Comparative Genomic Analysis of SARS-CoV-2 Delta and Omicron Variants in Ghana

Farah Ossama Ahmed Sabra
Systems & Biomedical Engineering. Faculty of Engineering
Cairo University
Giza, Egypt
farah.sabra03@eng-st.cu.edu.eg

Shehab Mohamed Ibrahim Mohamed
Systems & Biomedical Engineering. Faculty of Engineering
Cairo University
Giza, Egypt
shehabhegab20@gmail.com

Camellia Marwan Hussein
Systems & Biomedical Engineering. Faculty of Engineering
Cairo University
Giza, Egypt
Camelliamarwan@gmail.com

Aya Eyad Aly
Systems & Biomedical Engineering. Faculty of Engineering
Cairo University
Giza, Egypt
ayaeyad87@gmail.com

1. Introduction

This study investigates the comparative characteristics of the SARS-CoV-2 Delta and Omicron variants in Ghana. Focusing on a representative sample of 10 Delta variant sequences and 10 Omicron variant sequences, we aim to identify key differences between these variants within the Ghanaian context. Specifically, we will construct a consensus sequence from the Delta samples, perform a multiple sequence alignment on the Omicron samples, and analyze the evolutionary relationships between both sets of sequences using a phylogenetic tree. Furthermore, we will compare the chemical composition (C, G, T, and A) and CG content between the two variants, highlighting any notable variations. Finally, we will pinpoint the dissimilar regions within the Omicron sequences in comparison to the Delta consensus sequence, potentially shedding light on the underlying genetic mechanisms contributing to the distinct characteristics of each variant in the Ghanaian population.

2. Methods

2.1 Dataset

The dataset comprises SARS-CoV-2 sequences collected from Ghana, including 15 sequences for each of the Delta and Omicron variants. For analysis, 10 sequences per variant were selected. The sequences are provided in FASTA format and contain nucleotide (DNA) data.

2.2 Tools and Libraries

Biopython: Employed for sequence manipulation, generation of consensus sequences, and construction of phylogenetic trees. **MAFFT** a multiple sequence alignment tool used to align Omicron sequences. **Distance-based methods** (e.g., Neighbor-Joining) were used to build the phylogenetic tree. **Matplotlib and Phylo (Biopython)**: Utilized for visualizing results.

2.3 Algorithms and Parameters

2.3.1 Reference Sequence Analysis

A. Consensus Sequence Generation

The most frequent nucleotide at each position across the 10 Delta sequences was selected to generate the consensus sequence for the Delta variant. Gaps (represented by "-") were excluded from the final consensus sequence to ensure continuity and accuracy. This approach provides a representative sequence that captures the predominant genetic features of the Delta variant.

B. Multiple Sequence Alignment (MAFFT)

Multiple sequence alignment was performed using MAFFT (Multiple Alignment using Fast Fourier Transform), a widely used tool for aligning biological sequences. MAFFT employs a progressive alignment method, which is optimized using the Fast Fourier Transform (FFT) to enhance computational efficiency and alignment accuracy. The default settings (--auto) were applied, allowing the software to automatically select the most appropriate algorithm based on the input data. This ensures an optimal balance between alignment quality and computational speed.

C. Dissimilar Regions Extraction

To identify regions of genetic divergence between the Delta and Omicron variants, the Delta consensus and the alignment of 10 Omicron sequences extracted by the previous steps were utilized. Using the MAFFT --add option, the single consensus was added to the already existing alignment to allow column-by-column comparison.

The columns in the multiple sequence alignment where the Delta consensus sequence differed from the dominant character of Omicron sequences were extracted along with their indices, classified according to the type of variation (insertion, deletion or substitution) and grouped according to the affected nucleotides. The dissimilar columns were also divided into separate columns and whole regions. These comparisons highlight key genomic regions and trends that may contribute to the functional and phenotypic differences between the two variants.

2.3.2 Phylogenetic Tree

A. Sequence Alignment

Multiple sequence alignment of 10 Delta sequences and 10 Omicron sequences was performed using **MAFFT** to prepare for genetic distance calculations and phylogenetic tree construction, which would help elucidate the **evolutionary relationships** between variants.

B. Genetic Distance Matrix

Genetic distances between the aligned sequences were calculated using the **Distance Calculator** from the Biopython library. computing pairwise distances using an **'identity'** matrix. In this matrix, distances are calculated based on the proportion of nucleotide differences between sequences. This measure directly reflects the number of differences between sequences, making it straightforward and interpretable.

C. The Tree Construction

The phylogenetic tree was constructed using the **Neighbor-Joining (NJ) method**. This method is a distance-based approach that creates a tree by iteratively grouping the closest pair of operational taxonomic units (OTUs), building up the tree until all OTUs are included. This method was implemented using the DistanceTreeConstructor from the Biopython library, which takes the previously calculated distance matrix as input to construct the tree.
Some important tree terminologies:

Node: This represents a point on a phylogenetic tree where a lineage splits or diverges. Nodes are critical for understanding evolutionary relationships because each node corresponds to a hypothetical common ancestor from which descendant groups evolved. The node itself is often a point at the end of a branch or at a branching point connecting multiple branches.

Clades: A clade is a group consisting of an ancestor and all its descendants, a single "branch" on the tree of life. Clades are monophyletic, meaning they contain an ancestor, all of its descendants, and no other organisms. Identifying clades is essential for classifying the sequences based on common ancestry rather than just physical similarities.

Taxa: This refers to the different groups or categories in the tree, such as species, genera, families, etc. Taxa represent the actual sequences being studied and placed in the phylogenetic tree, and their relationships are depicted by the structure of the tree—where they appear relative to nodes and within clades.

2.3.3 Nucleotide Content Analysis

Comparing the average percentages of nucleotide constituents (A, T, G, C) and CG content between Delta and Omicron SARS-CoV-2 sequences.

A. Nucleotide Count and Percentage Calculation

Using the **Bio.SeqIO** module from the Biopython library, nucleotide composition was analyzed for each sequence. The nucleotide counts—adenine (A), thymine (T), guanine (G), and cytosine (C)—were determined using the `count` function. The percentage of each nucleotide was then calculated relative to the total sequence length. Additionally, the CG content was computed as the sum of the percentages of cytosine and guanine.

B. Average Nucleotide Composition

After processing individual sequences, averages of A, T, G, and C percentages and CG content were calculated across all sequences within each group (Delta and Omicron). These averages represent the overall nucleotide composition for each group.

C. Comparison of Groups

The averaged values for both Delta and Omicron groups were tabulated for direct comparison.

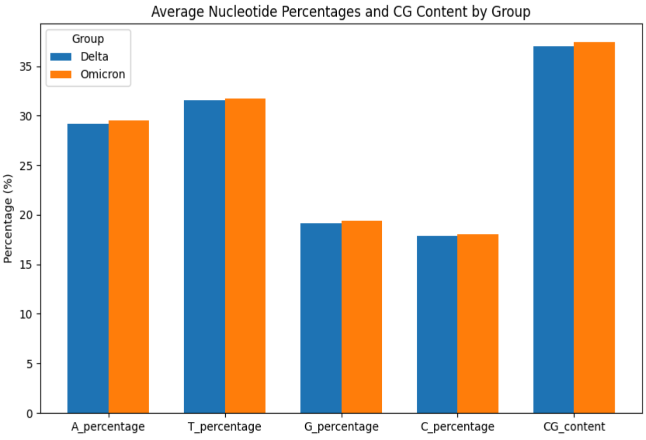


Fig 1. Average nucleotide percentages (A, T, G, C) and CG content in Delta and Omicron SARS-CoV-2 sequences

3. Results

3.1 Consensus Sequence and Alignment

The Delta consensus sequence was constructed by identifying the most frequent nucleotide at each position. Omicron sequences were aligned using MAFFT, revealing conserved and variable regions. Both variants showed a high intra-variant similarity with an average match rate of **0.99**.

3.2 Nucleotide Composition and CG Content

Group	Delta	Omicron
A_percentage	29.168	29.497
T_percentage	31.544	31.763
G_percentage	19.109	19.369
C_percentage	17.886	18.065
CG_content	36.995	37.435

Table 1. The Nucleotide Percentages of the Two Variants

The nucleotide composition of Delta and Omicron variants showed high similarity, with minor variations. Delta had slightly lower A (29.168%) and T (31.544%) percentages compared to Omicron (29.497% and 31.763%). Similarly, G (19.109%) and C (17.886%) percentages in Delta were marginally lower than in Omicron (19.369% and 18.065%). The CG content was nearly identical, at 36.995% for Delta and 37.435% for Omicron, reflecting their genetic similarity and comparable stability.

3.3 Dissimilar Columns and Regions

A total of 309 dissimilar columns were identified between the Delta consensus sequence and the aligned Omicron sequences. These regions likely represent key genetic differences contributing to the unique characteristics of the Omicron variant.

These discrepancies were categorized into insertions, deletions and substitutions, taking the delta consensus as the reference sequence, with the distributions in figure 2. From the visualizations, it is evident that **deletion** of nucleotides was the most prevalent type of mismatch, followed by nucleotide substitutions in the Delta variant sequences.

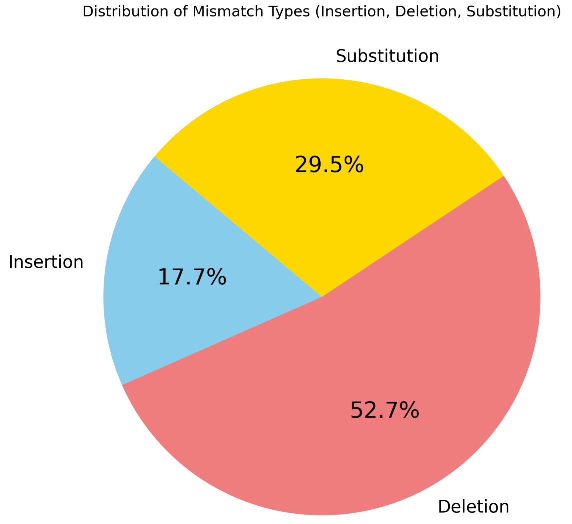


Fig 2. The Distribution of the Mismatch Types

The most frequently **inserted** nucleotide in the delta sequences was the **T nucleotide**, and the most commonly **deleted** from them was the A nucleotide.

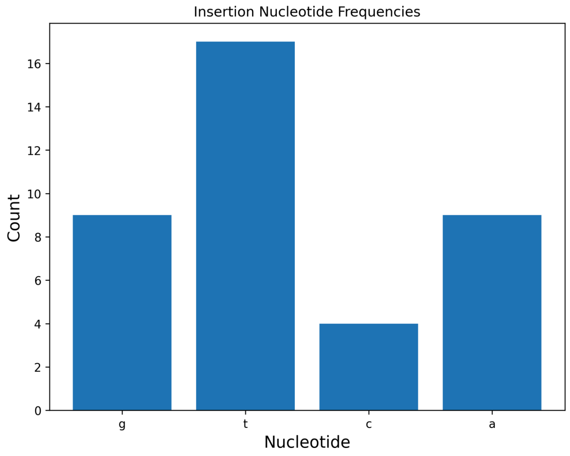


Fig 3. The Distribution of the Inserted Nucleotides

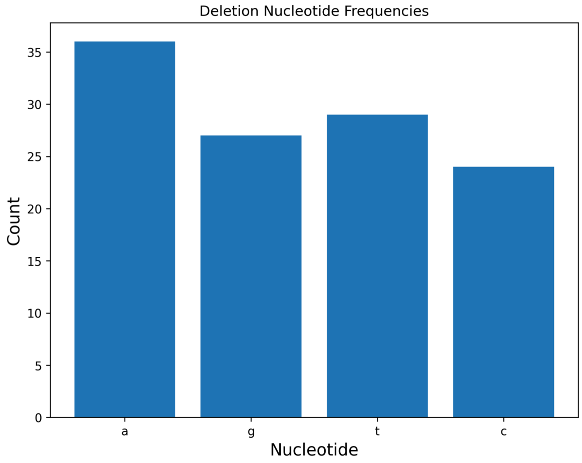


Fig 4. The Distribution of the Deleted Nucleotides

From the substitution heatmap, the most common with frequency 18.46% is nucleotide C substituting T in Delta, followed by T substituting C and G substituting A with percentages 16.92% and 15.38% respectively.

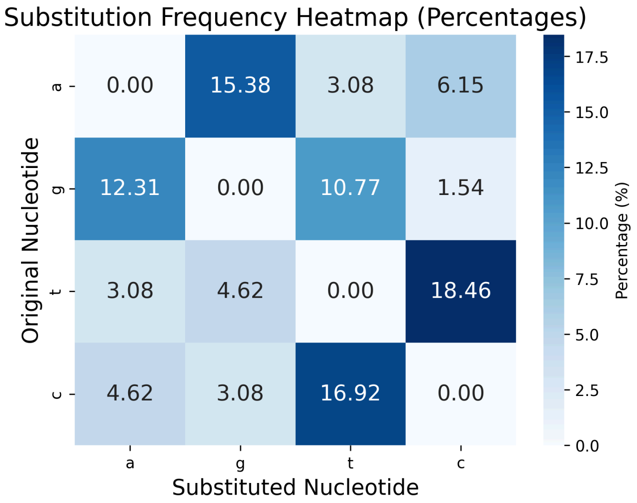


Fig 5. The Percentages of the Substitutions across Nucleotides

These 309 varying columns were arranged into contiguous regions and separate columns, 12 regions of consecutive columns were found, covering 80.2% of the number of dissimilar columns with the shortest region consisting of 2 nucleotides and the longest of 103. These regions varied in length as follows:

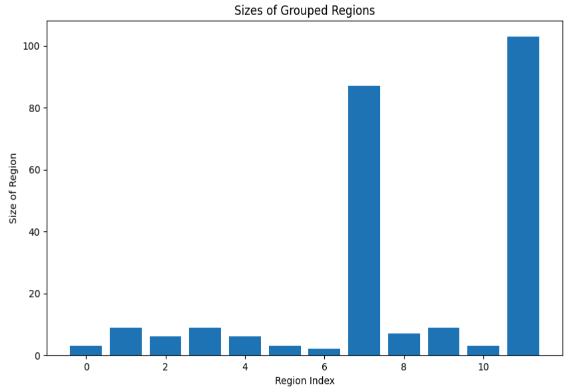


Fig 6. The Lengths of Dissimilar Regions between the Delta Consensus and Omicron Sequences

The positions of the dissimilar columns across the alignment of 11 sequences, the Delta consensus and the 10 aligned Omicron sequences, can be observed in the following graph. Note that the length of the final alignment is 29800 nucleotides.

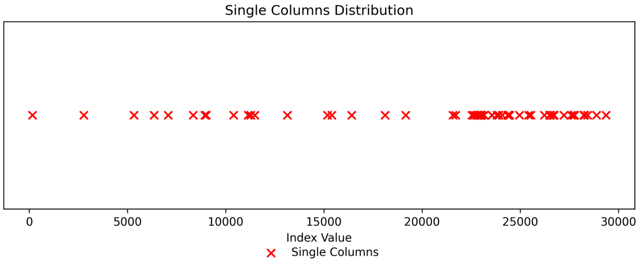


Fig 7. The indices of the single dissimilar columns (SNPs) of the Delta variant.

While the positions and lengths of the dissimilar regions, of more than one consecutive column, can be visualized in the following figure:

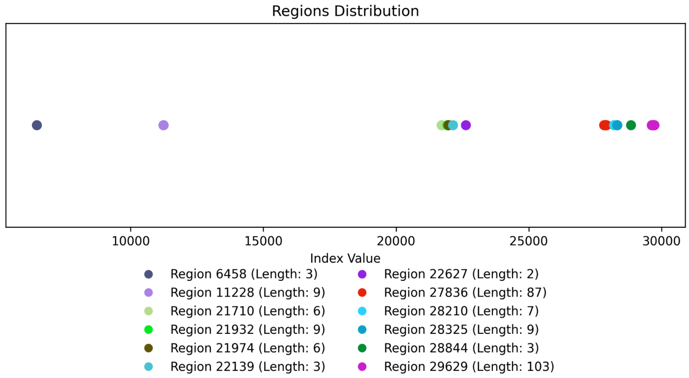


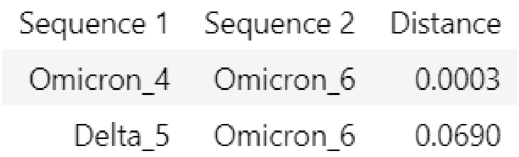
Fig 8. The Positions and lengths of Dissimilar Regions in the final alignment

Notably, the majority of these mismatches were concentrated in the **final third of the sequences**, specifically beyond index 21,000. Additionally, **two large regions of divergence**, each spanning over **80** contiguous nucleotides, were prominently identified across mismatching regions.

3.4 Phylogenetic Tree Results

A. Key findings from the Distance Matrix

The **minimum genetic distance (0.0002)** observed was between **Omicron_4 and Omicron_6**, indicating a **high genetic similarity**.
The **maximum genetic distance (0.0689)** was found between **Delta_5 and Omicron_6**, illustrating **significant genetic divergence**.



Sequence 1	Sequence 2	Distance
Omicron_4	Omicron_6	0.0003
Delta_5	Omicron_6	0.0690

Fig 9. The Min and Max distances extracted from Matrix

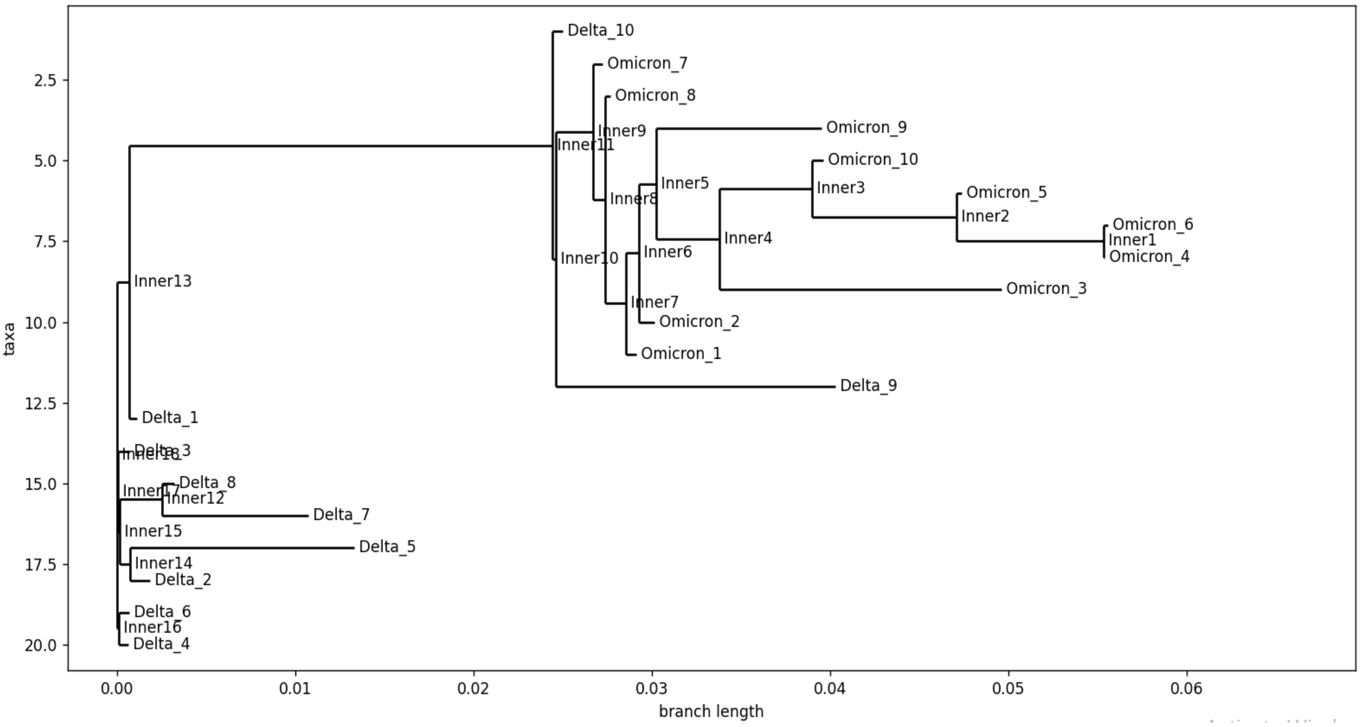


Fig 10. The Full Phylogenetic Tree Constructed

B. *Key findings from the Phylogenetic Tree*

The tree consists of **38 clades** in which each may have one descendant or more than one coming from the same ancestor.
Clustering of Omicron Variants: These sequences form a dense cluster with smaller branch lengths among them, suggesting close genetic relationships.
Branching of Delta Variants: Delta variants appear more dispersed. Delta_10 and Delta_9 are relatively close to the Omicron cluster.On the other hand, sequences like Delta_7 and Delta_8 form another sub-cluster, and Delta_5 and Delta_2 are closely related with another small branch length

The use of MAFFT for alignment and Biopython for analysis proved effective, providing reliable and reproducible results.

Phylogenetic trees offer critical insights into disease transmission and mutation rates, serving as key tools for epidemiological analysis. By mapping how variants spread and evolve, these trees are vital for public health planning and responses. Analyzing sequence clustering within these trees helps direct targeted surveillance and control efforts, Additionally, the identification of highly divergent sequences informs the design and updates of vaccines.

4. DISCUSSION

The study compared the Delta and Omicron variants using genomic data from Ghana. The phylogenetic tree confirmed the evolutionary divergence between the two variants, consistent with global observations.

The nucleotide composition analysis revealed minimal differences, suggesting that the functional differences between the variants may arise from specific mutations rather than overall nucleotide bias.

The dissimilar region comparison identified trends in dissimilarity, most notably insertion of the T nucleotide in the Delta sequences, deletion of the A nucleotides, although less prominent, and the substitution of the T with C nucleotides. As well as two large regions of variation, starting at indices 29629 and 27836 of lengths 103 and 87 respectively. These regions could be potential targets for further investigation into the biological and clinical implications of these mutations.

5. CONCLUSION

This study provides a comprehensive comparison of the SARS-CoV-2 Delta and Omicron variants using genomic data from Ghana. The results highlight the genetic divergence between the variants and identify key dissimilar regions and genetic content that may contribute to their distinct characteristics. It also shows which variants might be developing in similar ways and which are becoming more different. This can help in tracking the virus's spread and evolution and could be crucial for developing strategies to combat the virus, like improving treatments or adjusting vaccines to better match the most current forms of the virus. The methods and tools used in this study can be applied to analyze other viral variants and datasets, contributing to a better understanding of viral evolution and its impact on public health.

6. References

Katoh, Kazutaka, and Daron M. Standley. "MAFFT multiple sequence alignment software version 7: improvements in performance and usability." Molecular biology and evolution 30.4 (2013): 772-780.

Felsenstein, Joseph. "Confidence limits on phylogenies: an approach using the bootstrap." evolution 39.4 (1985): 783-791.

Cock, Peter JA, et al. "Biopython: freely available Python tools for computational molecular biology and bioinformatics." Bioinformatics 25.11 (2009): 1422.

Kluyver, Thomas, et al. "Jupyter Notebooks–a publishing format for reproducible computational workflows."

7. Member Contributions

Shehab Mohamed: Developed the consensus sequence by performing Multiple Sequence Alignment (MSA) and comparing the sequences position by position. Additionally, conducted MSA on the other 10 sequences..

Aya Eyad: Aligned the consensus sequence with the MSA output from Step 2 and performed column-by-column comparisons to identify the dissimilar regions. Analysis, comparisons and visualizations of dissimilar regions. Code Documentation and scripting.

Camellia Marwan: Applied sequence alignment between the 20 sequences, constructed the distance matrix and visualized the phylogenetic tree and extracted the most important conclusions.

Farah Osama: Analyzed and compared the CG content and the percentage of nucleotides (A, T, G, C) between the two classes.Additionally showing the analysis visually.

8. Associated Output files

FASTA files:

- **delta_consesnsus.fasta**: Carries a single record which is the extracted delta consensus from the 10 sequences.
- **aligned_omicron.fasta**: Carries the 10 aligned omicron sequences out of the MSA step.
- **aligned_all_delta_omicron_sequences.fasta:**Carries the alignment of the 20 sequences produced by aligning 10 Delta sequences and 10 Omicron sequences., used in the phylogenetic tree.
- **delta_omicron_aligned.fasta:** Carries the alignment of 11 sequences, produced by aligning the Delta consensus to the already aligned alignment of 10 Omicron Files.
- **dissimilar_cols.fasta:** Carries 11 sequences, which are the alignment of 10 Omicron sequences and the Delta consensus, but only the dissimilar columns from delta_omicron_aligned.fasta file, the indices of these columns indelta_omicron_aligned.fasta can be found in indices_dissimilar_cols.txt.

CSV files:

- **delta_group_analysis.csv: omicron_group_analysis.csv**
- **seq_names.csv:** The corresponding sequence ids and their names in the phylogenetic tree.

PNG files:

- **comparison_plot:**
- **insertion_nucleotide_frequencies:** Frequency of insertion per nucleotide.
- **deletion_nucleotide_frequencies:** Frequency of deletion per nucleotide.
- **grouped_vs_single_pie_chart:** Pie chart carrying the percentage of length of the two categories (dissimilar region vs dissimilar residue) over the number of all dissimilar columns.
- **mismatch_types_pie_chart:** Pie chart carrying percentages of deletions, insertions and substitutions in dissimilar columns.
- **Phylo_Tree:** shows a mixture of 20 Delta and Omicron sequences with various branch lengths suggesting different evolutionary distances.
- **region_sizes_bar_plot:** The lengths of dissimilar regions between the Delta consensus and the alignment of 10 Omicron sequences.
- **substitution_heatmap:** The percentages of different substitution combinations for dissimilar columns.
- **single_columns_scatter_plot**: The scatter plot of the indices of SNPs in the alignment of 11 sequences.
- **regions_scatter_plot**: The scatter plot of the indices and lengths of dissimilar regions in the alignment of 11 sequences.

Text files:

- **indices_dissimilar_cols.txt:** The indices of the dissimilar columns in the alignment of 11 sequences (delta consensus and 10 Omicron sequences).

Others:

- **phylo_tree.nwk.pdf**