Predicting New COVID-19 Mutations and Generating Predicted Sequences as Time Series Data

Given Sample Experiment Lists

(Experiments 4, 5 ,6 are the most important ones)

**********Evolution of Virus with Time **********

Experiment 1:
Perform Multiple Sequence Alignment (MSA) of the unique membrane protein sequences of SARS-CoV-2 from Dec 2019 to Dec 2022.

Goals:

1. Understand how the membrane protein evolved with time. We like to see how mutations change over time.

Methods:

1. The MSA can be performed using MAFFT and visualized using Jalview.
2. Report the occurrences of each mutation in a table format.
3. Create a visualization to show how the mutation positions change over time for membrane proteins.

Experiment 2:
Perform Multiple Sequence Alignment (MSA) of unique spike protein sequences of SARS-CoV-2 from Dec 2019 and Dec 2022.

Goals:

1. Understand how the spike protein evolved with time. We like to see how mutations change over time.

Methods:

1. The MSA can be performed using MAFFT and visualized using Jalview.
2. Report the occurrences of each mutation in a table format.
3. Create a visualization to show how the mutation positions change over time for spike protein.

**********Variant Analysis **********

Experiment 4:

Build phylogenetic trees of different variants of SARS-CoV-2 and map them back to the complete phylogenetic tree of all sequences.

Goal:

1. Check if the pattern of individual variant phylogenetic trees is preserved in the complete phylogenetic tree that contains sequences from all variants.

Methods:

1. Collect sequences of each variant from the GISAID EpiCoV database.
2. Perform Multiple Sequence alignments and create a phylogenetic tree for each variant.
3. Repeat the process with sequences from all variants and validate the pattern.

**********Time Series Forecasting **********

Experiment 4:
Create a Dataset with Mutation position and sequence Data to perform Time series forecasting.

Goals:

1. Perform time series forecasting with LSTM network to predict upcoming mutations of SARS-CoV-2 based on previous sequences. For example, predict 7 days of mutations based on the past 90 days of SARS-CoV-2 sequences.

Methods:

1. Collect all sequences from GISAID between December 2020 and Dec 2022 for each variant. (Need to register and login to access the EpiCoV database.
2. You can use the web scraping technique to collect all sequences at once and download the FASTA format and sequence technology metadata.
3. Then, use Coronavirus Genotyping Tool to find the nucleotide and protein mutations for each sequence.
4. Create a dataset with all collection dates and mutation sites such that if the sequence of a particular date contains that mutation, it will contain 1 otherwise, it will contain 0. Here is the Sample file.

********** Time Series Mutations **********

Experiment 5:
Create a visualization to compare the mutation rate of different variants of SARS-CoV-2.
Goal:

1. Understand the rate of mutation change for each variant over the period.

Methods:

1. Collect sequences of each variant from the GISAID EpiCoV database.
2. Calculate the new number of mutations for each day.
3. Plot the graph to show mutation rates over time.

Experiment 6:
Find the reoccurring mutation of SARS-CoV-2.
Re-occurred mutation: We'll identify mutations as re-occurring mutations if they happened earlier, then probably disappeared and came back in some other virus sequences.

Goal: Understand the nature of mutations to identify their re-occurrence. We need to analyze the functionality of those mutations.

Methods:
1. Collect SARS-CoV-2 sequences from GISAID EpiCoV Database.
2. Find mutations in each sequence using MSA and Coronavirus Genotyping Tool.
3. Analyze the occurrence of mutations over time and identify re-occurring mutations.

Experiment 8:
Correlate mutations with time and correlate mutations themselves.
Co-mutation: If two sites have mutation occurrences simultaneously, we call those mutations co-mutations.

Goal:
1. Find out the functionality associated with co-mutation.

Methods:
1. Collect SARS-CoV-2 sequences from GISAID EpiCoV Database.
2. Find mutations in each sequence using MSA and Coronavirus Genotyping Tool.
3. Analyze the occurrence of mutations over time and identify co-mutations.