Assignment 6:

Sequence alignment, dissecting protein sequences into domains 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:

MAFFT and muscle alignment of intein free and intein containing phage proteins

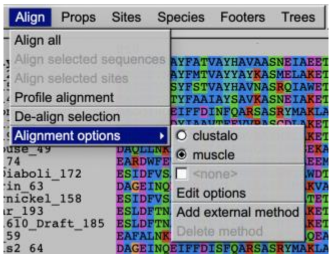
We will work on the intein containing and intein free homologs from lab4. Open the two phams (or the one pham, if it contains both the intein containing and intein free homologs) in a text editor and copy them into a single multiple fasta file (call it **`my_phams.fasta`**).

If you did not succeed with the exercise in lab 4 use ONE of the two files linked here:
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_2023/lab_5/Violet10_Topgun_11_combined_small_unaligned.fasta : about 200 very similar Terminase sequences

The Phams homologous to Racecar-193 and C3PO_75:
https://j.p.gogarten.uconn.edu/mcb3421_2023/lab_5/Racecar_193__C3PO_75_combinedPhams.fasta : only a few but rather divergent sequences, one with a rather short intein. .

A) 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 the alignment look?
Your answer --->

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

B) Aligning the sequences using mafft:

Use filezilla and transfer the **unaligned** file (my_phams.fasta) into the lab5 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:

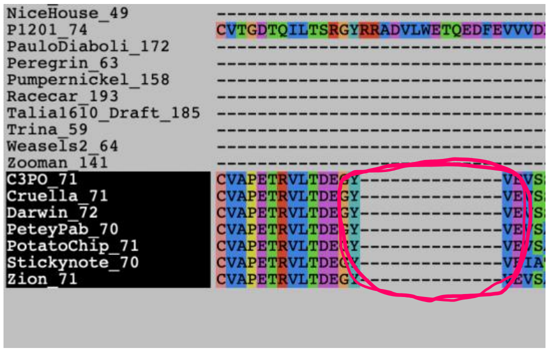
```
mafft --auto --reorder --maxiterate 1000 my_phams.fasta >
my_phams_mafft.fasta
```

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

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

Your answer --->

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.

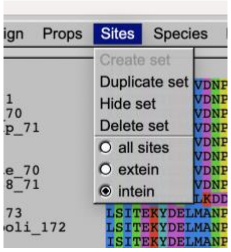
Note, the C-extein now will be out of alignment, you will need to insert gaps in another part (using the spacebar), or delete gaps in all the other sequences.

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

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

To create sets of sites the correspond to the extein and the intein.
First go to Sites 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) 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 Xes to the right. Only the X under the intein should remain.

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
Also save the file as a multiple fasta file: under File > Save selection > enter a name e.g. my_phams_Extein_and_Intein.fst > select fasta > save .

Check in a text editor that the intein and extein fast files are in fasta format.

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

- 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 (-)

D) Predicting structures of intein and extein in alphafold

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

On your laptop open chimeraX (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 open, 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 (use chimera (without the x) and check out

- 1AM2 (mini intein),
- 1lws (yeast V-ATPases intein bound to DNA),
- 1jva (yeast V-ATPases intein with ATPases subunit before splicing),
- 1um2 (yeast V-ATPases intein with ATPases subunit after splicing (chain A and B) (see https://pdb101.rcsb.org/motm/131 for more info on the last two structures; if you select jsmol, this starts a little web embedded structure viewer)
- 1AT0 (Drosophila autocatalytic domain of the hedgehog protein)

Coloring in secondary structure might help.

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 in chimera, color by secondary structure (Tools > Depiction > Secondary structure). You can compare it to some of the above listed intein / hedgehog structures.

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

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

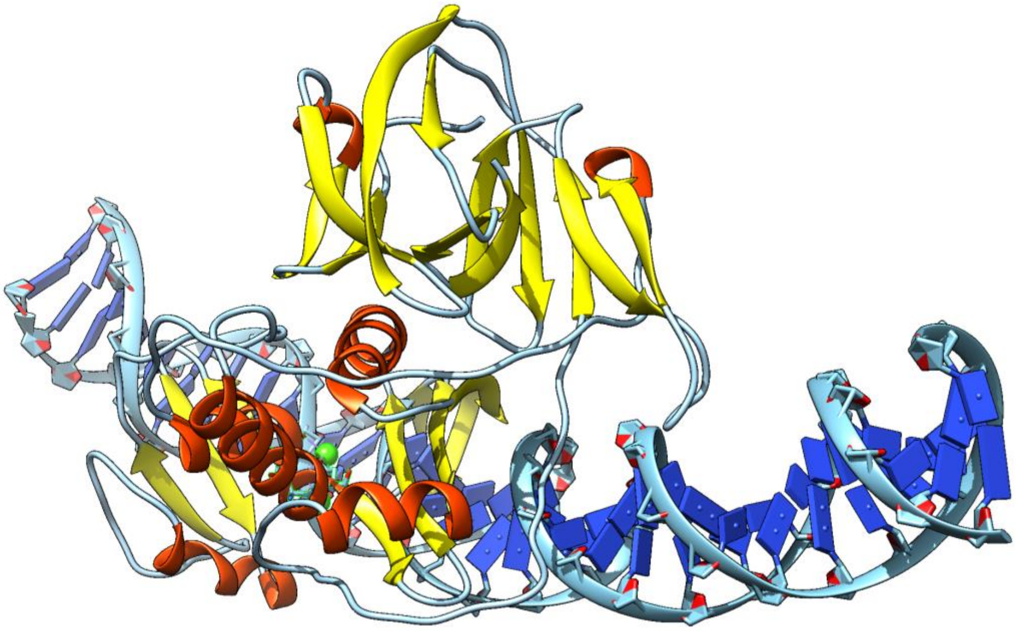
Does predicted structure from the shorter sequence suggest that this is a mini intein (i.e. only containing the splicing domain)?
Your answer --->

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

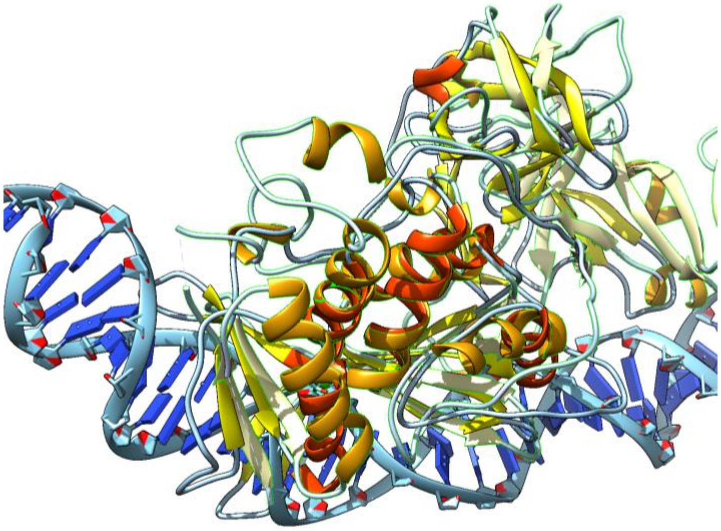
Discuss your findings.
Your answer --->

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



Part of the HE align nicely (orange alpha helices are from 1lws)