CSE185 Lab 7 Report - Code Documentation (10 pts)

- Document any commands used or additional analysis steps below!
- You should include enough detail that the instructors (or your future self) could come back to this several months from now and know exactly what you did and why you did it.
- We will not run this notebook, but will look back to see what you did especially if you end up with different answers.

For grading purposes only - Do not copy or edit this cell!

Part 1 - Downloading the data

awk '{print "wget -P lab7 "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nuccore&id=%22$1%22&rettype=fasta&retmode=text%5C.fasta%22%22%7D' ~/public/lab7/lab7_accessions.txt | sh && cat lab7/*.fasta > ~/lab7/lab7_virus_genomes_unedited.fa

Code to remove empty lines from fasta file

grep -v '^$' ~/lab7/lab7_virus_genomes_unedited.fa > ~/lab7/lab7_virus_genomes.fa

Part 2 - Performing Multiple Sequence Alignment

2.1 Install mafft

In [ ]:

```
# Command to git clone mafft
git clone https://gitlab.com/sysimm/mafft.git

# To edit topline in Makefile to install to home directory
vi Makefile

# commands to make and install
make clean
make
make install
```

2.2 Run mafft

In [ ]:

```
# Command to run mafft interactively
mafft
```

I used lab7_virus_genomes.fa as the input file and chose Output format as 5. Phylip format / Sorted and used the default strategy of 3. FFT-NS-2 (default). I then had the output file as lab7_virus_genomes.aln.

Part 3 - Building a tree

3.1 Install RaxML

In [ ]:

```
# Install the .tar.gz file for v8.2.12
gunzip standard-RAxML-8.2.12.tar.gz
tar xf standard-RAxML-8.2.12.tar

# Next are the commands to compile it for the SSE3.PTHREADS
make -f Makefile.gcc
rm *.o

make -f Makefile.SSE3.gcc
rm *.o

make -f Makefile.PTHREADS.gcc
rm *.o

make -f Makefile.SSE3.PTHREADS.gcc

# copy binary file to local path
cp raxmlHPC-PTHREADS-SSE3 ~/local/bin/raxml
```

```
# if raxml command isn't showing up
export PATH=$PATH:$HOME/local/bin
```

3.2 Run RaxML

In [ ]:

```
# Simple ML Search
raxml -s lab7_virus_genomes.aln -m GTRGAMMA -p 12345 -n T1

# Bootstrapping Search
raxml -s lab7_virus_genomes.aln -# 100 -m GTRGAMMA -p 12345 -b 12345 -n T2

# Bipartitions
raxml -m GTRCAT -p 12345 -f b -t RAxML_bestTree.T1 -z RAxML_bootstrap.T2 -n T3
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js