Spatio-Temporal Analysis of COVID-19 Genomic Sequence Similarity

CL726: Introduction to Genomics

Course Project Report

Members

Adit Agrawal: 200040009

Atharva Jadhav: 200020033

Gaurav Awasthi: 200020036

Prakriti Shetty: 200020095

Guide: Prof. Sarika Mehra

Department of Chemical Engineering
Indian Institute of Technology Bombay

02 May 2024

Table of Contents

Introduction

The global COVID-19 pandemic caused by the Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) highlighted the importance of viral genomics in understanding and combatting infectious diseases. Since its initial emergence, researchers across the world have focused on characterizing the SARS-CoV-2 genome. This project delves into this crucial area of research by leveraging the vast collection of SARS-CoV-2 genome sequences deposited in the National Center for Biotechnology Information (NCBI) database.

Our project utilizes alignment algorithms and powerful computational tools that allow us to compare the genetic sequences of SARS-CoV-2 isolates from different geographical regions and time points. Through this comparative analysis, we aim to identify and characterize mutations within the viral genome. Understanding these mutations is essential for uncovering how the virus has evolved during the pandemic. Specifically, we are interested in exploring how these genetic variations might influence the virus's transmissibility, virulence, or ability to evade the immune system.

The findings from this project can contribute to our overall understanding of SARS-CoV-2. By shining light on the evolutionary patterns of the virus, effective public health initiatives can be strategized. This knowledge can be used to guide strategies for containment, develop more targeted vaccines, and design effective therapeutic interventions. Ultimately, this project represents a small but crucial step towards mitigating the ongoing COVID-19 pandemic and preparing for future emerging infectious diseases.

Methodology

The steps we followed in the analysis were:

1) Data Collection: The first step is collecting relevant data from the NCBI list. We were interested in two main areas of research: the first to trace out how the virus travelled, the second to compare sequences from various times and places to the known sequences for the Omicron and Delta variant.

2) Alignment Technique: Once we sourced the data, the next step was to align each sequence with all the other sequences to determine similarity. From this, it is possible to hypothesize the path COVID-19 travelled and the mutations that occurred along the way.

3) Clustering: Clustering can be done once the alignment of the sequences is completed since one of the inputs to the clustering algorithm is the similarity score. Clustering enables one to determine various groups in the selections that are sufficiently close to each other to yield meaningful information.

4) Tabulation: The results and process were tabulated.

Multiple Sequence Alignment (MSA)
It's one of the most essential tools in molecular biology. It's useful for finding highly conserved sub-regions or embedded patterns of a set of biological resources. It performs successive pair-wise alignments, builds consensus sequences and further aligns them. It uses N-D matrix instead of ones used in class. Computational run-time $O(2^k n^k)$. Clustalw was one of the initial algorithms.
Key Evaluation Parameters include:

- Scoring Matrix
- Gap Open Penalty
- Gap Extension Penalty

The alignment algorithm we used was MAFFT (Multiple Alignment using Fast Fourier Transform), a multiple alignment program that significantly reduces computer operating time by using the fast Fourier transform technique to approximate a near solution to the alignment problem at hand. The resource we used for MAFFT is available via the European Bioinformatics Institute website (https://www.ebi.ac.uk/jdispatcher/msa/mafft). We had originally planned to use the Needle algorithm available on Galaxy, but the viral genome proved to be too long for the server to handle, resulting in a timeout error. In contrast, MAFFT aligned each of the 20 sequences with each other in less than a minute, a huge improvement over Needle.

The similarity measure that we used was the similarity %. We did not use the score because the sequence length varied across records due to various factors such as mutations, human errors, unread nucleotides and such. We also did not use identity % because similarity % is a more holistic approach that considers gaps.

Analysis 1: Similarity of Random Sequences with Delta and Omicron Variants

In this section, we collect 20 random samples from the NCBI database and attempt to compare them to the Delta and Omicron variants. Both variants were prominent in different countries at different times.

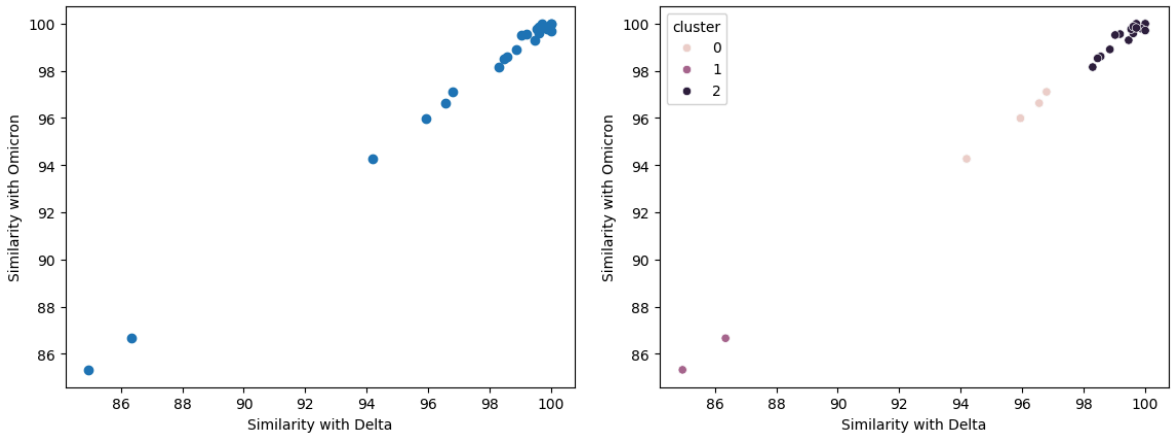
We chose samples taken from Bangladesh in May 2021 as a reference sequence for the Delta variant since it was the dominant variant in the Indian subcontinent at the time. Similarly, samples collected in Bangladesh in Jan 2022 are chosen as the reference for the Omicron variant.

For instance, Omicron was the dominant variant in the USA by Christmas 2021, so samples collected then should align better with the reference Omicron than samples from the USA in, say, 2020. To this end, as mentioned above, we perform a pairwise alignment using the MAFFT algorithm and calculate the similarity scores.

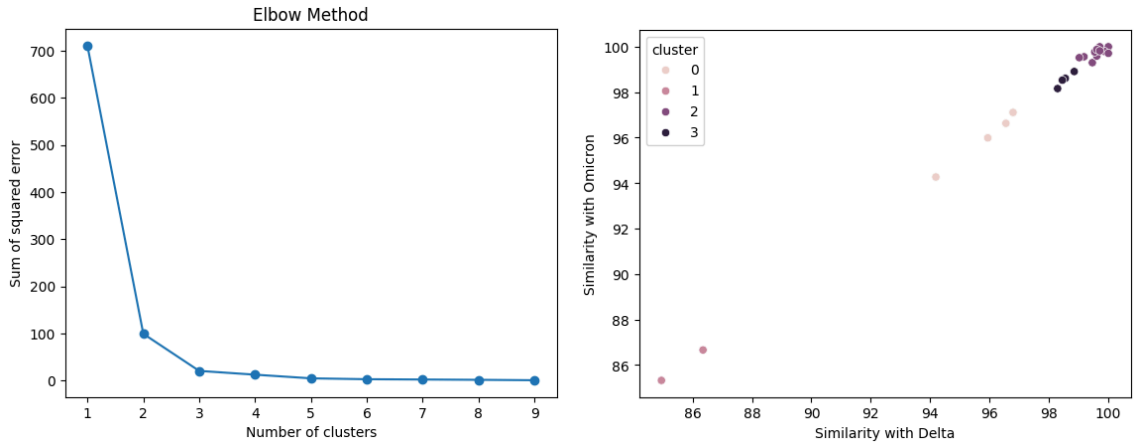
ID	Geography	Collection Date	Similarity with Delta	Similarity with Omicron
MT843234.1	Italy	2019-12-18	100	100
MZ223389.1	Italy	2019-10-17	100	100
MT019532.1	Wuhan	2019-12-30	99.87	99.79
OY390535.1	USA	NA	95.94	95.99
OX683594.1	UK: England	NA	98.85	98.91
HG999976.1	UK: Oxford	NA	99.59	99.67
OY887427.1	UK: England	2020-10-01	94.19	94.27
OX656416.1	UK: England	2020-12-25	98.55	98.61
MZ875753.1	USA: Colorado	2021-08-03	99.46	99.3
OU349741.1	UK: England	2021-06-17	99.61	99.59
MZ437368.1	Bangladesh/Delta	2021-05-18	100	99.71
OA979203.1	UK: England	2020-06-29	98.45	98.53
PP275463.1	USA: North Carolina	2024-02-06	99.54	99.77
OZ023853.1	UK: England	2024-03-06	99.18	99.56
OZ020766.1	UK: Scotland	2024-02-09	86.35	86.66
OY286551.1	UK: England	2023-04-30	99.02	99.52
ON026026.1	Bangladesh/Omicron	2022-01-13	99.71	100

OQ705450.1	USA: New York	2023-03-12	99.61	99.88
PP419738.1	USA: New York	2023-08-27	96.79	97.11
OL985535.1	USA: Vermont	2021-12-11	98.29	98.16
OM498708.1	USA: Tennessee	2022-01-18	96.55	96.63
OM409105.1	USA: Virginia	2022-01-09	99.71	99.82
OY530643.1	UK: Wales	2022-02-14	84.95	85.32

The scatter plot of the two scores was then as follows:



Based on visual inspection, we cluster the above data into three groups, as shown in the plot above on the right. We plot the SSE against the errors and use the elbow method to obtain the optimal number of clusters. We thus choose 4 clusters.



The following sequences are present in each of the clusters:

Cluster 0: This is the intermediate cluster, which lies between roughly 96-98% similarity for both the variants. Most of these results seem physically reasonable, as the Delta variant became dominant in the UK and USA around May 2021. The Omicron variant started spreading in later

2021, and thus again the samples collected are not expected to be very similar to that. By our hypothesis, however, the Tennessee sample should've been more similar to Omicron than it is.

ID	Geography	Collection Date	Similarity with Delta	Similarity with Omicron
OY390535.1	USA	NA	95.94	95.99
OY887427.1	UK: England	2020-10-01	94.19	94.27
PP419738.1	USA: New York	2023-08-27	96.79	97.11
OM498708.1	USA: Tennessee	2022-01-18	96.55	96.63

Cluster 1:

These are the most uncorrelated with the others. The sample collected in 2024 seems reasonable as the virus may have mutated again in the two years since Omiron dominated in 2022. Again, however, we expect the Wales samples to be more similar to Omiron than what is shown here.

ID	Geography	Collection Date	Similarity with Delta	Similarity with Omicron
OZ020766.1	UK: Scotland	2024-02-09	86.35	86.66
OY530643.1	UK: Wales	2022-02-14	84.95	85.32

Cluster 2:

These are the most highly correlated with both Delta and Omicron variants. Since most of the collection dates range between 2019 and 2021 the results seem physically reasonable. Even in the US and UK as of 2024, the dominant variants are subvariants of Omicron[5] and thus the high similarity score with Omicron is expected.

ID	Geography	Collection Date	Similarity with Delta	Similarity with Omicron
MT843234.1	Italy	2019-12-18	100	100
MZ223389.1	Italy	2019-10-17	100	100
MT019532.1	Wuhan	2019-12-30	99.87	99.79
HG999976.1	UK: Oxford	NA	99.59	99.67
MZ875753.1	USA: Colorado	2021-08-03	99.46	99.3
OU349741.1	UK: England	2021-06-17	99.61	99.59
MZ437368.1	Bangladesh/Delta	2021-05-18	100	99.71

PP275463.1	USA: North Carolina	2024-02-06	99.54	99.77
OZ023853.1	UK: England	2024-03-06	99.18	99.56
OY286551.1	UK: England	2023-04-30	99.02	99.52
ON026026.1	Bangladesh/Omicron	2022-01-13	99.71	100
OQ705450.1	USA: New York	2023-03-12	99.61	99.88
OM409105.1	USA: Virginia	2022-01-09	99.71	99.82

Cluster 3: All samples in these clusters represent high similarity with Delta and Omicron, though not as high as the previous cluster. By the collection dates, the results seem to be physically reasonable.

ID	Geography	Collection Date	Similarity with Delta	Similarity with Omicron
OX 683594.1	UK: England	NA	98.85	98.91
OX656416.1	UK: England	2020-12-25	98.55	98.61
OA979203.1	UK: England	2020-06-29	98.45	98.53
OL985535.1	USA: Vermont	2021-12-11	98.29	98.16

Analysis 2: Validation of Spread of Virus across Geographies

In this section, we attempt a combined spatial-temporal analysis of similarity scores to validate the common perception of the spread of the virus, ie, from China to US and Europe to India. This is based on publicly available information and news sources, along with the date of detection of the first case in that country. Since the virus is expected to mutate with time, we assume that if the virus spread from country A to B to C; the sequence in B will be more similar to A than C is to A. C, however, will be expected to be similar to B.

The following table gives the alignment results:

	NC_045512.2	**PP380203.1**	**PP380279.1**	**OR655373.1**	**OR357639.1**	**OY740686.1**	**OX327063.1**	**PP222830.1**	**PP222833.1**	
NC_045512.2	100	99.9	99.79	98.18	99.77	94.14	99.55	99.77	99.77	**2019 China (RefSeq)**
PP380203.1	99.9	100	99.72	98.1	99.7	94.5	99.69	99.7	99.7	**2021 China**
PP380279.1	99.79	99.72	100	98.03	99.83	94.5	99.81	99.95	99.96	**2022 China**
OR655373.1	98.18	98.1	98.03	100	98.01	94.36	98.01	99.81	98.01	**2021 India**
OR357639.1	99.77	99.7	99.83	98.01	100	94.44	99.91	99.81	99.81	**2022 India**
OY740686.1	94.14	94.5	94.5	94.36	94.44	100	94.25	94.21	94.22	**2021 Switzerland**
OX327063.1	99.55	99.69	99.81	98.01	99.91	94.25	100	99.7	99.71	**2022 Switzerland**
PP222830.1	99.77	99.7	99.95	99.81	99.81	94.21	99.7	100	99.98	**2021 USA**
PP222833.1	99.77	99.7	99.96	98.01	99.81	94.22	99.71	99.98	100	**2022 USA**

As expected, China 2022 is not as similar to China 2019 as China 2021 is. While unexpected, there is a consistent trend that the 2022 samples from India, Switzerland, and the US are more similar to 2019 China than the 2021 samples from the same countries.

Other observations are as follows:

1. The samples in India in the two years are not that close, probably due to the difference between Delta and Omicron strains, which were dominant at that time.
2. The USA in 2021 and 2022 are very similar, but Switzerland is quite different.
3. China in 2021 shows almost an equally good match with both India and China in 2022.
4. India and Switzerland in 2022 are quite similar, probably because the Omiron variant had become dominant in both countries by then

Analysis 3: Constructing a Phylogenetic Tree

For this analysis, we use the dissimilarity matrix method.

Let n be a positive integer.
A distance matrix of order n (also called a dissimilarity matrix of order n) is a matrix D of size (n,n) which satisfies:

1. $D_{i,j} > 0$ for all i,j = {1,2…n} with i!=j
2. $D_{i,j} = 0$ for all i,j = {1,2…n} with i=j
3. $D_{i,j} = D_{J,I}$ for all i,j = {1,2…n}

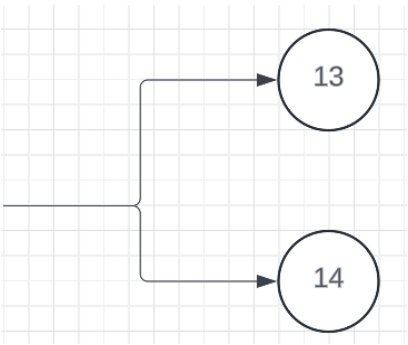
We used 21 sequences out of the 23 used for analysis 1 because the first two sequences' dissimilarity scores did not agree with the conditions imposed on the validity of the dissimilarity matrix.

We'll now begin a step-by-step approach towards building our phylogenetic tree.

Initial Dissimilarity Matrix:

		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
		MT019532.1	OY390535.1	OX683594.1	HG999976.1	OY887427.1	OX656416.1	MZ875753.1	OU349741.1	MZ437368.1	OA979203.1	PP275463.1	OZ023853.1	OZ020766.1	OY286551.1	ON026026.1	OQ705450.1	PP419738.1	OL985535.1	OM498708.1	OM409105.1	OY530643.1
3	MT019532.1	0																				
4	OY390535.1	4	0																			
5	OX683594.1	1.12	2.94	0																		
6	HG999976.1	0.39	3.67	0.76	0																	
7	OY887427.1	5.79	5.64	4.68	5.43	0																
8	OX656416.1	1.42	4.7	1.79	1.06	4.46	0															
9	MZ875753.1	0.55	4.27	1.35	0.61	5.85	1.65	0														
10	OU349741.1	0.48	3.76	0.85	0.11	5.5	1.14	0.51	0													
11	MZ437368.1	0.13	4.06	1.15	0.41	5.81	1.45	0.54	0.39	0												
12	OA979203.1	1.53	4.81	1.9	1.14	6.56	2.19	1.75	1.25	1.55	0											
13	PP275463.1	0.41	4.07	1.14	0.39	5.79	1.44	0.87	0.44	0.46	1.53	0										
14	OZ023853.1	0.84	4.15	1.23	0.46	5.86	1.53	1	0.53	0.82	1.62	0.06	0									
15	OZ020766.1	13.67	15.02	14.01	13.3	18.63	14.32	13.85	13.32	13.65	14.44	13.02	13.01	0								
16	OY286551.1	0.98	4.25	1.35	0.81	5.97	1.86	1.21	0.87	0.98	1.95	0.57	0.74	13.56	0							
17	ON026026.1	0.21	4.01	1.09	0.33	5.73	1.39	0.7	0.41	0.29	1.47	0.23	0.44	13.34	0.48	0						
18	OQ705450.1	0.33	4	1.08	0.32	5.73	1.38	0.8	0.38	0.39	1.46	0.27	0.37	13.21	0.36	0.12	0					
19	PP419738.1	3.16	6.9	3.98	3.23	6.64	4.28	3.63	3.29	3.21	4.37	3.03	3.19	16.06	3.13	2.89	2.81	0				
20	OL985535.1	1.72	4.38	2.66	1.93	6.12	2.95	2	1.83	1.71	3.06	1.99	2.33	15.15	2.48	1.84	1.92	4.74	0			
21	OM498708.1	3.39	7.07	4.12	3.37	8.83	4.43	3.86	3.42	3.45	4.51	3.54	3.54	14.57	3.84	3.37	3.46	4.77	4.98	0		
22	OM409105.1	0.21	3.87	0.94	0.19	5.63	1.24	0.7	0.26	0.29	1.33	0.35	0.35	13.12	0.66	0.18	0.26	3.11	1.82	3.21	0	
23	OY530643.1	14.98	16.13	14.83	14.77	16.48	15.82	15.09	14.83	15.05	15.82	14.65	14.71	17.97	14.52	14.68	14.63	17.46	14.89	17.47	14.53	0

The least coefficient is for the 13-14 combination.

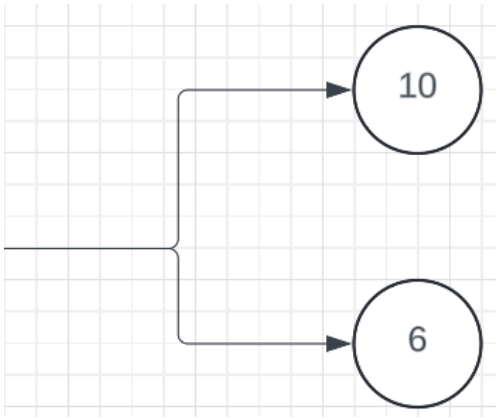


To get the updated dissimilarity matrix, we assume single linkage as the distance between the newly formed cluster and other variants; hence our governing updation will be the min $d_{i,j}$ of 13 v/s 14 with all the other sequences.

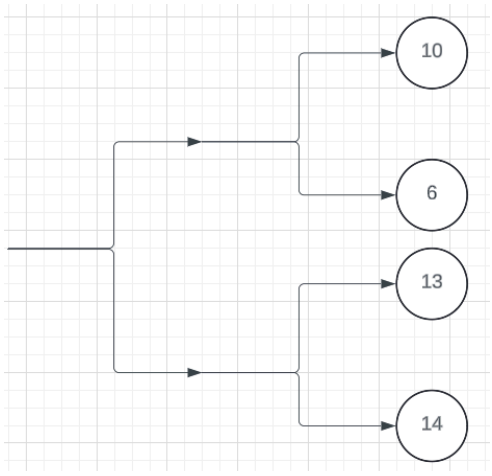
First updated dissimilarity matrix:

		3	4	5	6	7	8	9	10	11	12	13-14	15	16	17	18	19	20	21	22	23
		MT019532.1	OY390535.1	OX683594.1	HG999976.1	OY887427.1	OX656416.1	MZ875753.1	OU349741.1	MZ437368.1	OA979203.1	PP275463.1	OZ020766.1	OY286551.1	ON026026.1	OQ705450.1	PP419738.1	OL985535.1	OM498708.1	OM409105.1	OY530643.1
3	MT019532.1	0																			
4	OY390535.1	4	0																		
5	OX683594.1	1.12	2.94	0																	
6	HG999976.1	0.39	3.67	0.76	0																
7	OY887427.1	5.79	5.64	4.68	5.43	0															
8	OX656416.1	1.42	4.7	1.79	1.06	4.46	0														
9	MZ875753.1	0.55	4.27	1.35	0.61	5.85	1.65	0													
10	OU349741.1	0.48	3.76	0.85	0.11	5.5	1.14	0.51	0												
11	MZ437368.1	0.13	4.06	1.15	0.41	5.81	1.45	0.54	0.39	0											
12	OA979203.1	1.53	4.81	1.9	1.14	6.56	2.19	1.75	1.25	1.55	0										
13-14	PP275463.1	0.41	4.07	1.14	0.39	5.79	1.44	0.87	0.44	0.46	1.53	0									
15	OZ020766.1	13.67	15.02	14.01	13.3	18.63	14.32	13.85	13.32	13.65	14.44	13.01	0								
16	OY286551.1	0.98	4.25	1.35	0.81	5.97	1.86	1.21	0.87	0.98	1.95	0.57	13.56	0							
17	ON026026.1	0.21	4.01	1.09	0.33	5.73	1.39	0.7	0.41	0.29	1.47	0.23	13.34	0.48	0						
18	OQ705450.1	0.33	4	1.08	0.32	5.73	1.38	0.8	0.38	0.39	1.46	0.27	13.21	0.36	0.12	0					
19	PP419738.1	3.16	6.9	3.98	3.23	6.64	4.28	3.63	3.29	3.21	4.37	3.03	16.06	3.13	2.89	2.81	0				
20	OL985535.1	1.72	4.38	2.66	1.93	6.12	2.95	2	1.83	1.71	3.06	1.99	15.15	2.48	1.84	1.92	4.74	0			
21	OM498708.1	3.39	7.07	4.12	3.37	8.83	4.43	3.86	3.42	3.45	4.51	3.54	14.57	3.84	3.37	3.46	4.77	4.98	0		
22	OM409105.1	0.21	3.87	0.94	0.19	5.63	1.24	0.7	0.26	0.29	1.33	0.35	13.12	0.66	0.18	0.26	3.11	1.82	3.21	0	
23	OY530643.1	14.98	16.13	14.83	14.77	16.48	15.82	15.09	14.83	15.05	15.82	14.65	17.97	14.52	14.68	14.63	17.46	14.89	17.47	14.53	0

The least coefficient is for the 10-6 combination.



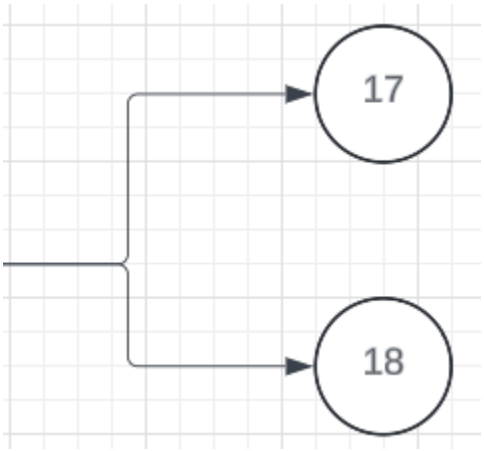
So our combined flowchart looks like:



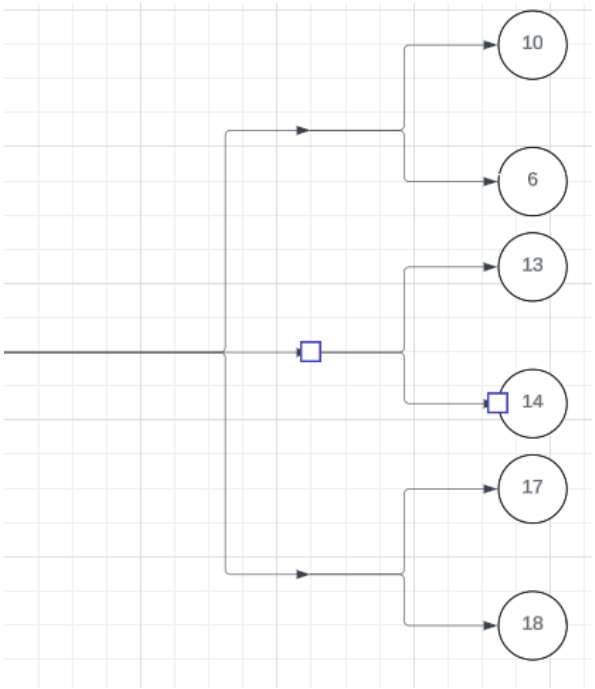
Second updated dissimilarity matrix

		3	4	5	10-6	7	8	9	11	12	13-14	15	16	17	18	19	20	21	22	23
		MT019532.1	OY390535.1	OX683594.1	HG999976.1	OY887427.1	OX656416.1	MZ875753.1	MZ437368.1	OA979203.1	PP275463.1	OZ020766.1	OY286551.1	ON026026.1	OQ705450.1	PP419738.1	OL985535.1	OM498708.1	OM409105.1	OY530643.1
3	MT019532.1	0																		
4	OY390535.1	4	0																	
5	OX683594.1	1.12	2.94	0																
10-6	HG999976.1	0.39	3.67	0.76	0															
7	OY887427.1	5.79	5.64	4.68	5.43	0														
8	OX656416.1	1.42	4.7	1.79	1.06	4.46	0													
9	MZ875753.1	0.55	4.27	1.35	0.61	5.85	1.65	0												
11	MZ437368.1	0.13	4.06	1.15	0.41	5.81	1.45	0.54	0											
12	OA979203.1	1.53	4.81	1.9	1.14	6.56	2.19	1.75	1.55	0										
13-14	PP275463.1	0.41	4.07	1.14	0.39	5.79	1.44	0.87	0.46	1.53	0									
15	OZ020766.1	13.67	15.02	14.01	13.3	18.63	14.32	13.85	13.65	14.44	13.01	0								
16	OY286551.1	0.98	4.25	1.35	0.81	5.97	1.86	1.21	0.98	1.95	0.57	13.56	0							
17	ON026026.1	0.21	4.01	1.09	0.33	5.73	1.39	0.7	0.29	1.47	0.23	13.34	0.48	0						
18	OQ705450.1	0.33	4	1.08	0.32	5.73	1.38	0.8	0.39	1.46	0.27	13.21	0.36	0.12	0					
19	PP419738.1	3.16	6.9	3.98	3.23	6.64	4.28	3.63	3.21	4.37	3.03	16.06	3.13	2.89	2.81	0				
20	OL985535.1	1.72	4.38	2.66	1.93	6.12	2.95	2	1.71	3.06	1.99	15.15	2.48	1.84	1.92	4.74	0			
21	OM498708.1	3.39	7.07	4.12	3.37	8.83	4.43	3.86	3.45	4.51	3.54	14.57	3.84	3.37	3.46	4.77	4.98	0		
22	OM409105.1	0.21	3.87	0.94	0.19	5.63	1.24	0.7	0.29	1.33	0.35	13.12	0.66	0.18	0.26	3.11	1.82	3.21	0	
23	OY530643.1	14.98	16.13	14.83	14.77	16.48	15.82	15.09	15.05	15.82	14.65	17.97	14.52	14.68	14.63	17.46	14.89	17.47	14.53	0

The least coefficient is for the 17-18 combination.



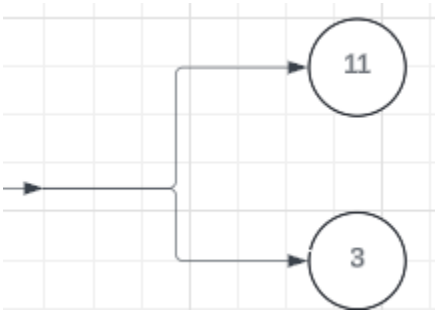
So our combined flowchart looks like:



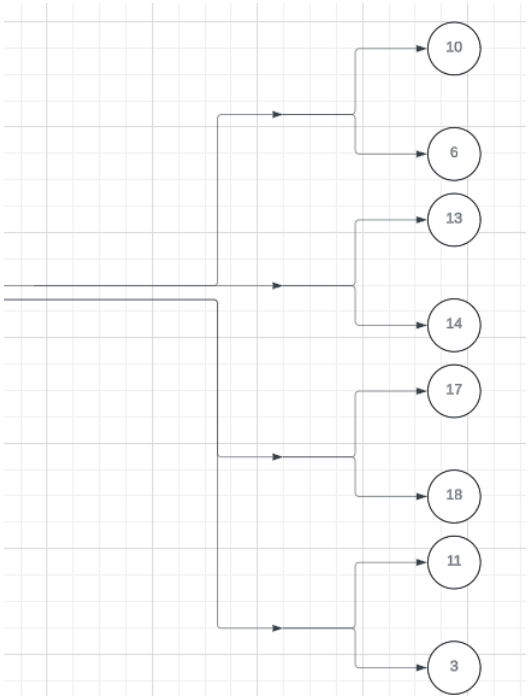
The third updated dissimilarity matrix:

		3	4	5	10-6	7	8	9	11	12	13-14	15	16	17-18	19	20	21	22	23
		MT019532.1	OY390535.1	OX683594.1	HG999976.1	OY887427.1	OX656416.1	MZ875753.1	MZ437368.1	OA979203.1	PP275463.1	OZ020766.1	OY286551.1	ON026026.1	PP419738.1	OL985535.1	OM498708.1	OM409105.1	OY530643.1
3	MT019532.1	0																	
4	OY390535.1	4	0																
5	OX683594.1	1.12	2.94	0															
10-6	HG999976.1	0.39	3.67	0.76	0														
7	OY887427.1	5.79	5.64	4.68	5.43	0													
8	OX656416.1	1.42	4.7	1.79	1.06	4.46	0												
9	MZ875753.1	0.55	4.27	1.35	0.61	5.85	1.65	0											
11	MZ437368.1	0.13	4.06	1.15	0.41	5.81	1.45	0.54	0										
12	OA979203.1	1.53	4.81	1.9	1.14	6.56	2.19	1.75	1.55	0									
13-14	PP275463.1	0.41	4.07	1.14	0.39	5.79	1.44	0.87	0.46	1.53	0								
15	OZ020766.1	13.67	15.02	14.01	13.3	18.63	14.32	13.85	13.65	14.44	13.01	0							
16	OY286551.1	0.98	4.25	1.35	0.81	5.97	1.86	1.21	0.98	1.95	0.57	13.56	0						
17-18	ON026026.1	0.21	4	1.08	0.32	5.73	1.38	0.7	0.29	1.46	0.23	13.21	0.36	0					
19	PP419738.1	3.16	6.9	3.96	3.23	6.64	4.28	3.63	3.21	4.37	3.03	16.06	3.13	2.81	0				
20	OL985535.1	1.72	4.38	2.66	1.93	6.12	2.95	2	1.71	3.06	1.99	15.15	2.48	1.84	4.74	0			
21	OM498708.1	3.39	7.07	4.12	3.37	8.83	4.43	3.86	3.45	4.51	3.54	14.57	3.84	3.37	4.77	4.98	0		
22	OM409105.1	0.21	3.87	0.94	0.19	5.63	1.24	0.7	0.29	1.33	0.35	13.12	0.66	0.18	3.11	1.82	3.21	0	
23	OY530643.1	14.98	16.13	14.83	14.77	16.48	15.82	15.09	15.05	15.82	14.65	17.97	14.52	14.68	17.46	14.89	17.47	14.53	0

The least coefficient is for the 11-3 combination.



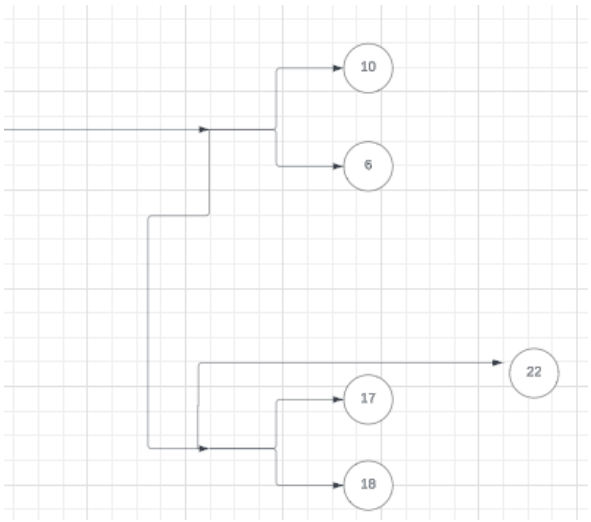
So our combined flowchart looks like:



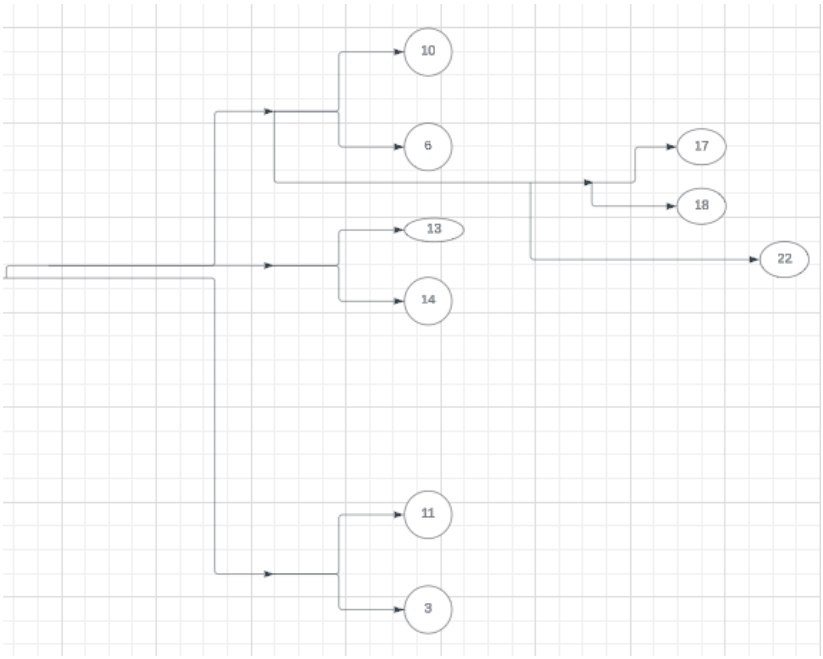
The fourth updated dissimilarity matrix:

		11-3	4	5	10-6	7	8	9	12	13-14	15	16	17-18	19	20	21	22	23
		MT019532.1	OY390535.1	OX683594.1	HG999976.1	OY887427.1	OX656416.1	MZ875753.1	OA979203.1	PP275463.1	OZ020766.1	OY286551.1	ON026026.1	PP419738.1	OL985535.1	OM498708.1	OM409105.1	OY530643.1
11-3	MT019532.1	0																
4	OY390535.1	4	0															
5	OX683594.1	1.12	2.94	0														
10-6	HG999976.1	0.39	3.67	0.76	0													
7	OY887427.1	5.79	5.64	4.68	5.43	0												
8	OX656416.1	1.42	4.7	1.79	1.06	4.46	0											
9	MZ875753.1	0.55	4.27	1.35	0.61	5.85	1.65	0										
12	OA979203.1	1.53	4.81	1.9	1.14	6.56	2.19	1.75	0									
13-14	PP275463.1	0.41	4.07	1.14	0.39	5.79	1.44	0.87	1.53	0								
15	OZ020766.1	13.65	15.02	14.01	13.3	18.63	14.32	13.85	14.44	13.01	0							
16	OY286551.1	0.98	4.25	1.35	0.81	5.97	1.86	1.21	1.95	0.57	13.56	0						
17-18	ON026026.1	0.21	4	1.08	0.32	5.73	1.38	0.7	1.46	0.23	13.21	0.36	0					
19	PP419738.1	3.16	6.9	3.98	3.23	6.64	4.28	3.63	4.37	3.03	16.06	3.13	2.81	0				
20	OL985535.1	1.71	4.38	2.66	1.93	6.12	2.95	2	3.06	1.99	15.15	2.48	1.84	4.74	0			
21	OM498708.1	3.39	7.07	4.12	3.37	8.83	4.43	3.86	4.51	3.54	14.57	3.84	3.37	4.77	4.98	0		
22	OM409105.1	0.21	3.87	0.94	0.19	5.63	1.24	0.7	1.33	0.35	13.12	0.66	0.18	3.11	1.82	3.21	0	
23	OY530643.1	14.98	16.13	14.83	14.77	16.48	15.82	15.09	15.82	14.65	17.97	14.52	14.68	17.46	14.89	17.47	14.53	0

The least coefficient is for the 10-6 and 17-18-22 combination.



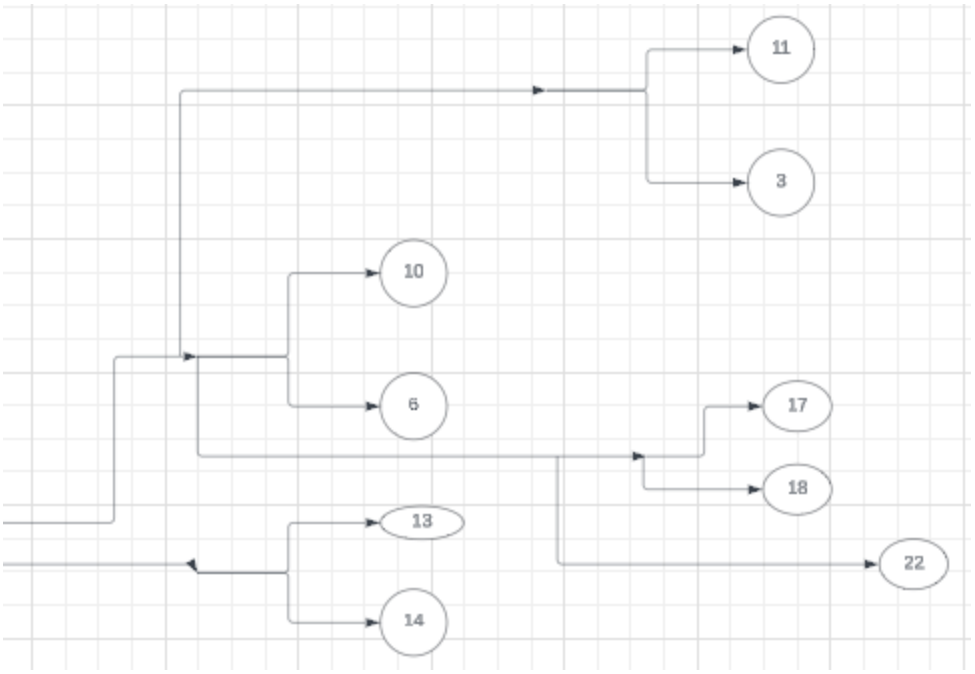
So our combined flowchart looks like:



Fifth updated dissimilarity matrix:

		11-3	4	5	7	8	9	12	13-14	15	16	19	20	21	10-6-17-18-22	23
		MT019532.1	OY390535.1	OX683594.1	OY887427.1	OX656416.1	MZ875753.1	OA979203.1	PP275463.1	OZ020766.1	OY286551.1	PP419738.1	OL985535.1	OM498708.1	OM409105.1	OY530643.1
11-3	MT019532.1	0														
4	OY390535.1	4	0													
5	OX683594.1	1.12	2.94	0												
10-6-17-18-22	HG999976.1	0.21	3.67	0.76												
7	OY887427.1	5.79	5.64	4.68	0											
8	OX656416.1	1.42	4.7	1.79	4.46	0										
9	MZ875753.1	0.55	4.27	1.35	5.85	1.65	0									
12	OA979203.1	1.53	4.81	1.9	6.56	2.19	1.75	0								
13-14	PP275463.1	0.41	4.07	1.14	5.79	1.44	0.87	1.53	0							
15	OZ020766.1	13.65	15.02	14.01	18.63	14.32	13.85	14.44	13.01	0						
16	OY286551.1	0.98	4.25	1.35	5.97	1.86	1.21	1.95	0.57	13.56	0					
19	PP419738.1	3.16	6.9	3.98	6.64	4.28	3.63	4.37	3.03	16.06	3.13	0				
20	OL985535.1	1.71	4.38	2.66	6.12	2.95	2	3.06	1.99	15.15	2.48	4.74	0			
21	OM498708.1	3.39	7.07	4.12	8.83	4.43	3.86	4.51	3.54	14.57	3.84	4.77	4.98	0		
23	OY530643.1	14.98	16.13	14.83	16.48	15.82	15.09	15.82	14.65	17.97	14.52	17.46	14.89	17.47	14.53	0

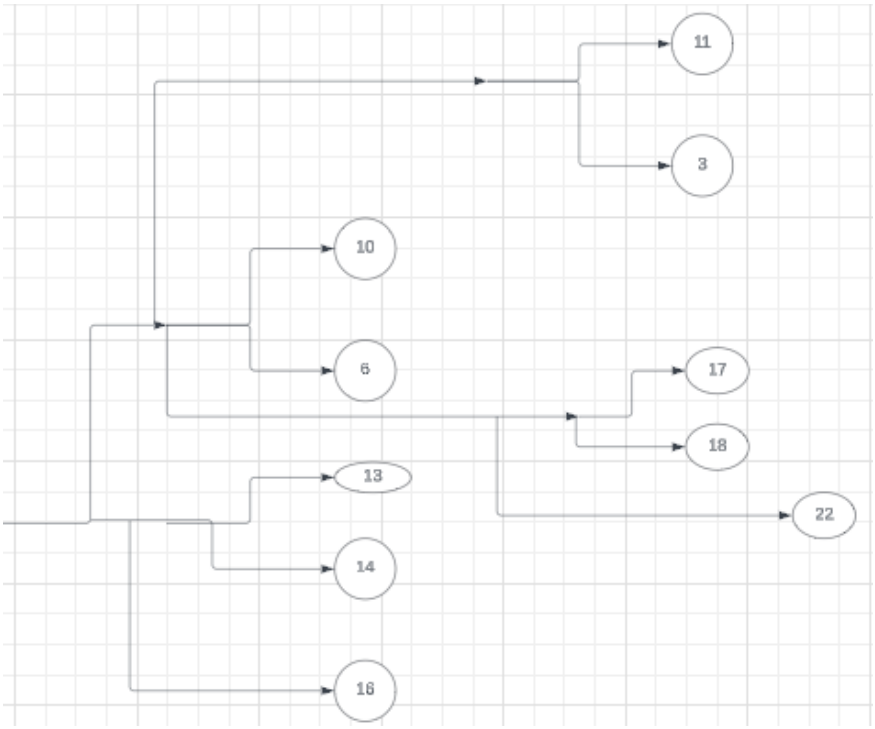
The least coefficient is for the 10-6-17-18-22 and 11-3 combination.
So our combined flowchart looks like:



The sixth updated dissimilarity matrix:

		4	5	7	8	9	12	13-14	15	16	19	20	21	10-6-17-18-22-11-3	23
		OY390535.1	OX683594.1	OY887427.1	OX656416.1	MZ875753.1	OA979203.1	PP275463.1	OZ020766.1	OY286551.1	PP419738.1	OL985535.1	OM498708.1	OM409105.1	OY530643.1
10-6-17-18-22-11-3	MT019532.1														
4	OY390535.1	0													
5	OX683594.1	2.94	0												
7	OY887427.1	5.64	4.68	0											
8	OX656416.1	4.7	1.79	4.46	0										
9	MZ875753.1	4.27	1.35	5.85	1.65	0									
12	OA979203.1	4.81	1.9	6.56	2.19	1.75	0								
13-14	PP275463.1	4.07	1.14	5.79	1.44	0.87	1.53	0							
15	OZ020766.1	15.02	14.01	18.63	14.32	13.85	14.44	13.01	0						
16	OY286551.1	4.25	1.35	5.97	1.86	1.21	1.95	0.57	13.56	0					
19	PP419738.1	6.9	3.98	6.64	4.28	3.63	4.37	3.03	16.06	3.13	0				
20	OL985535.1	4.38	2.66	6.12	2.95	2	3.06	1.99	15.15	2.48	4.74	0			
21	OM498708.1	7.07	4.12	8.83	4.43	3.86	4.51	3.54	14.57	3.84	4.77	4.98	0		
23	OY530643.1	16.13	14.83	16.48	15.82	15.09	15.82	14.65	17.97	14.52	17.46	14.89	17.47	14.53	0

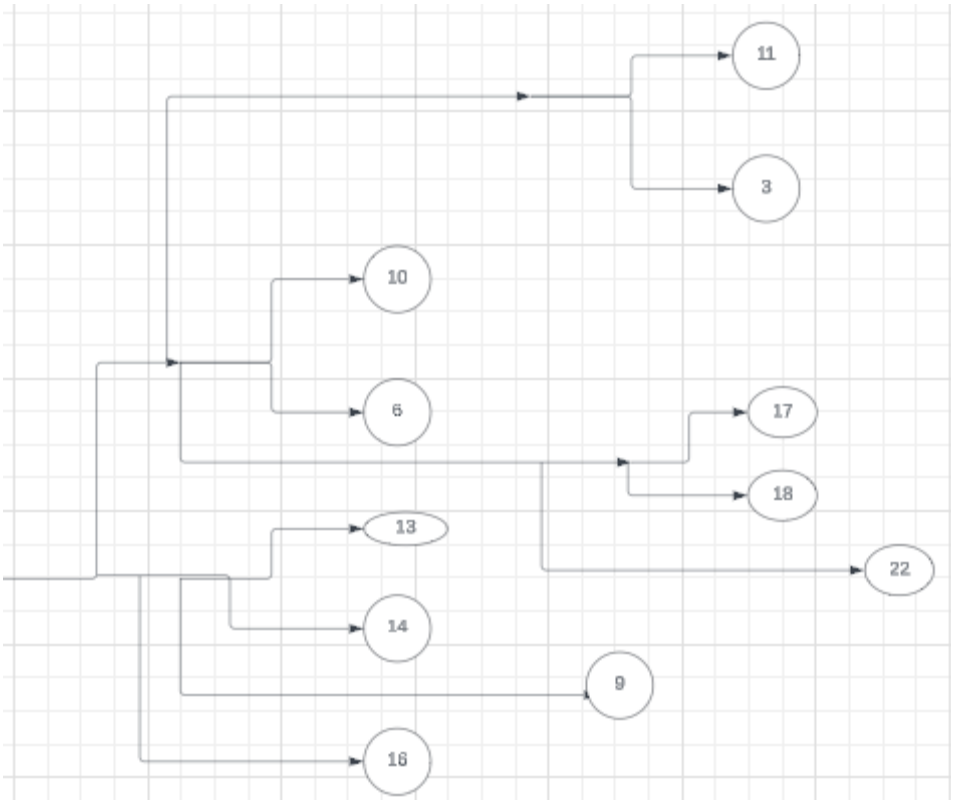
The least coefficient is for the 16 and 13-14 combination.
So our combined flowchart looks like:



The seventh updated dissimilarity matrix:

		4	5	7	8	9	12	15	16	19	20	21	)-6-17-18-22-11	23
		OY390535.1	OX683594.1	OY887427.1	OX656416.1	MZ875753.1	OA979203.1	OZ020766.1	OY286551.1	PP419738.1	OL985535.1	OM498708.1	OM409105.1	OY530643.1
10-6-17-18-22-11-3	MT019532.1													
4	OY390535.1	0												
5	OX683594.1	2.94	0											
7	OY887427.1	5.64	4.68	0										
8	OX656416.1	4.7	1.79	4.46	0									
9	MZ875753.1	4.27	1.35	5.85	1.65	0								
12	OA979203.1	4.81	1.9	6.56	2.19	1.75	0							
13-14	PP275463.1	4.07	1.14	5.79	1.44	0.87	1.53							
15	OZ020766.1	15.02	14.01	18.63	14.32	13.85	14.44	0						
19	PP419738.1	6.9	3.98	6.64	4.28	3.63	4.37	16.06	3.13	0				
20	OL985535.1	4.38	2.66	6.12	2.95	2	3.06	15.15	2.48	4.74	0			
21	OM498708.1	7.07	4.12	8.83	4.43	3.86	4.51	14.57	3.84	4.77	4.98	0		
23	OY530643.1	16.13	14.83	16.48	15.82	15.09	15.82	17.97	14.52	17.46	14.89	17.47	14.53	0

The least coefficient is for the 9 and 13-14 combination.
So our combined flowchart looks like:


11
3
10
6
17
18
13
22
14
9
16

Conclusions

Through the analysis undertaken above, we discovered that

1) The genetic records match the timeline of COVID-19 infection and mutation as reported in the news.

2) The fast rate of mutation of viruses means that there is inevitably some messy data to work with, and a few unexpected results popped out, such as some areas of the USA being closer to Omicron during the Delta wave than they should be.

3) The Wuhan sample collected in 2019 is closer in similarity to the older waves than it is to the newer waves, giving us the ability to measure the ‘rate’ of mutation in the COVID virus.

4) Our phylogenetic tree further validated this, with samples collected at longer time durations being farther in the genetic tree. A similar pattern was observed in regards to geography.

Bibliography

1. Jahan M, Nasif MAO, Rahmat R, Islam SRU, Munshi SU, Ahmed MS. Genome Sequencing of Omicron Variants of SARS-CoV-2 Circulating in Bangladesh during the Third Wave of the COVID-19 Pandemic. Microbiol Resour Announc. 2022 Jul 21;11(7):e0038122. doi: 10.1128/mra.00381-22. Epub 2022 May 31. PMID: 35638826; PMCID: PMC9302173.
2. Banu TA, Sarkar MMH, Akter S, Goswami B, Jahan I, Osman E, Uzzaman MS, Habib MA, Mahmud ASM, Uddin MM, Nafisa T, Molla MMA, Yeasmin M, Akram A, Khan MS. Genome Sequencing of the SARS-CoV-2 Delta (B.1.617.2) Variant of Concern Detected in Bangladesh. Microbiol Resour Announc. 2021 Dec 2;10(48):e0084921. doi: 10.1128/MRA.00849-21. Epub 2021 Dec 2. PMID: 34854726; PMCID: PMC8638595.
3. Jansson, J. (2008). Phylogenetic Tree Construction from a Distance Matrix. In Springer eBooks (pp. 651–653). https://doi.org/10.1007/978-0-387-30162-4_292
4. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013 Apr;30(4):772-80. doi: 10.1093/molbev/mst010. Epub 2013 Jan 16. PMID: 23329690; PMCID: PMC3603318.
5. What COVID-19 variants are going around in April 2024? (2024, April 26). Nebraska Medicine Omaha, NE. https://www.nebraskamed.com/COVID/what-covid-19-variants-are-going-around
6. Madeira F, Pearce M, Tivey ARN, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Research. 2022 Jul;50(W1):W276-W279. DOI: 10.1093/nar/gkac240. PMID: 35412617; PMCID: PMC9252731.

Appendix

The code used for Analysis 1 can be accessed here (with LDAP):
https://colab.research.google.com/drive/1qo9m7eTqNlq0pSA7ezrDo8S4habYikC9?usp=sharing

The alignment results for the various analyses can be found here:
https://docs.google.com/spreadsheets/d/1Con0XDPzsKZZ6YUysPV-4Eb0AoU4zhmSJ0YxfzkUirw/edit
https://docs.google.com/spreadsheets/d/1Im0bDnyfd8lTLS6NtTo5l5B4Jqam4Zdp9swQX-iVoY0/edit

The datasets used can be found here:

First analysis:
https://drive.google.com/file/d/16x4yQEIKVku8zx74fc7gI-H-SV6evV2g/view?usp=drive_link
https://drive.google.com/file/d/19TNKHnKpG2ypQvcdxT1YojyHiYT5UzWy/view?usp=drive_link

Second analysis:
https://drive.google.com/file/d/1FOFCagxfBHWx_DxXq2D2MOVme7okqRpx/view?usp=drive_link