**Programming exercise, due January 20th 2023:**

The task is to write a single python script that runs a simple bioinformatic pipeline. Your script will loop through input files and call basic bioinformatic tools that you will need to have installed. You are responsible for installing the tools yourself. The input files will be a set of 6 bacterial genomes from the family Campylobacteraceae that I will provide you (on the moodle). You will also select at least 5 proteins you are interested in online. Below are the steps that your script will carry out, along with an example command for the corresponding bioinformatic tools (in grey, no example will be provided for the tree construction and visualization steps). At the end, you will have constructed at least 5 protein trees. You will then submit your script and images of the trees, along with a short interpretation (5 sentences max) of your results. Each step will be marked below (max. 10 points per step) for a total of 80 points.

1. Predict genes using **prodigal** and export the predicted amino acids sequences using prodigal.

prodigal -i input\_file.fasta -a protein\_file.fasta

1. Make blast database(s) from the amino acid sequences using **makeblastdb**.

makeblastdb -in protein\_file.fasta -dbtype prot -out protein\_db

1. For the following step, you will need to do some research a select a set of at least 5 proteins and then obtain representative sequence from the *Campylobacter jejuni* NCTC11168 genome. There are slides on the moodle showing you how to do this on NCBI. Save these proteins into a file that will be used to search the amino acids sequences from your genomes. Then use **blastp** to then search for these proteins in your genomes and output your results to a table.

blastp -query query\_proteins\_from\_ncbi.fasta -db protein\_db db -max\_target\_seqs 1 -evalue 1E-5 -outfmt "6 qseqid sseqid pident evalue sstart send length" -out blastp.results.txt

1. Use the output table(s) to select homologues and combine the homologues into individual fasta files for each query sequence. You may do this by hand and will receive 50% of the points for this step. If you automate selecting and compiling the fasta files in the python script, you will receive full points.
2. For each of the fasta files of protein homologues, run **MUSCLE** to align then.

muscle -in homologous\_proteins.fasta -out aligned\_homologous\_proteins.afa

1. Make a phylogenetic tree using **iqtree.** This step can be conducted online to receive 50% of the points (<http://www.iqtree.org/>). If you automatic the generation of trees, as above, with your python script you will receive full points. You can find example commands for running iqtree on their website.
2. Visualize each of the trees using any tree visualizing software you want. Provide an imagine (a screen shot is fine) of each of the trees and give them a title.
3. Provide a brief interpretation of all your results (5 sentences max). Submit your script, your tree images and your interpretation.