**Day 4: Sequence dataset**

**General instruction:** For this activity, you will work in parallel and as a group. Assign a color (grey, green, blue, and pink) to each member based on the alphabetical order of your names. For instance, in a group with Marco, Ana, Peter, and Sam, the alphabetical order would be Ana, Marco, Peter, and Sam. Thus, Ana would be assigned grey, Marco green, Peter blue, and Sam pink.

GROUP MEMBERS NAME:

Grey:

Green:

Blue:

Pink:

The teaching team will circle through the different breakout rooms. If you need assistance, click the *Ask for help* icon and we will come to you as soon as possible.

**Part I: Making names that make sense!**

1. You will use the file **sequences.fa** on the Class website to make a multiple sequence alignment and later, to reconstruct a phylogenetic tree for these sequences.

The format of this file is called FASTA and it ends in *.fa*

*Open sequences.fa in your text editor (NotePad++ for Windows or BBEdit for Mac)*

*The first line starts with >*

*We call > the carrot. After the carrot, the name of a sequence is written. On the next line, the amino acid or nucleotide sequence is written in one letter abbreviations. Another > means another sequence name followed by its sequence and so on.*

FASTA format is a common sequence format. All sequences must have a unique sequence name.

**Use Ctrl+F to figure out how many sequences are in sequences.fa, what should you search for? How many sequences?**

1. FASTA format is important when we make multiple sequence alignments and build trees, but we will also need other sequence formats. Some of these sequence formats will sometimes truncate the sequence names. If uninformative sequence names such as accession codes like P18459 and Q9PU40 are used, these are often harder to make sense of than a name that includes the protein name and the species that the sequence is from. For instance, Tyrosine\_hydroxylase\_Drosophila and Tyrosine\_hydroxylase\_Chicken would be more informative, but these names are long. Thus, we will recode the sequence names (the headers) in a way that makes more sense even when the names are truncated. This will take a little bit of time, but each group will do it together. It is a good time investment because it will make the rest of your analysis easier.

For example, the full-length headers:

>tr|Q9PU40|Q9PU40\_CHICK **T**yrosine **h**ydroxylase OS=**G**allus **g**allus GN=tyrosine hydroxy…

can be renamed as shown in **BOLD** letters immediately after the carrots (below). Note the **space** between the new and old name. The space is important and should be kept: anything after the space will be ignored for some formats. However, keeping this information in your original file is important because it keeps track of how you renamed your sequences which is important for any activity in class and for research.

Your new short names must have **underscores instead of spaces:**

>**TH\_GG** tr|Q9PU40|Q9PU40\_CHICK **T**yrosine **h**ydroxylase OS=**G**allus **g**allus GN=tyrosine hydro…

It helps knowing something about the proteins in your dataset when renaming them. For the two sequences above one says Tyrosine 3-monooxygease and the other Tyrosine hydroxylase. From your Uniprot activity where you worked with these proteins, you may remember that these are the same and we can call both Tyrosine hydroxylase, or TH for short.

The species in sequence.fa (your dataset) are:

Homo sapiens HUMAN [HS]

Mus musculus MOUSE [MM]

Monodelphis domestica MONDO [MD]

Gallus gallus CHICK [GG]

Danio rerio DANRE [DR]

Drosophila melanogaster DROME [DM]

Dictyostelium discoideum DICDI [DD]

**How many sequences from each species are there in the file (use Ctrl+F)?**

1. **Have one member in the group copy all sequences to a shared Google Doc**
2. **Divide the number of sequences by species as evenly as you can among the group members.**
3. **Rename your assigned sequences in Google doc using the following rules:**
	1. **Use the species abbreviation within the brackets above.**
	2. **To rename the protein names, use the 2 or 3 letter codes within the brackets below:**

Phenylalanine hydroxylase or phenylalanine monooxygenase [PAH]

Tyrosine hydrolase or tyrosine monooxygenase [TH]

Tryptophan hydroxylase or tryptophan monooxygenase [TPH]

Tryptophan hydroxylase 1 [TPH1] or tyrosine hydroxylase-like [THL]

Henna [PAH]

Uncharacterized protein [UP]

1. **Copy all sequences from the Google doc with the new names into your own texteditor and save as sequences\_edited\_names.fa**
2. **Download this document and complete Part II individually (but you may still discuss with the group).**

**Part II: Making a Multiple sequence alignment**

1. Open JalView
2. Open **sequences\_edited\_names.fa** in JalView (just like you open a word file in Word. If you cannot see the file sequences\_edited\_names.fa that you just downloaded,change from FASTA to All Files:

****

1. Your sequences are not aligned and they will look like this, except that your names are smarter. Go to Color and Color by Taylor.





Color by Taylor colors the amino acids based on their properties.

Your sequences should now be very colorful, but they are not aligned and it is a busy and straining to look at the sequences this way. So, let’s align them.

1. Go to Webservices, select Muscle with Defaults to align your sequences using Muscle. Make sure that all or none of the sequences are selected before executing the command.



1. Note that the alignment opens in a new Window. Color it by Taylor. What stands out when you look at the alignment? Are any of the sequences messing with the alignment such that it is making it necessary to have many gaps in the alignment?
2. When you have identified the one sequence that seems to mess up the alignment, click on its name and copy the sequence.
3. Go NCBI BLAST and perform a search against the refseq\_proteins database, but use the Organism field below the database field to specify the species that this sequence is from. By specifying the species your results will be much easier to for this task. What is the accession number for the top hit? What is the Description name for this protein?
4. What is the accession number of the best hit that may be TH, TPH, or PAHs?
5. Briefly reflect over your BLAST results and take a guess at what you think may have happened before you continue.
6. Click on the accession number of the top hit to get to its GenPept page. Here, click on the GeneID (located just above the sequence at the bottom of the GenPept page). What is the GeneID?
7. This takes you to the Gene page. Close to the top, you will find Genomic context. This shows where in the genome this gene is found. It also shows its the neighboring genes.
8. Now what do you think may have happened to the long sequence in your dataset?
9. Replace the long sequence in your dataset with the first hit in your dataset that may be TPH (copy its FASTA file and replace it in your dataset using a text editor). Rename it to match your other sequence.
10. Open the fixed file in JalView and align it using Muscle with default. Save it twice, first as FASTA format and second as phylip format.
11. Use the Phylip file to determine how long your alignment is and how many sequences are in your alignment.

Length of MSA:

Number of sequences in MSA:

**Save this document and submit to Canvas for day 4.**

**Save all files in your Bioinformatics folder for next time.**