ReadMe File

Group: 4 | Group Name : Go Corona

Project Drive Link

https://drive.google.com/drive/folders/1nGe6umANoqar3DZWGkftu1YDN2sm1QWi?usp=sharing

Running the Notebook

- **Input Files**
  US_aligned_sequence.fasta - USA dataset
  India_aligned_sequence.fasta - India dataset
  India_metadata.tsv - India metadata
  USA_metadata.csv - USA metadata
  decoder, encoder - model to reduce dimension
  output.vcf - contains position of SNPs
  model1_checkpoint.h5
  model2_checkpoint.h5
  model3_checkpoint.h5 - LSTM models

- **Output Files**

  ARIMA.txt - result using only ARIMA model
  ARIMA_SNP.txt - result using ARIMA and SNP position
  Neural.txt - result using only Neural network
  Neural_SNP.txt - result using Neural and SNP position

- **How to Run**
  Flu_Sequence_Detector.ipynb - Python notebook should be loaded on jupyter but Google colab is recommended as it contains all the libraries preinstalled. Also, you can use GPU power to get results faster.
  You need the following steps to run program
  1. All the input file should be loaded and then run the colab runtime.
  2. Change the names of file according to dataset that you want to predict

For USA

fasta file -> US_aligned_sequence.fasta

metadata file -> USA_metadata.csv

For India

fasta file -> India_aligned_sequence.fasta

metadata file -> India_metadata.tsv

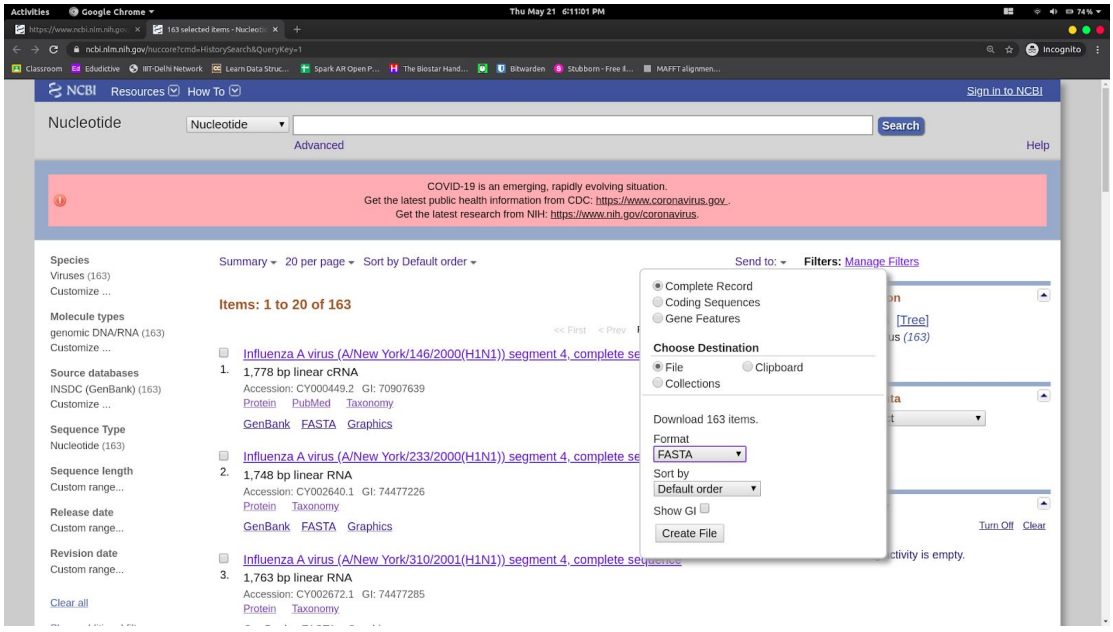
3. Decoder, encoder and LSTM model is for USA data. If you are using indian data then you should make sure to make it again, code is included in the main python file.

4. There are some cells specifically for the USA and India respectively. So run accordingly

Obtaining SNPs

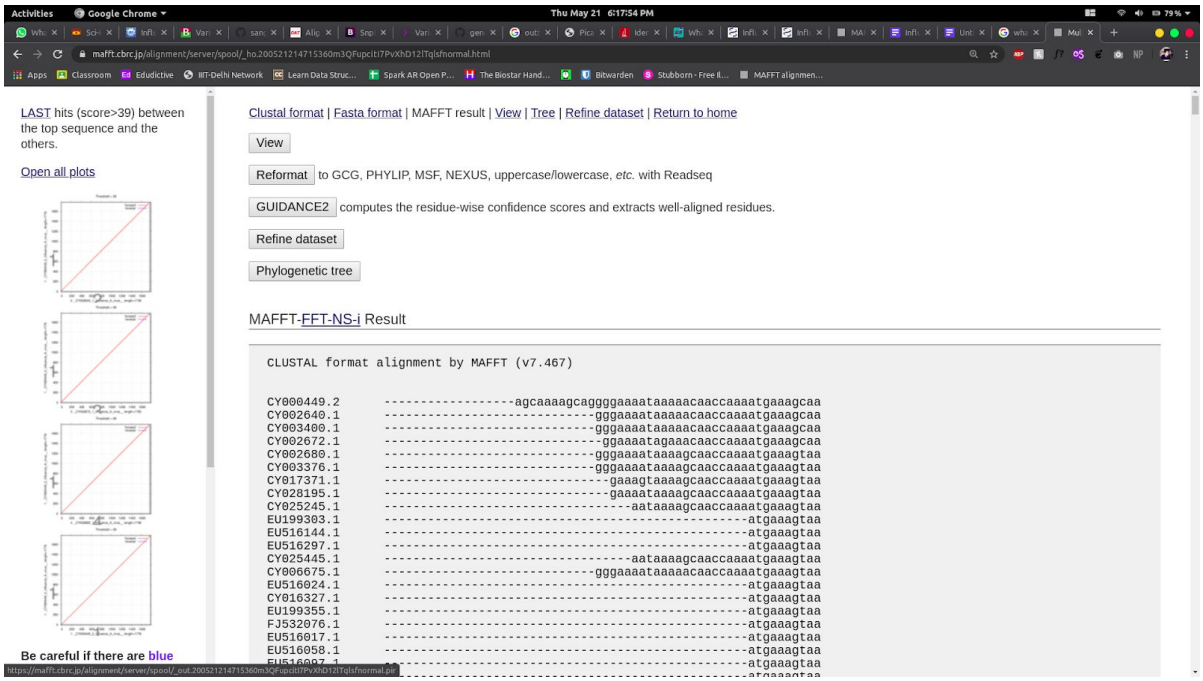
1. Getting the combined FASTA sequences

- For getting combined fasta sequences, we can use the Batch Entrez tool from NCBI.
- Go to https://www.ncbi.nlm.nih.gov/sites/batchentrez and select the **filtered-accessions.txt** file for **Nucleotide Database.**
- The mentioned 163 accessions in filtered-accessions.txt are unique to a specific interval to time to avoid redundant sequences to align.
- Download the accessions in FASTA format as shown below. The file is already included in the project with the name **sequence.fasta**



2. Aligning the sequences using MAFFT

- Head over to https://mafft.cbrc.jp/alignment/server/ for accessing the MAFFT server.
- Select the sequence.fasta file and just submit the job to execute.
- We get the alignments as shown. Download the alignments in the fasta format and save it as **aligned-output.fasta**



3. Extracting SNPs

- Now when we have got the alignment file with us, we need to extract the SNPs from it. For this we will be using **snp-sites**. SNP-sites can rapidly extract SNPs from a multi-FASTA alignment using modest resources and can output results in multiple formats for downstream analysis in a time and space efficient manner.
- For installations instructions, go to their github repository: https://github.com/sanger-pathogens/snp-sites
- In Debian environment we executed the simple command to install snp-sites:

```
sudo apt-get install snp-sites
```

- Now in the directory where, you have downloaded the **aligned-output.fasta.** Execute the following command:

```
snp-sites -mvp -o output aligned-output.fasta
```

-m	output a multi fasta alignment file (default)
-v	output a VCF file
-p	output a phylip file
output	output file name

- We now have 3 new files including .vcf file (shown below) which will be used to get the SNP positions in using scikit-allele library in the notebook itself.