NORTHERN ILLINOIS UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE

BIO-INFORMATICS -I

TERM PROJECT

Final Report

Phylogenetic Analysis of Multi SARS-Like Genomic Segments

Semester: FALL 2023

Team Members:(Team 8)

Vikramaditya Reddy Varkala Z1973679
Edara Sai Subash Z1948230
Spandana Regatti Z1981721

Professor: Dr. Minmei Hou

Abstract

This study conducts a comprehensive phylogenetic analysis of multiple SARS-like sequences to understand their evolutionary relationships. Employing the UPGMA and Neighbor-Joining (NJ) algorithms, we analyzed these sequences to identify conserved regions and evolutionary links. Our approach included data preparation, sequence processing, and phylogenetic analysis, culminating in detailed visualization and post-processing of the results. The study provides insights into the genetic diversity and evolutionary patterns among various SARS-like viruses, contributing to the broader understanding of their origins and mutations.

Introduction

The emergence and spread of SARS-like viruses have prompted significant concern in global health. Understanding the phylogenetic relationships among these viruses is essential for tracking their evolution and informing public health strategies. This report details our bioinformatics approach to analyze multiple SARS-like sequences, aiming to elucidate their evolutionary connections and variances.

Work Environment, Programming Language, Library Packages and Methods:

Work Environment: Visual Studio

Programming Language:

We have used Python for this project development for its extensive support for scientific computing and data analysis. Its readability and the availability of specialized libraries makes it ideal for this project.

Libraries:

BioPython: A powerful tool for biological computation. You can use several modules from BioPython

Bio.Phylo.TreeConstruction: For constructing phylogenetic trees using algorithms like UPGMA and Neighbor-Joining

Bio.Align: For handling multiple sequence alignments.

Bio.AlignIO: For reading and writing sequence alignments.

Bio.SeqRecord: For handling individual sequence records.

Bio.Seq: For handling biological sequences.

Matplotlib: A comprehensive library for creating static, interactive visualizations in Python. Used here for plotting phylogenetic trees.

DendroPy: A Python library for phylogenetic computing. It is used for tree comparison and handling phylogenetic trees in various formats.

OS: A standard Python library for interacting with the operating system, like file and directory manipulation.

Random: This module is used for generating pseudo-random numbers, here for setting a seed for reproducibility of results.

Methods used in the Code :

deletePreviousFiles: This function removes files in a given directory, ensuring a clean start for the analysis.

get_gene_positions: Reads a text file to extract gene positions for SARS-CoV-2.

get_genome_list: Extracts a list of genome IDs from the alignment file.

extract_gene_by_position: Extracts specific gene sequences from the whole genome sequences based on given start and end positions.

get_conserved_regions: Identifies and extracts conserved regions from the multiple sequence alignment (MSA).

Phylogenetic Tree Construction: The script constructs phylogenetic trees using two methods:

UPGMA (Unweighted Pair Group Method with Arithmetic Mean

Neighbor-Joining (NJ)

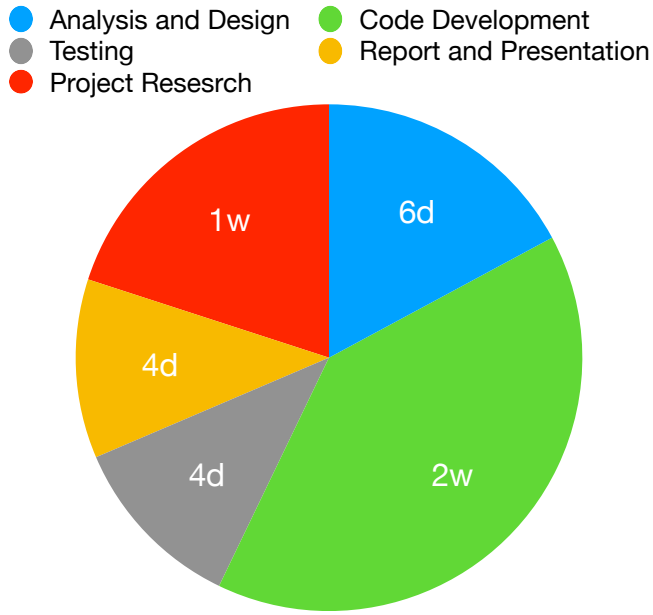
Visualization and Output: Trees are visualized using Matplotlib and saved as PNG images. The trees are also written in Newick format.

DendroPy Processing: The Newick format trees are further processed using DendroPy for additional analysis.

Team Members Contribution:

The overall efforts and contributions from team members are detailed below.

Timeline:



Individual Contributions:

	Vikramaditya	Subash	Spandana
Analysis and Design	30%	35%	35%
Code Development	35%	35%	30%
Testing	35%	35%	30%
Report and Presentation	35%	30%	35%
Project Research	33%	33%	34%

Initial Analysis

Source of Data File:

We have downloaded **multi.sars-like.maf** and **sarsCov2structure.txt** from department server using FileZilla provided by Professor. Minmei Hou, **sarsCov2structure.txt** contains information about the position of different SARS genomic segments that helps with to identify the type of sequence we are dealing with in multi-sars-like.maf file.

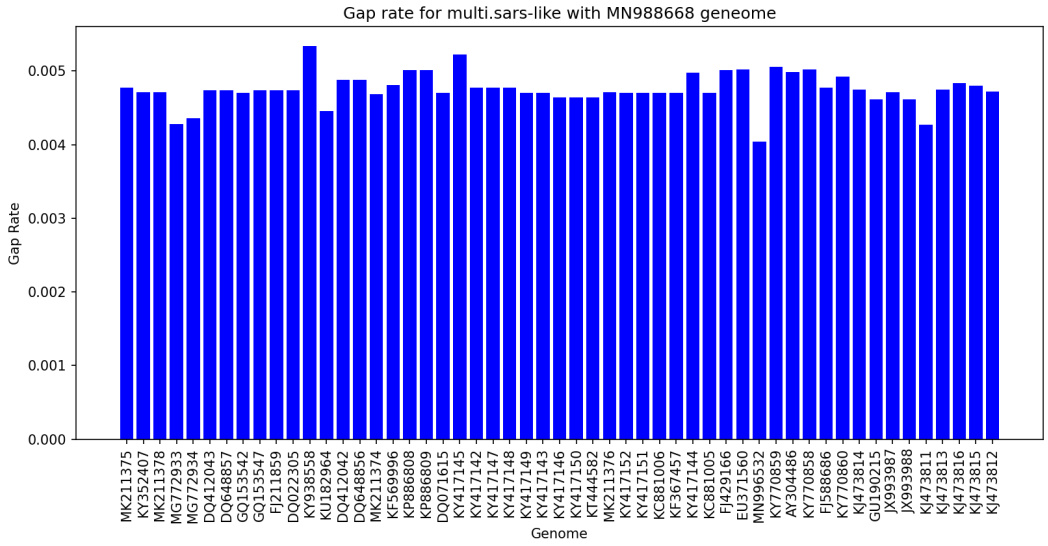
Initial Data Analysis:

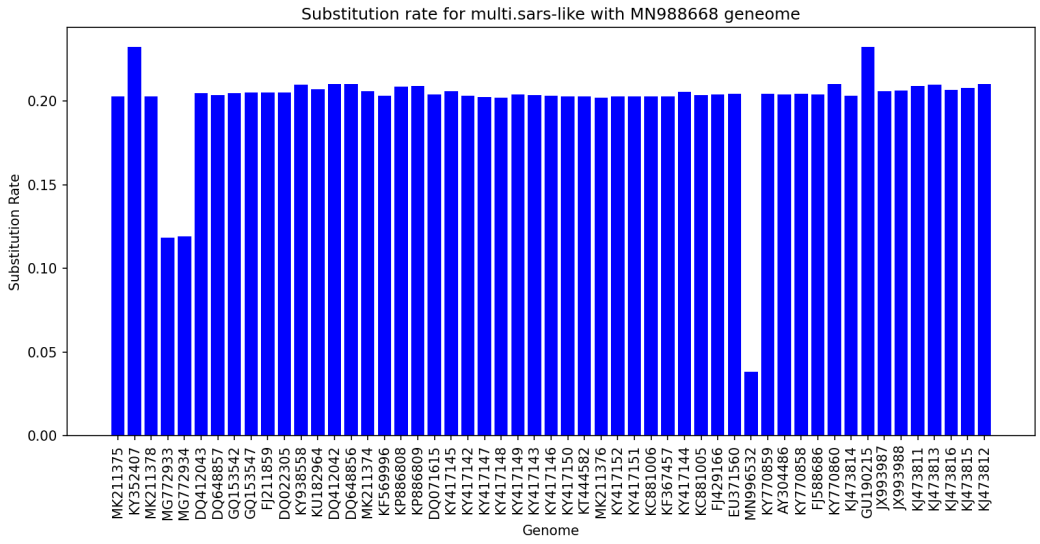
We have compared MN988668 variant with rest of the 53 variants and found out Gap Rate and substitution rate.

Substitution rate = number-of- mismatches / (number-of-mismatches + number-of-matches)

Gap Rate = number of gaps / (Number of matches + number of mismatches + number of gaps)

Below are the outputs(gap rates and substitution rates in the form visualizations of the analysis.





Data, Methods,Preliminary Analysis and Results

Data Preparation:

Our analysis began with preparing the data, involving cleaning directories, reading genome lists, and extracting gene positions from various SARS-like sequences.

Sequence Processing:

The sequences were processed by extracting gene sequences for each virus and identifying conserved regions, focusing our analysis on critical genetic segments.

Phylogenetic Analysis:

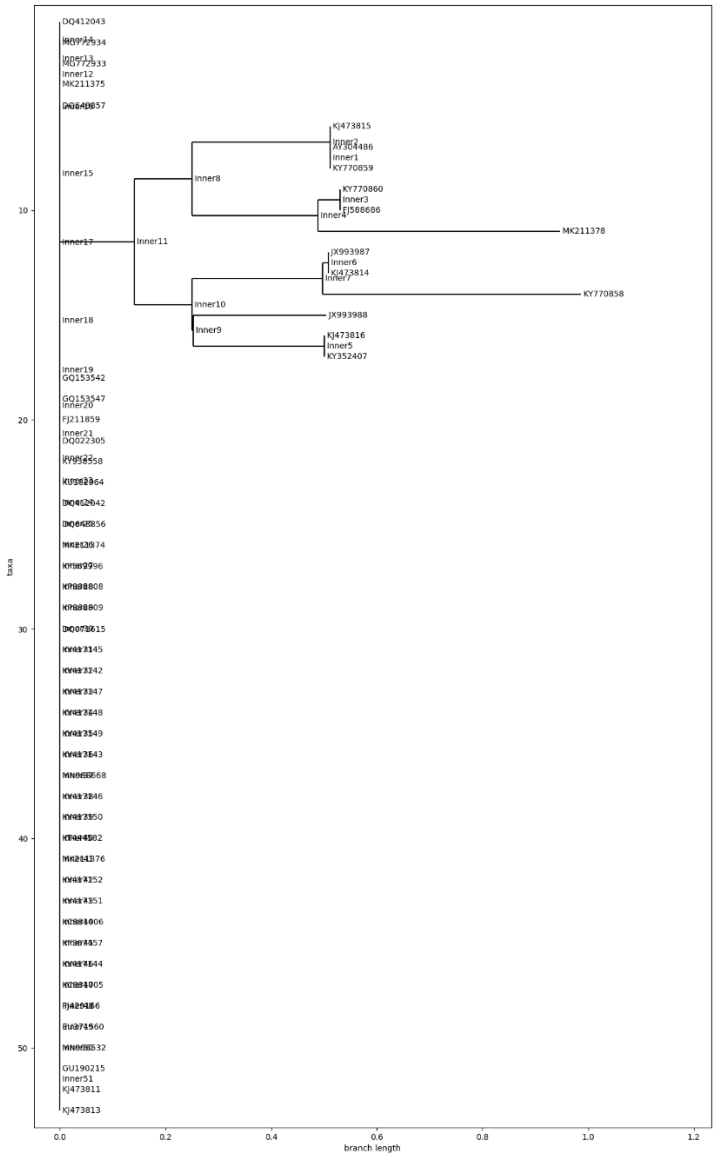
Phylogenetics is the study of the evolutionary history and relationships among individuals or groups of organisms. These relationships are discovered through molecular sequencing data and morphological data matrices.

Neighbor-Joining (NJ) Method

Description: The NJ method is a bottom-up, agglomerative clustering technique used in phylogenetics. It is best known for its use in the creation of phylogenetic trees (dendrograms) based on genetic distance data. NJ's key advantage is its efficiency in handling large datasets, making it suitable for our comprehensive analysis of SARS-like sequences.

Application: In our analysis, NJ was applied to construct phylogenetic trees from a distance matrix, computed from the aligned sequences of the SARS-like viruses. This method helped reveal the evolutionary relationships based on genetic distances.

Results Interpretation: The resulting NJ trees provided a snapshot of the evolutionary history of the viruses, revealing the relative genetic distances and potential common ancestors.



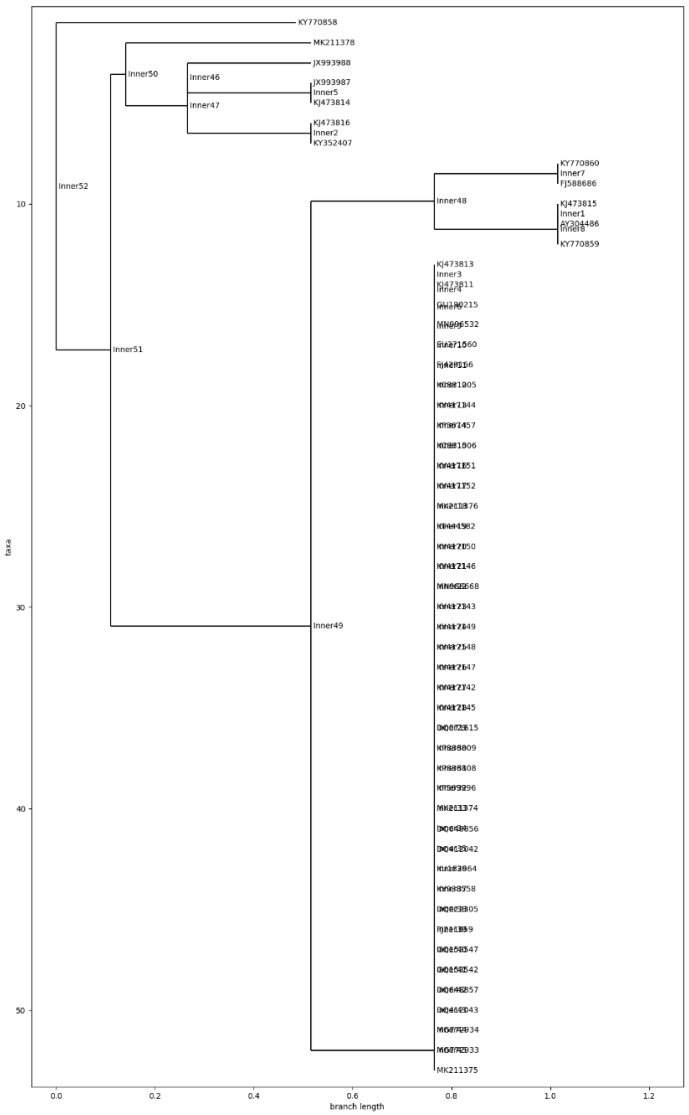
Sample Output(NJ) - S Type

Unweighted Pair Group Method with Arithmetic Mean (UPGMA):

Description: UPGMA is a simpler method of phylogenetic tree construction that assumes a constant rate of evolution (molecular clock hypothesis). It's an algorithm used for hierarchical clustering in phylogenetics that aggregates species into progressively larger clusters.

Application: We utilized UPGMA to build phylogenetic trees, where the distance between groups will be calculated between members of the groups. This method was particularly useful in providing an alternative perspective on the evolutionary relationships, complementing the NJ method.

Results Interpretation: The resulting UPGMA trees provided a snapshot of the evolutionary history of the viruses, revealing the relative genetic distances and potential common ancestors.



Sample Output(UPGMA) - S Type

Robinson Foulds distance:

The Robinson-Foulds distance, named after David F. Robinson and Leslie R. Foulds who introduced it in 1981, is a measure used to compare phylogenetic trees. It's a method of quantifying how similar or dissimilar two phylogenetic trees are.The core idea behind the RF distance is to compare the partitions (bi partitions) of a set of taxa (species) that are created by the branching points (internal nodes) of each tree.
In simple terms, each internal node in a phylogenetic tree divides the tree into two groups of taxa, and these groups are compared between trees.A lower RF distance indicates higher similarity.
The Image on the right gives the distances between gene pairs using UPGMA and NJ method.

Robinson Foulds distance Data Frame

	Genomes	UPGMA	NJ
0	Whole-orf1a	5.4864	4.7700
1	Whole-orf1b	5.7565	4.9897
2	Whole-S	5.3777	4.2771
3	Whole-E	4.0964	3.6677
4	Whole-M	5.2058	4.4073
5	Whole-N	5.0856	3.9348
6	orf1a-orf1b	2.6753	2.2834
7	orf1a-S	6.0207	5.1363
8	orf1a-E	5.9857	5.7474
9	orf1a-M	5.4437	4.7951
10	orf1a-N	5.0115	4.2971
11	orf1b-S	6.1852	5.1017
12	orf1b-E	5.8848	5.5425
13	orf1b-M	5.4222	4.9381
14	orf1b-N	5.6074	4.4426
15	S-E	4.2813	3.6316
16	S-M	5.7240	4.8861
17	S-N	6.1602	5.0197
18	E-M	4.7760	4.7183
19	E-N	5.6289	5.1523
20	M-N	5.5299	5.5533

distance outputs(UPGMA AND NJ)

Visualizing the Robinson Distance:

To compare the genetic distances across all gene types and whole genomes, we have used heat maps. These heat maps will visually represent the Robinson-Foulds , allowing for an immediate grasp of the genetic variability and similarities among the sequences.
This approach will enable us to identify patterns and outliers in the data effectively.

Image 1 depicts a heat map which is comparing UPGMA Robinson-Foulds distance among whole genome and other sub gene types.The S type and E type has highest Robinson-Foulds distance of **6.4100**. The orf1a and orf1b type has lowest Robinson-Foulds distance of **2.0000**.

Image2 depicts a heat map which is comparing Nj Robinson-Foulds distance among whole genome and other sub gene types.The orf1a type and M type has highest Robinson-Foulds distance of **5.6900**.The orf1a and orf1b type has lowest Robinson-Foulds distance of **1.6900**.

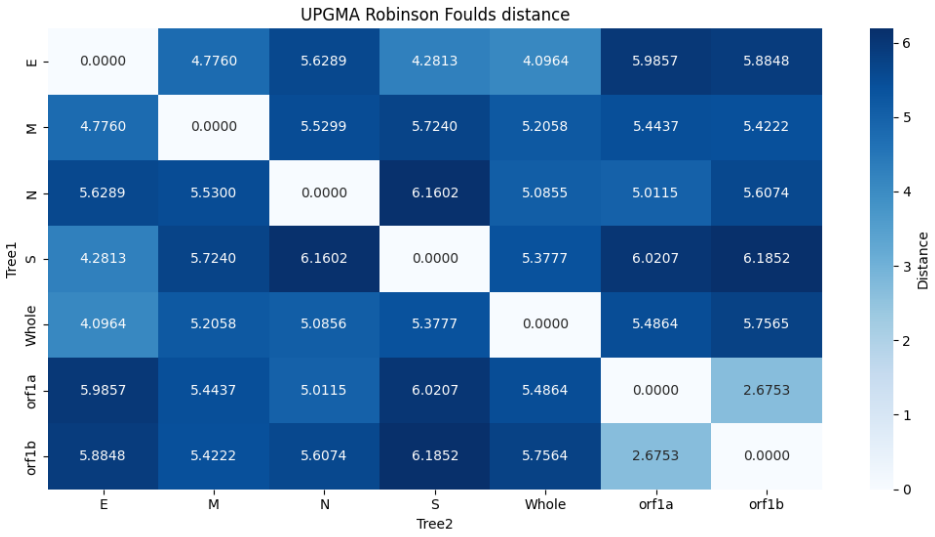


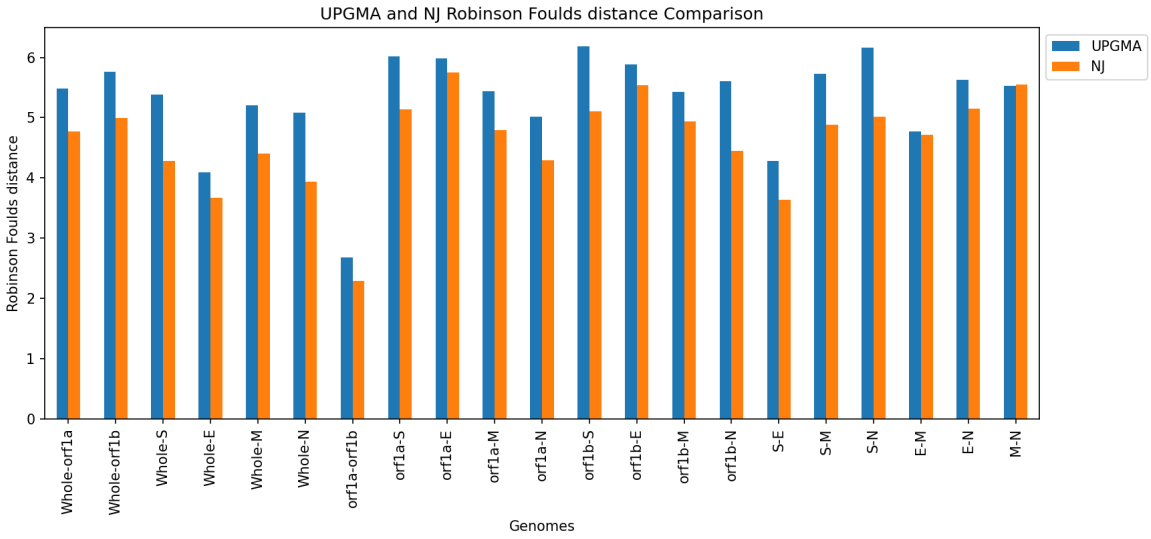
Image 1(UPGMA)

Image 2(NJ)

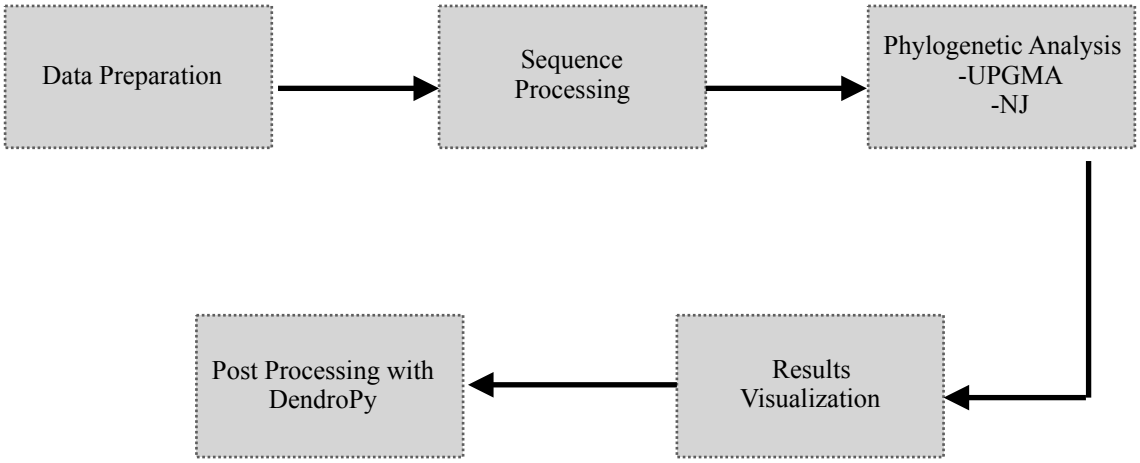
Comparative analysis between 2 methods:

We have used Bar graph to compare the RESULTS obtained from the UPGMA and NJ methods. This comparative analysis will help in understanding how each method portrays the phylogenetic relationships and the implications of their inherent assumptions.These bar graph will provide a clear visual representation of the differences and similarities in the phylogenetic trees generated by both methods

This Bar Graph below shows the comparison of Robinson-Foulds distance of Genomes using both the UPGMA and NJ distance methods .In this bar graph , we can see that the Robinson-Foulds distance of NJ method is less than the UPGMA method.So , the NJ method is better at comparing similarities between two sub-genome types compared to UPGMA.



Phylogenetic Analysis Workflow:



Interactive Tree of Life tool:

Interactive Tree Of Life (iTOL) is an online tool designed for the display, annotation, and management of phylogenetic trees.

We have used **Interactive Tree Of Life** tool to create this circular representation of whole genome.

Image1 below is the circular representation of Whole genome Phylogenetic tree using NJ Method.

Image2 below is the circular representation of Whole genome Phylogenetic tree using UPGMA Method.

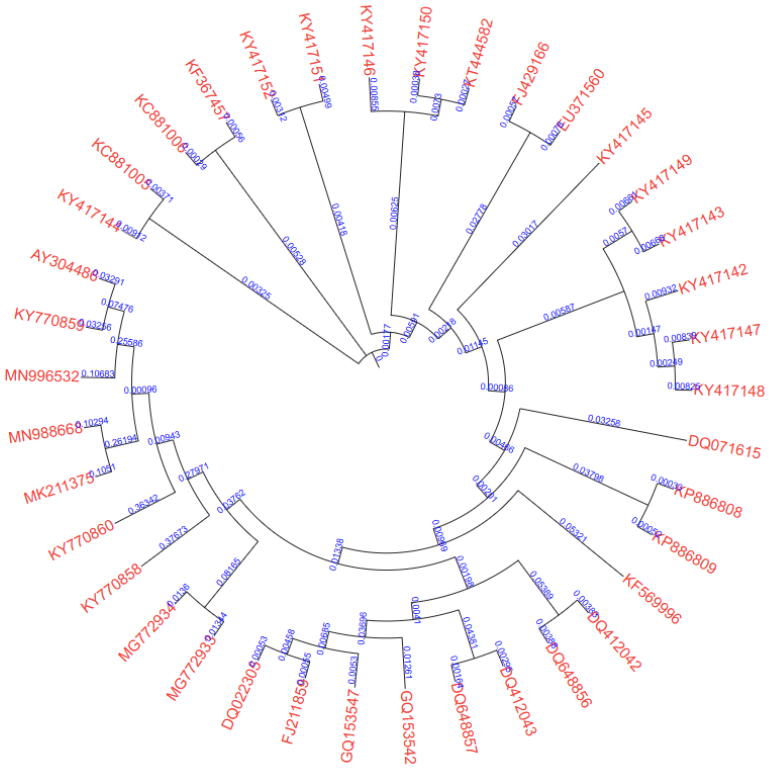


Image1(NJ_Whole)

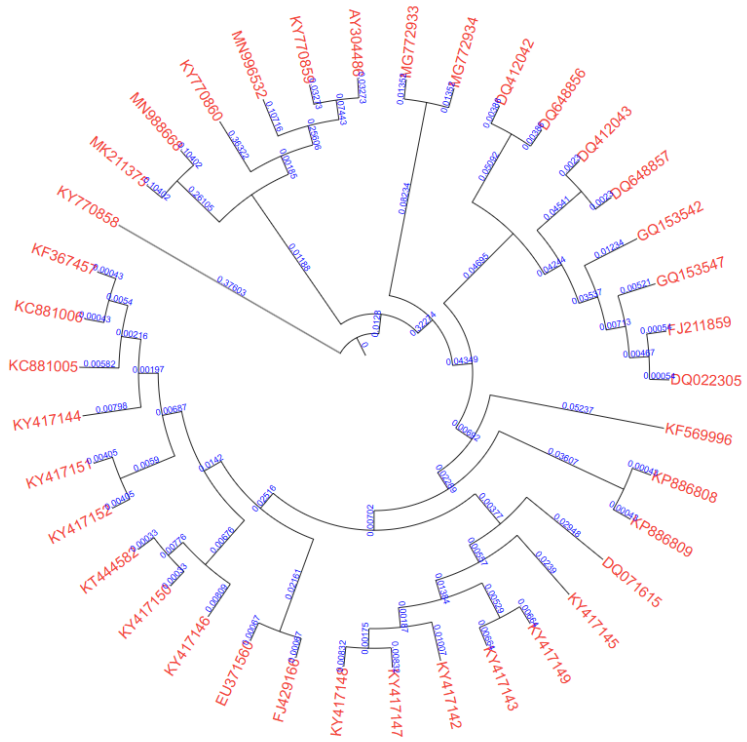


image2(UPGMA_whole)

Conclusion:

This study effectively analyzed Multi SARS like genomic sequences, extracting gene locations and identified conserved areas through the use of bioinformatics techniques. Using the UPGMA and NJ algorithms, phylogenetic trees were built, revealing information about the evolutionary relationships between various viral strains.We were able to find NJ tree was better at finding similarity between 2 gene types.

Our study also highlighted the impact of using conserved regions versus whole genomes for phylogenetic analysis. Trees built using conserved regions were more divergent, underscoring the importance of selecting appropriate genomic regions for accurate evolutionary interpretations.

Overall, this project not only provided insights into the genetic dynamics of SARS but also demonstrated the critical role of methodological choices in phylogenetic analysis.

The robustness of these phylogenetic conclusions was confirmed by a comparison analysis using the Robinson Foulds measure. The project's overall findings demonstrate how computational biology can effectively comprehend the evolution of viruses.

References:

https://itol.embl.de

https://biopython.org/wiki/Documentation

https://dendropy.org/#documentation

https://en.wikipedia.org/wiki/Robinson–Foulds_metric

https://matplotlib.org/stable/gallery/images_contours_and_fields/image_annotated_heatmap.html