**BIOMI 609 Computational Genomics and Bioinformatics**

**Spring 2022**

**San Diego State University**

**Dr. Arun Sethuraman**

**Lab 3 - Phylogenomic Reconstruction**

This week, we learned the theory behind phylogenetic reconstruction - we’ll put these skills to practice in today’s lab, by constructing a multiple sequence alignment of SARS-CoV2 whole genomes, followed by phylogenomic reconstruction. Prior to this, we have to first download and install a few things:

**Exercise 0: Installing MAFFT, RAxML**

First we will install the ultra-fast multiple sequence aligner, MAFFT (<https://mafft.cbrc.jp/alignment/software/>). To do this, move into your Tools folder on JetStream, then:

wget <https://mafft.cbrc.jp/alignment/software/mafft-7.490-with-extensions-src.tgz>

tar -zxvf mafft\*.tgz

cd maff\*

cd core

make

sudo make install

#This should automatically add mafft to your path, so you don’t have to do this separately

Thereon, we will install RAxML from source.

#Make sure that you’re inside your Tools folder

git clone <https://github.com/stamatak/standard-RAxML.git>

cd standard-RAxML

make -f Makefile.MPI.gcc

**Exercise 1: Constructing a Multiple Sequence Alignment**

Recall that last week, we’ve already obtained multiple sequences from SARS-CoV2 (saved as deltas.fasta in my case). We will work with that as input to MAFFT to construct a multiple sequence alignment. To construct the alignment, the command is as simple as:

mafft deltas.fasta > deltas\_aligned.fasta

MAFFT is extremely fast, and should generate an alignment within seconds, in this case stored into deltas\_aligned.fasta

**Exercise 2: Constructing a phylogeny using RAxML**

Once the alignment is completed, the next step is to construct and visualize a phylogeny. This is super easy using RAxML as well!

raxmlHPC -s deltas\_aligned.fasta -m GTRGAMMA -p 12345 -n gamma\_covid

#-m specifies the model

#-n is just a name for the run

#-p is a random seed assigned to this run

This run shouldn’t take long either - you’ll notice that the output files have a “RAxML” prefix, with the best tree printed out in a file called “RAxML\_bestTree\_gamma\_covid”. Now open this, copy the NEWICK string printed, and upload it into the online tree viewer: <http://etetoolkit.org/treeview/>

Then hit “View Tree”. Voila!

**Exercise 3: Working with larger datasets**

Here comes the fun part - the whole goal of you running smaller data is to pave the way for analyses using large datasets. Now consider that you want to reconstruct the phylogeny of an entire lineage of SARS-CoV2 genomes, millions of which are publicly available via NCBI. So navigate to <https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/sars-cov-2> inside your VM, which is an interactive hub for all individual protein sequences from millions of genomes. It’s not feasible or smart in the least to attempt to build a phylogeny using whole genomes, but we can very well do this for a single gene/protein sequence. So let’s try that!

Click on “All proteins”, and this should take you to an interactive table - refine the results to obtain sequences from one particular geographical region, and for one protein only (e.g.orf1a). In my example, I used orf1a protein sequences from across Europe (ended up being 2718 total sequences). I downloaded this by clicking on the “Download” button (see below) by the “SARS-CoV-2 Data Hub” icon. Select “Protein”, say “Download all records”, then “Build custom” to pick at least the name of the geographical region from which you are analyzing data from. This will permit us to make sense of the phylogeny later (see screenshot below). Then click on “Download”.

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

In my case, I ended up with a file, that I just renamed to orf1a\_europe.fasta to make things easier. I created a new folder called Week7, moved this file into that. Note that this is still a pretty gigantic multiple sequence FASTA file (~12 Mb). So we will run MAFFT in parallel, and using the reference genome as a “guide”, to make things more efficient. To do this, we first also need to download the orf1a nucleotide sequence for the Wuhan (reference) strain. I just searched for this on GenBank <https://www.ncbi.nlm.nih.gov/gene/43740578>

Navigate down to “Reference assembly”,

mafft --thread 5 orf1a\_europe.fasta > orf1a\_europe\_aligned.fasta

Thereon run RAXML (note that this will take a long time):

raxmlHPC -T 6 -m PROTGAMMAGTR -s orf1a\_europe\_aligned.fasta -p 12345 -n orf1a\_europe

Thereon visualize the tree as before, or you can also install and use FigTree for this purpose, which allows a much better visualization. <https://github.com/rambaut/figtree/releases>