighbor joining (NJ).
Load my_phams_Extein_and_Intein.mase into seaview. Select all sequences and the extein sites. Then select Trees > Distance methods > place check marks as below and click on Go



● NJ ○ BioNJ ○ Save to File

Distance Poisson

Ignore all gap sites □

# of replicates

☑ Bootstrap 100

□ Transfer Bootstrap Expectation method

□ Show bootstrap trees

Optimize branch lengths

○ User tree:

Seaview comes with its own tree viewer. Explore the Br support, Re-root and Swap options. Note these features do not change the tree, only the way the tree is depicted.

The subtree option is particularly useful, if one wants to focus on part of the tree and does not want to include long branches leading to divergent homologs (e.g., if all intein are in a group of more closely related extein, saving this part of the tree as a subtree is a good idea. IMPORTANT: Take notes on what you were doing (how the tree was calculated, and how a subtree was selected).

Save the tree(s) as unrooted trees (in the File menu)

If your my_phams_Extein_and_Intein.mase file contains more than one intein sequence, repeat the tree building for the intein sequences (select Sites > intein, and select all intein containing sequences.

Which dataset did you work on (give the phage name and gene number of one intein and one intein free sequence):
Your answer:

Do the intein containing sequences form a clan in the extein tree?
Your answer:

G) Predicting structures of intein and extein in alphafold

Open intein.fa in a text editor. Delete all the gap symbols "-".

On your laptop open chimeraX1.8 (you can download the program from https://www.cgl.ucsf.edu/chimerax/)
Somewhere along the line you will need to login with your google account!

Under tools > Structure Prediction select alphafold.

In the windows that opens, paste the amino acid sequence for which you want to predict the structure into the sequence window (select paste, if this is not the default).
DO NOT PASTE THE ANNOTATION LINE, only the aa sequence, no gaps, no * at the end.
Click on "predict". Careful with too much clicking. There will be a window popping up that says .... Run Anyway. Now you wait .....

In the meantime, you could reacquaint yourself with the structure of intein – Go over your results from lab3.

Once alphfold is done, the predicted structure is displayed in chimeraX. The coloring scheme reflects the reliability of the reconstruction. Dark blue is reliable, yellow and red are uncertain.

How certain is alphafold of the quality of the reconstructed structure?
Your answer --->

Save the structure from chimeraX as a chimeraX session (rotate it so that the putative splicing domain is visible), as a png image, and as a pdb file. Give it the name of the phage_gene_number_intein. (e.g., C3PO_75_intein.pdb or C3PO_75_intein.png).

Open the pdb file color by secondary structure.
You can compare your structure to a mini intein, i.e., an intein without a Homing ENdonuclease (HEN) (6RIY or 1AM2), the 17-kDA fragment of hedgehog C-terminal auto processing domain (1AT0 or 6TYY), an intein with homing endonuclease, bound to its target cleavage site (1LWS), or the HEN from an intron bound to its target site (6X1J).
Note that the latter two HENs resulted from an internal duplication, whereas the HENs from the phages likely have only one LAGLIDADG domain and likely form a homodimer to cleave the DNA -- see 1BP7 for an example).

Is the predicted structure similar to that of an intein structure determined via x-ray crystallography?
Your answer --->

Do you recognize the self-splicing domains in your intein?
Your answer --->

Do you recognize a homing endonuclease domain in your intein?
Your answer --->

Does your HEN domain contain one (as in 1BP7) or two LADLIDADG domains (as in 6X1J) ?
Your answer --->

If your intein alignment includes a shorter sequence (a candidate for a mini intein), repeat the alphafold prediction for this sequence.

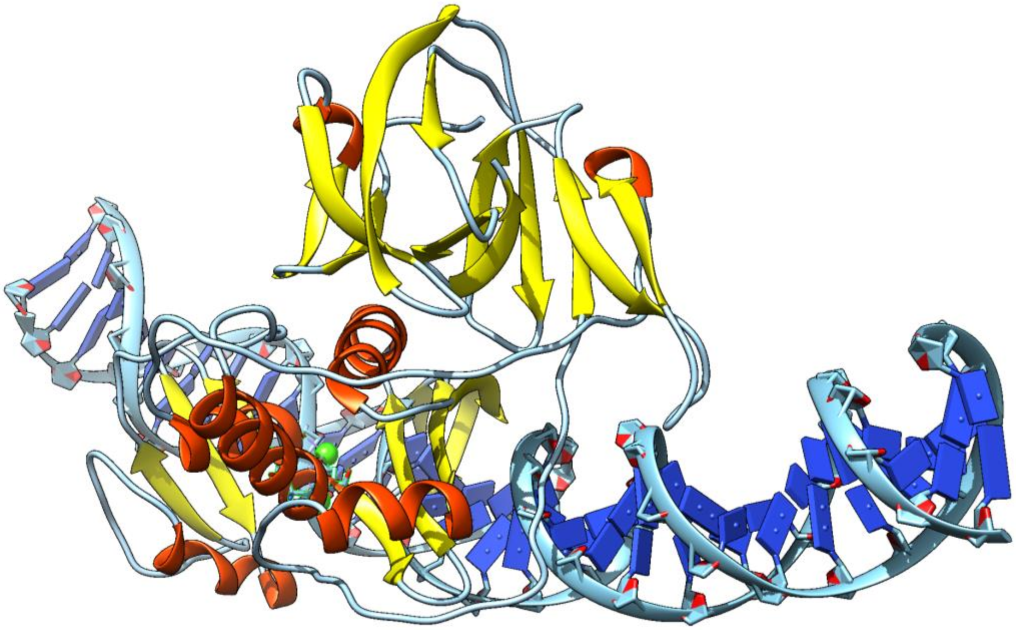
Does predicted structure from the shorter sequence suggest that this is a mini intein (i.e. only containing the splicing domain)?
Your answer --->

If you have time and interest, repeat the structure prediction for the extein + intein sequence, the uninvaded extein, and the uninvaded homolog.

Discuss your findings.
Your answer --->

How far did you get in today’s exercise?
Your answer --->

Safe the files in a safe place, we will need them again.



The C3PO_75 intein aligned to 1LWS, with the 1lws protein hidden