# Developing Mutation Detection Algorithms for SARS-CoV-2 Covid 19 Genomic Sequences

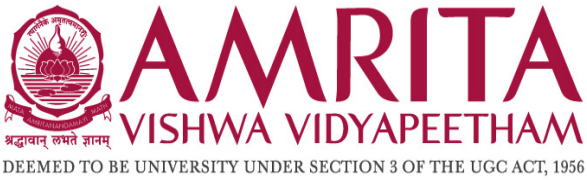
PROJECT REPORT

*Submitted by*

**Group A9**
**Abhishek Karthik J (CB.SC.U4AIE23010)**
**Keerthivasan S V (CB.SC.U4AIE23037)**
**Mothishwaran M P (CB.SC.U4AIE23041)**
**Mopuru Sai Bavesh Reddy (CB.SC.U4AIE23044)**

*in partial fulfillment for the award of the degree of*

**BACHELOR OF TECHNOLOGY**
**IN**
**COMPUTER SCIENCE ENGINEERING**
**(ARTIFICIAL INTELLIGENCE)**




**COMPUTER SCIENCE ENGINEERING - ARTIFICIAL INTELIIGENCE**

**AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE**

**AMRITA VISHWA VIDYAPEETHAM**

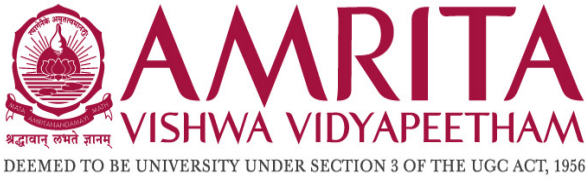
COIMBATORE - 641 112 (INDIA)

**APRIL - 2025**

**COMPUTER SCIENCE ENGINEERING - ARTIFICIAL INTELLIGENCE**

**AMRITA VISHWA VIDYAPEETHAM**

COIMBATORE - 641 112




# BONAFIDE CERTIFICATE

This is to certify that the report entitled **"Developing Mutation Detection Algorithms for SARS-CoV-2 Covid 19 Genomic Sequences"** submitted by **Abhishek Karthik J (CB.SC.U4AIE23010)**,**Keerthivasan S V (CB.SC.U4AIE23037)**,**Mothishwaran M P (CB.SC.U4AIE23041)**,**Mopuru Sai Bavesh Reddy (CB.SC.U4AIE23044)** in partial fulfillment of the requirements for the courses 22B10211 Intelligence of Biological Systems (IBS2) and 22AlE212 Design and Analysis of Algorithms (DAA) in the **Degree of Bachelor of Technology** in **"COMPUTER SCIENCE ENGINEERING - ARTIFICIAL INTELLIGENCE"** is a bonafide record of the work carried out by them under our guidance at Amrita School of Artificial Intelligence, Coimbatore.

**Dr. Harishchander & Ms. Rema M**
Project Guides
Designation: Professor & Associate Professor

*Submitted for the university examination held on 10th of April 2025*

**INTERNAL EXAMINER** **EXTERNAL EXAMINER**

## AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE

# AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112

## DECLARATION

We, **Abhishek Karthik J (CB.SC.U4AIE23010)**,**Keerthivasan S V (CB.SC.U4AIE23037)**,**Mothishwaran M P (CB.SC.U4AIE23041)**,**Mopuru Sai Bavesh Reddy (CB.SC.U4AIE23044)** hereby declare that this project entitled **"Developing Mutation Detection Algorithms for SARS-CoV-2 Covid 19 Genomic Sequences"**, is the record of the original work done by us under the guidance of **Dr. Harishchander & Ms. Rema M**, Amrita School of Artificial Intelligence, Coimbatore. To the best of our knowledge, this work has not formed the basis for the award of any degree/diploma/ associateship/fellowship/or a similar award to any candidate in any university.

**Place: Coimbatore** **Signature of the Students Date: 11-04-2025**

COUNTERSIGNED

Dr. K.P.Soman

Professor and Dean

Amrita School of Artificial Intelligence

Amrita Vishwa Vidyapeetham

# Contents

# List of Figures

# List of Tables

# Acknowledgement

We would like to express our sincere appreciation to all those who have helped and guided us throughout this project. Most importantly, we are significantly grateful to our project guides, **Guide Name 1 & Guide Name 2**, for their constant encouragement, expert advice, and constructive criticism. Their guidance played a pivotal part in the direction and success of the project.

We would also like to express our gratitude to the staff and faculty at Amrita School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore, for providing us with the facilities, resources, and learning environment necessary for the successful completion of our research. Their support and expertise were invaluable in overcoming the technical challenges we encountered.

We appreciate our peers and teammates for their cooperation, motivation, and enthusiasm. Their help and support enabled us to overcome obstacles and achieve our project objectives. The collaborative atmosphere fostered by our team was essential to the success of this research.

We would also like to acknowledge the various research groups and organizations that have made SARS-CoV-2 genomic data publicly available, which was essential for

our analysis.

Lastly, we would like to thank our families for supporting us throughout, for being patient with us, and for believing in us along the way during our learning process. This entire project has been a valuable experience, and we thank all those who have been involved in the process to its successful conclusion.

# List of Abbreviations

| | | |
|---|---|---|
| SARS-CoV-2 | - | Severe Acute Respiratory Syndrome Coronavirus 2 |
| MSA | - | Multiple Sequence Alignment |
| MAFFT | - | Multiple Alignment using Fast Fourier Transform |
| PRANK | - | Probabilistic Alignment Kit |
| FFT | - | Fast Fourier Transform |
| CSV | - | Comma-Separated Values |
| UPGMA | - | Unweighted Pair Group Method with Arithmetic Mean |
| NJ | - | Neighbor-Joining |
| HRM | - | High-Resolution Melting |
| RT-PCR | - | Reverse Transcription Polymerase Chain Reaction |
| CNN | - | Convolutional Neural Network |
| RNA | - | Ribonucleic Acid |
| VOC | - | Variant of Concern |

# Abstract

This research project focuses on developing and implementing multiple sequence alignment (MSA) algorithms for detecting mutations in SARS-CoV-2 genomic sequences. The emergence of SARS-CoV-2 variants with potentially increased transmissibility and impact on vaccine efficacy highlights the need for robust computational methods to analyze genomic variations.

Our study implements two MSA algorithms from scratch: Multiple Alignment using Fast Fourier Transform (MAFFT) and Probabilistic Alignment Kit (PRANK). We apply these algorithms to analyze genomic sequences collected from three U.S. states: California, Texas, and New York. The implementation includes data collection, preprocessing, alignment, consensus sequence generation, and phylogenetic analysis.

The research demonstrates significant differences in mutation patterns across geographic regions. Our analysis reveals that California and Texas sequences exhibit high similarity (90.08%) using MAFFT, while New York sequences show greater divergence. Contrastingly, PRANK-based analysis indicates more uniform divergence across all states. The study also analyzes evolutionary drift from the Wuhan reference strain through phylogenetic tree construction and compares regional consensus sequences with

known Variants of Concern (VOCs) including Alpha, Beta, Delta, Omicron, and the Omicron subvariant XBB.1.5 ("Kraken").

This work contributes to our understanding of SARS-CoV-2 genomic variations and demonstrates the efficacy of computational approaches in tracking viral evolution. The methodologies developed can be extended to other viral genomic analyses and may aid in epidemiological research and public health response strategies.

# Chapter 1

# Introduction

The emergence of SARS-CoV-2, the virus responsible for the COVID-19 pandemic, has prompted extensive research efforts to understand its genomic structure, evolution, and mutations. As viruses replicate, they accumulate genetic changes that can affect their transmissibility, virulence, and response to vaccines or treatments. Tracking these mutations is crucial for public health surveillance and pandemic response strategies.

This research focuses on developing and implementing multiple sequence alignment (MSA) algorithms to detect and analyze mutations in SARS-CoV-2 genomic sequences. We specifically examine sequences from three U.S. states—California, Texas, and New York—to understand regional variations in viral evolution and compare them with known Variants of Concern (VOCs).

The study implements two MSA algorithms from scratch: Multiple Alignment using Fast Fourier Transform (MAFFT) and Probabilistic Alignment Kit (PRANK). These algorithms allow for efficient and accurate alignment of multiple genomic sequences, enabling the identification of conserved regions and mutations.

Our methodology encompasses several key stages: data collection and filtering, se-

quence preprocessing and cleaning, multiple sequence alignment using MAFFT and PRANK, consensus sequence generation, and phylogenetic analysis. By comparing the results from different algorithms and across different regions, we aim to gain insights into the evolutionary patterns of SARS-CoV-2.

The findings of this research contribute to our understanding of SARS-CoV-2 genomic variations and demonstrate the effectiveness of computational approaches in tracking viral evolution. The methodologies developed can be extended to other viral genomic analyses and may aid in epidemiological research and public health response strategies.

## 1.1 Literature Survey

Our review of recent literature reveals significant advancements in SARS-CoV-2 mutation detection and analysis:

- "An Update on Detection Technologies for SARS-CoV-2 Variants of Concern" highlighted the effectiveness of RT-PCR assays in identifying variants by targeting known mutation sites, enabling quicker detection compared to sequencing methods.

- "Enhanced Detection and Molecular Modeling of Adaptive Mutations in SARS-CoV-2" (2023) applied a site-by-site analysis to identify nucleotide and amino acid sites under selection, providing insights into adaptive mutations that could influence viral behavior.

- "Global Landscape of SARS-CoV-2 Mutations and Conserved Regions" (2023) conducted a comprehensive analysis of SARS-CoV-2 genomes to identify mutations and conserved regions, providing a systematic resource for novel sequence features, aiding in the development of vaccines and therapeutics by highlighting conserved regions less prone to mutation.

- "Tracking Mutational Semantics of SARS-CoV-2 Genomes" (2022) employed natural language processing algorithms to analyze SARS-CoV-2 genome mutations, demonstrating that NLP-based approaches can effectively process genomic data, revealing characteristics and evolutionary patterns of the virus.

- "Rapid Detection of Predominant SARS-CoV-2 Variants Using High-Resolution Melting Analysis" (2023) developed a multiplex high-resolution melting (HRM) method for SARS-CoV-2 detection and mutation site identification, establishing a one-step multiplex HRM assay capable of detecting SARS-CoV-2 and identifying specific mutation sites, facilitating rapid variant analysis.

Our research builds upon these foundations while implementing novel approaches to multiple sequence alignment and phylogenetic analysis of SARS-CoV-2 genomes.

## 1.2 Problem Statement

The rapid evolution of SARS-CoV-2 has resulted in multiple variants with potentially different transmissibility, virulence, and response to vaccines. Traditional methods for genomic sequence analysis often rely on existing tools without exploring the underlying

algorithms or customizing them for SARS-CoV-2-specific characteristics. This research addresses the need for:

- Implementation of multiple sequence alignment algorithms from scratch to provide greater transparency and customization for SARS-CoV-2 analysis
- Comparison of different alignment approaches (MAFFT and PRANK) to determine their relative effectiveness for viral mutation detection
- Regional analysis of SARS-CoV-2 sequences to understand geographic patterns of viral evolution
- Development of computational methods to identify mutations and assess their potential impact
- Comparison of regional sequences with known Variants of Concern to identify regions at higher risk

## 1.3 Objectives

- Develop MSA-based approaches (MAFFT and PRANK) from scratch to analyze SARS-CoV-2 genomic sequences
- Generate consensus viral sequences from multiple samples across California, New York, and Texas
- Compare these sequences with the Wuhan reference strain to track evolutionary divergence

- Identify mutation hotspots and regional variation patterns
- Construct phylogenetic trees to visualize evolutionary relationships between regional strains
- Implement efficient algorithms for sequence alignment, mutation detection, and phylogenetic analysis
- Compare regional consensus sequences with known Variants of Concern to assess potential public health risks
- Analyze the likelihood of disease spread between states based on sequence similarity

## 1.4 Organization of the Report

The structure of this report is outlined below:

- **Chapter 1: Introduction**

  Provides the background, literature survey, problem statement, and objectives of the project.

- **Chapter 2: Background**

  Discusses the key concepts including SARS-CoV-2 genomic characteristics, multiple sequence alignment techniques, and phylogenetic analysis methods.

- **Chapter 3: Proposed Work**

Describes the methodology for data collection, preprocessing, implementation of MAFFT and PRANK algorithms, and phylogenetic analysis.

- **Chapter 4: Results and Inferences**

  Presents the findings from our sequence alignments, consensus sequence comparisons, phylogenetic analyses, and comparisons with Variants of Concern, along with their implications.

- **Chapter 5: Conclusion and Future Work**

  Summarizes the key findings and suggests directions for future research.

# Chapter 2

# Background

## 2.1 SARS-CoV-2 Genomic Characteristics

SARS-CoV-2 is a positive-sense single-stranded RNA virus with a genome of approximately 30,000 nucleotides. The genome encodes structural proteins including spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins, as well as several non-structural proteins with functions in viral replication and host immune evasion.

The virus exhibits a mutation rate typical of RNA viruses, which contributes to its genetic diversity and evolution. These mutations can affect viral properties such as transmissibility, virulence, and immune escape, making genomic surveillance an essential component of pandemic response.

The SARS-CoV-2 genome is organized into specific functional regions. The 5' end of the genome contains the ORF1a and ORF1b regions, which encode the replicase polyproteins that are processed into non-structural proteins essential for viral replication. The 3' end contains genes encoding the structural proteins, with the spike protein gene being particularly important as it determines host cell tropism and is a primary

target for host immunity.

The virus undergoes natural selection pressure that shapes its evolution. Mutations that confer advantages in transmission, replication efficiency, or immune evasion tend to be positively selected. This selection process has led to the emergence of multiple variants of concern throughout the pandemic, each characterized by specific constellations of mutations, particularly in the spike protein region.

Understanding these genomic characteristics and tracking their changes over time and across geographic regions is crucial for public health surveillance, vaccine development, and therapeutic strategies. Our research focuses on developing computational methods to efficiently analyze these genomic variations and identify patterns that may have epidemiological significance.

## 2.2 Variants of Concern (VOCs)

Throughout the COVID-19 pandemic, certain SARS-CoV-2 variants have been designated as Variants of Concern (VOCs) due to their potential impact on transmissibility, disease severity, or effectiveness of vaccines and therapeutics. This study focuses on five significant VOCs:

- **Alpha (B.1.1.7)**: First identified in the United Kingdom in September 2020, Alpha is characterized by multiple mutations in the spike protein, including N501Y, which enhances binding to human ACE2 receptors. This variant demonstrated approximately 50% increased transmissibility compared to previous strains and was associated with increased disease severity in some studies.

- **Beta (B.1.351)**: Originating in South Africa in May 2020, the Beta variant contains multiple mutations in the spike protein, including K417N, E484K, and N501Y. These mutations have been linked to reduced neutralization by antibodies generated through previous infection or vaccination, raising concerns about immune escape.

- **Delta (B.1.617.2)**: Identified in India in October 2020, Delta became globally dominant by mid-2021 due to its substantially increased transmissibility. Key mutations include L452R and P681R in the spike protein, which enhance viral replication and cell entry. Delta was associated with higher viral loads, increased disease severity, and some reduction in vaccine effectiveness.

- **Omicron (B.1.1.529)**: Detected in November 2021, Omicron represents a significant evolutionary jump with over 30 mutations in the spike protein alone. These extensive changes resulted in enhanced transmissibility and substantial immune escape from both natural and vaccine-induced immunity. While generally causing less severe disease than Delta, Omicron's extreme transmissibility led to significant global impact.

- **Omicron Subvariant XBB.1.5 ("Kraken")**: Emerging in late 2022, XBB.1.5 is a recombinant of two BA.2 sublineages with additional mutations, particularly F486P in the spike protein. This mutation enhances binding to ACE2 receptors while maintaining the immune evasion capabilities of its parent lineages, making it one of the most transmissible variants to date.

These VOCs have significantly impacted the course of the pandemic, necessitating adjustments in public health strategies, vaccine development, and therapeutic approaches. By comparing genomic sequences from different U.S. states with these known VOCs, our study aims to identify regions potentially harboring variants with similar characteristics, which could inform targeted surveillance and intervention efforts.

## 2.3 Multiple Sequence Alignment

Multiple Sequence Alignment (MSA) is a computational method used to align three or more biological sequences (DNA, RNA, or protein) to identify regions of similarity. These alignments can reveal conserved domains, functional regions, and evolutionary relationships between sequences.

MSA algorithms typically aim to maximize the overall similarity score of the alignment, which reflects the conservation of nucleotides or amino acids across all sequences at each position. Gaps are introduced to compensate for insertions or deletions that may have occurred during evolution. The alignment process seeks to place homologous positions in the same column, revealing the evolutionary relationships and functional constraints operating on the sequences.

MSA is a fundamental tool in bioinformatics with applications in:

- Phylogenetic analysis to infer evolutionary relationships
- Identification of conserved functional elements in genes or proteins
- Structural prediction of proteins

- Detection of mutations and their potential impact on function

In the context of SARS-CoV-2 research, MSA is particularly valuable for identifying variants, tracking mutation patterns, and understanding viral evolution across time and geography. The effectiveness of MSA depends on the alignment algorithm used, with different algorithms employing different approaches to balance accuracy and computational efficiency.

### 2.3.1 MAFFT Algorithm

MAFFT (Multiple Alignment using Fast Fourier Transform) is a widely-used algorithm for multiple sequence alignment. It aligns nucleotide or protein sequences to identify regions of similarity that may indicate functional, structural, or evolutionary relationships between the sequences.

Key features of MAFFT include:

- Speed: MAFFT is faster than many traditional alignment methods due to the use of FFT.
- Accuracy: It provides high-quality alignments by considering both local and global sequence similarities.
- Versatility: MAFFT can handle a wide range of sequence lengths and numbers of sequences.

The MAFFT algorithm consists of four main steps:

1. **Compute the Similarity Matrix using FFT ($O(n \log n)$)**: Instead of directly comparing sequences character by character, MAFFT uses Fast Fourier Transform to convert sequences into frequency representations. This is done by first encoding nucleotides numerically (e.g., A=1, C=2, G=3, T=4), then applying FFT to transform these numerical sequences into frequency components. The similarity between sequences can then be calculated by comparing these frequency spectra, which is much faster than direct comparison, especially for long sequences.

2. **Construct the Guide Tree ($O(N^2)$)**: Using the similarity matrix obtained in the previous step, MAFFT builds a hierarchical clustering of sequences called a guide tree. This tree determines the order in which sequences will be aligned. Sequences with higher similarity are clustered first, creating a strategy for progressive alignment. This is typically implemented using algorithms like UPGMA (Unweighted Pair Group Method with Arithmetic Mean) or Neighbor-Joining.

3. **Progressive Alignment ($O(N^2)$)**: Sequences are aligned progressively following the order defined by the guide tree. Starting from the most similar sequences, pairwise alignments are performed and extended as we move up the tree. This approach ensures that the most similar sequences are aligned first, reducing the chances of misalignment.

4. **Iterative Refinement ($O(N^3)$)**: To improve the initial alignment, MAFFT employs an iterative refinement strategy. This involves removing one sequence

at a time from the current alignment, realigning it to the remaining profile, and checking if the overall alignment score improves. This process is repeated until no further improvement is observed or a maximum number of iterations is reached.

### 2.3.2 PRANK Algorithm

PRANK (Probabilistic Alignment Kit) is a phylogeny-aware multiple sequence alignment algorithm. Unlike traditional methods, PRANK distinguishes between insertions and deletions during the alignment process, which helps in more accurately representing the evolutionary history of sequences.

Key features of PRANK include:

- Phylogeny-aware: PRANK uses evolutionary information to guide the alignment process
- Accurate indel handling: It distinguishes between insertions and deletions, reducing alignment errors
- Evolutionary modeling: Incorporates evolutionary models to better represent sequence changes over time
- Reduced alignment bias: Avoids systematic errors common in progressive alignment methods

The PRANK algorithm consists of four main steps:

1. **Construct a Guide Tree ($\mathbf{O(N^2)}$)**: PRANK begins by estimating evolutionary distances between sequences, typically using metrics like Hamming distance or

more sophisticated evolutionary models. These distances are then used to build a phylogenetic guide tree that represents the evolutionary relationships between sequences. The guide tree is crucial as it determines not just the order of alignment but also helps in tracking evolutionary events.

2. **Distinguish Insertions and Deletions ($O(N^2)$)**: A key innovation in PRANK is its ability to differentiate between insertions and deletions. As the alignment progresses, PRANK tracks the history of indels (insertions and deletions) and labels new gaps as either insertions or deletions based on the guide tree information. This prevents the same gap from being penalized multiple times in downstream alignments, which is a common problem in traditional progressive alignment methods.

3. **Phylogeny-Aware Progressive Alignment ($O(N^3)$)**: Sequences are aligned following the topology of the guide tree. At each internal node of the tree, PRANK reconstructs ancestral sequences based on the descendants. When aligning sequences or profiles, the algorithm considers the evolutionary changes that might have occurred since their divergence from their common ancestor, leading to more biologically accurate alignments.

4. **Using Posterior Probabilities for Alignment Decisions ($O(N^2)$)**: PRANK uses a probabilistic framework to make alignment decisions. It calculates the probability of different evolutionary scenarios (e.g., insertion, deletion, substitution) and chooses the one with the highest posterior probability. This approach min-

imizes systematic errors in gap placement, particularly in regions with multiple indels, resulting in alignments that better reflect the true evolutionary history.

## 2.4 Phylogenetic Analysis

Phylogenetic analysis is a method used to study the evolutionary relationships between different species, organisms, or genes. In the context of SARS-CoV-2, phylogenetic analysis helps in tracking the spread and evolution of the virus over time and across geographical regions.

Key methods in phylogenetic analysis include:

- Distance-based methods (e.g., Neighbor-Joining): These methods calculate a distance matrix between sequences and use it to construct a tree. The Neighbor-Joining algorithm, in particular, starts with a star-like tree and iteratively joins the closest pairs of nodes, recalculating distances at each step.

- Maximum likelihood: This method evaluates the probability of observing the given sequences under a specific evolutionary model. It selects the tree topology and branch lengths that maximize this probability.

- Bayesian inference: This approach uses Bayesian statistics to estimate the posterior probability of a tree given the sequence data and prior information about evolutionary processes.

Phylogenetic trees provide a visual representation of the evolutionary relationships

between sequences, with branch lengths often indicating the amount of genetic change or time since divergence.

In our research, we focus primarily on distance-based methods for phylogenetic analysis, specifically the Neighbor-Joining algorithm. This approach is computationally efficient and well-suited for analyzing large datasets of viral sequences. The phylogenetic trees generated from our analysis help visualize how SARS-CoV-2 sequences from different regions relate to each other and to the original Wuhan reference strain, providing insights into the patterns of viral evolution and spread.

# Chapter 3

# Proposed Work

## 3.1 Data Collection and Processing

### 3.1.1 Dataset Description

Our research utilizes a comprehensive dataset of SARS-CoV-2 genomic sequences collected from three U.S. states: California, Texas, and New York. The dataset contains over 20,000 complete genomic sequences with associated metadata including:

- Unique sequence identifiers/accession numbers
- Complete nucleotide sequences (approximately 30,000 bases per genome)
- Collection dates and geographic locations
- Host information
- Release dates and submitter details

This rich dataset provides a robust foundation for analyzing regional patterns in viral evolution and identifying state-specific mutation trends. The sequences represent

samples collected across different time periods of the pandemic, allowing for temporal analysis of viral mutations in addition to geographic comparisons.

### 3.1.2 Data Collection Methodology

The data collection process involved several key steps to ensure a comprehensive and representative dataset:

1. **Source Identification**: We identified reliable repositories of SARS-CoV-2 genomic data, focusing on databases with well-documented metadata and quality control measures.

2. **Data Extraction**: We extracted sequence data and metadata from these repositories, ensuring that all necessary fields were captured for subsequent analysis.

3. **Regional Filtering**: The dataset was filtered to specifically select sequences from our three target states: California, Texas, and New York. This filtering was based on the location information in the metadata.

4. **Format Standardization**: The collected data was converted into a standardized CSV format with unified field names and consistent data types to facilitate further processing.

5. **Quality Verification**: Initial quality checks were performed to identify and flag any obvious anomalies in the data, such as incomplete sequences or missing critical metadata.

Additionally, we collected reference sequences for key Variants of Concern, including Alpha (B.1.1.7), Beta (B.1.351), Delta (B.1.617.2), Omicron (B.1.1.529), and the

Omicron subvariant XBB.1.5 ("Kraken"). These reference sequences were essential for our comparative analysis of regional variants with known VOCs.

This methodical approach resulted in a well-structured dataset tailored to our research objectives, providing a solid foundation for the subsequent preprocessing and analysis steps.

### 3.1.3 Data Preprocessing and Cleaning

Before applying alignment algorithms, we implemented a rigorous preprocessing pipeline to ensure data quality and consistency:

1. **Sequence Validation**: Each nucleotide sequence was validated to ensure it contained only standard nucleotide characters (A, T, C, G). Sequences with excessive ambiguous nucleotides (e.g., N) were identified and flagged.

2. **Length Filtering**: Sequences shorter than 29,000 nucleotides were excluded from the analysis, as they likely represented incomplete or low-quality genomes. This threshold was chosen based on the typical length of complete SARS-CoV-2 genomes (approximately 30,000 nucleotides).

3. **Standardization**: All sequences were converted to uppercase to ensure consistent character case throughout the dataset.

4. **Ambiguity Removal**: Ambiguous nucleotides (e.g., N, R, Y) were removed from sequences to ensure that only definitive nucleotides (A, T, C, G) were used in the alignment process. This step was crucial for accurate mutation detection and comparison.

5. **Format Conversion**: The cleaned sequences were converted to FASTA format, with each sequence preceded by a header line containing the sequence identifier. The FASTA files were structured with line wrapping at 80 characters for better readability.

6. **State-Specific Files**: Separate FASTA files were created for each state (California, Texas, and New York) to facilitate state-specific analyses and comparisons.

This preprocessing pipeline ensured that our analyses were based on high-quality, complete genomic sequences, minimizing the risk of artifacts or false positives in our mutation detection results. The resulting cleaned FASTA files served as the input for our MAFFT and PRANK implementations.

## 3.2 Implementation of MAFFT Algorithm

### 3.2.1 Numerical Encoding and FFT Application

Our implementation of the MAFFT algorithm begins with the application of Fast Fourier Transform (FFT) to efficiently compare sequence similarities:

1. **Nucleotide Encoding**: Each nucleotide in the sequences is converted to a numerical value (A=1, C=2, G=3, T=4) to create a numerical signal suitable for FFT processing. This encoding preserves the sequential information while converting biological data into a mathematical form.

2. **Fourier Transform Application**: The FFT algorithm is applied to these numerical sequences, transforming them from the time domain to the frequency domain. This transformation captures periodic patterns in the sequence data that might not be apparent in the original form.

3. **Spectrum Analysis**: The magnitude spectrum of each transformed sequence is calculated, representing the frequency components present in the original sequence. These magnitude spectra serve as characteristic "fingerprints" of the sequences.

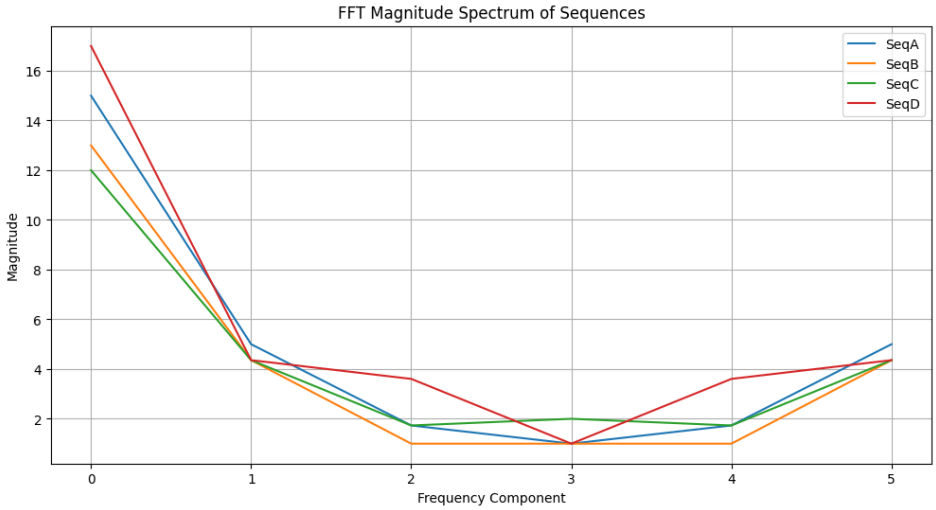



Figure 3.1: FFT Magnitude Spectrum of Sequences showing the frequency components for four sample sequences (SeqA, SeqB, SeqC, and SeqD). The distinct patterns in the magnitude spectrum allow for efficient sequence comparison.

4. **Fingerprint Comparison**: The similarity between sequences is assessed by comparing their frequency spectra rather than direct character-by-character comparison. This approach is particularly efficient for long sequences like viral genomes, as the computational complexity scales with O(n log n) rather than $O(n^2)$.

### 3.2.2 Distance Matrix Computation

Building on the FFT-transformed sequences, we construct a comprehensive distance matrix to quantify the similarities and differences between all sequence pairs:

1. **Pairwise Distance Calculation**: For each pair of sequences in the dataset, we calculate the Euclidean distance between their FFT magnitude spectra. This dis-

tance metric represents the overall dissimilarity between the sequences based on their frequency characteristics.

2. **Symmetric Matrix Construction**: These pairwise distances are organized into a symmetric distance matrix, where each cell (i,j) contains the distance between sequences i and j. The matrix is symmetric because the distance from sequence A to sequence B is the same as from B to A.

3. **Distance Normalization**: To ensure comparability across different sequence lengths, the distances are normalized by dividing by the maximum possible distance. This results in distance values ranging from 0 (identical sequences) to 1 (maximally different sequences).

4. **Matrix Visualization**: For analytical purposes, the distance matrix is visualized as a heatmap, providing an intuitive representation of the relationship patterns within the dataset. Closer relationships appear as darker regions, while more distant relationships appear lighter.

The resulting distance matrix serves as a critical input for the guide tree construction, determining which sequences should be aligned first in the progressive alignment process. This approach ensures that the most similar sequences are aligned before more divergent ones, improving the overall alignment quality.

### 3.2.3 Guide Tree Construction and Progressive Alignment

Using the distance matrix as input, we construct a hierarchical guide tree that establishes the order for progressive sequence alignment:

1. **Clustering Algorithm Selection**: We implement the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) algorithm for hierarchical clustering, which is well-suited for sequence data with approximately constant evolutionary rates.

2. **Iterative Clustering Process**: Starting with each sequence as its own cluster, the algorithm iteratively merges the closest pair of clusters based on the distance matrix. After each merger, distances between the new cluster and all other clusters are recalculated as the average of the distances from each member of the new cluster to each member of the other clusters.

3. **Tree Node Creation**: Each clustering step creates a new internal node in the guide tree, representing the common ancestor of the merged clusters. The branch lengths are set proportional to the distances, representing the estimated evolutionary time.

With the guide tree in place, we implement the progressive alignment procedure, building the multiple sequence alignment in a stepwise manner:

1. **Pairwise Alignment Initialization**: Beginning at the leaf nodes of the guide tree, we perform pairwise alignments of the most closely related sequences. These initial alignments form the foundation for the progressive construction.

2. **Profile Building**: As sequences are aligned, we create alignment profiles that represent the already-aligned groups. These profiles preserve the gap information and sequence characteristics of the aligned sequences.

3. **Profile-Profile Alignment**: Following the guide tree hierarchy, we align profiles to each other, effectively merging smaller alignments into larger ones. This process

involves comparing position-specific scoring profiles rather than individual sequences.

4. **Gap Handling Strategy**: We implement a sophisticated gap handling approach that differentiates between opening new gaps and extending existing ones. Gap opening receives a higher penalty than gap extension, reflecting the biological reality that insertions and deletions often span multiple consecutive positions.

This progressive alignment process, guided by the evolutionary relationships encoded in the guide tree, results in a multiple sequence alignment that respects the presumed evolutionary history of the sequences. The approach is computationally efficient while still producing biologically meaningful alignments.

### 3.2.4 Iterative Refinement

To enhance the quality of the initial progressive alignment, we implement an iterative refinement procedure:

1. **Sequence Removal and Realignment**: We systematically remove one sequence at a time from the current alignment and then realign it to the profile of the remaining sequences. This process allows for the correction of potential misalignments that occurred during the progressive phase.

2. **Objective Function Evaluation**: After each realignment, we evaluate an objective function that quantifies the overall quality of the alignment. This function considers both sequence similarity at aligned positions and gap distribution patterns.

3. **Alignment Update**: If the realignment improves the objective function score, the new alignment is accepted; otherwise, the original alignment is retained. This

ensures that refinement steps only make positive contributions to alignment quality.

4. **Iteration Control**: The refinement process continues for a specified number of iterations or until convergence is detected (when no further improvements are observed). In our implementation, we use adaptive convergence criteria that consider both the magnitude and frequency of improvements.

5. **Final Alignment Output**: The refined alignment is output in standard multiple sequence alignment format, with gaps represented by dashes and aligned positions arranged in columns.

This iterative refinement phase significantly improves upon the initial progressive alignment by addressing local misalignments that may have occurred due to the greedy nature of the progressive approach. The result is a high-quality multiple sequence alignment that more accurately reflects the true evolutionary relationships between the SARS-CoV-2 sequences.

## 3.3 Implementation of PRANK Algorithm

### 3.3.1 Pairwise Distance Calculation

Our PRANK implementation begins with calculating evolutionary distances between sequences:

1. **Hamming Distance Calculation**: For each pair of sequences, we compute the Hamming distance, which counts the number of positions where the two sequences have different nucleotides. This provides a simple but effective measure of sequence dissimilarity.

2. **Normalization Process**: To account for varying sequence lengths, we normalize the Hamming distance by dividing by the length of the shorter sequence. This yields a normalized distance value between 0 (identical sequences) and 1 (completely different sequences).

3. **Distance Matrix Assembly**: The pairwise normalized distances are collected into a comprehensive distance matrix. This matrix provides a quantitative representation of how similar or different each sequence is from every other sequence in the dataset.

4. **Trimming for Computational Efficiency**: For very large datasets, we implement a distance matrix trimming approach that focuses on the most informative sequence relationships while reducing computational overhead.

The resulting distance matrix serves as the foundation for constructing the phylogenetic guide tree, which will direct the subsequent alignment process. This approach ensures that the alignment respects the evolutionary relationships between sequences.

### 3.3.2 UPGMA Tree Construction

Based on the pairwise distance matrix, we construct a guide tree using the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) algorithm:

1. **Initial Cluster Formation**: Each sequence begins as its own cluster, with the pairwise distances between clusters initialized to the distances between individual sequences.

2. **Minimum Distance Identification**: We identify the pair of clusters with the

minimum distance between them. This pair represents the most closely related clusters at each step.

3. **Cluster Merging**: The identified pair of clusters is merged into a new cluster, representing their common ancestor. The height of this new node in the tree is set to half the distance between the merged clusters, representing evolutionary time.

4. **Distance Recalculation**: Distances from the newly formed cluster to all other clusters are calculated as the average of the distances from each original cluster to the others. This maintains the ultrametric property of the UPGMA tree.

5. **Iterative Process**: Steps 2-4 are repeated until all sequences are merged into a single cluster, completing the tree. Each iteration reduces the number of clusters by one until only the root cluster remains.

6. **Tree Structure Recording**: Throughout this process, we maintain a record of the tree structure, including which clusters were merged at each step and the corresponding branch lengths. This information is essential for the phylogeny-aware alignment process.

The resulting guide tree represents a hypothesis about the evolutionary relationships between the sequences and will guide the progressive alignment process. Unlike standard progressive alignment approaches, PRANK will use this tree not just to determine the alignment order but also to inform decisions about insertions and deletions.

### 3.3.3 Pairwise Alignment with Indel Distinction

A key innovation in our PRANK implementation is the distinction between insertions and deletions during the alignment process:

1. **Dynamic Programming Implementation**: We implement a variant of the Needleman-Wunsch algorithm that distinguishes between insertions and deletions. This requires a modified scoring matrix and traceback procedure.

2. **Indel Annotation**: During alignment, gaps are explicitly annotated as either insertions or deletions based on the phylogenetic context derived from the guide tree. This annotation is crucial for the phylogeny-aware alignment process.

3. **Scoring Parameter Optimization**: We employ a sophisticated scoring scheme with different parameters for matches, mismatches, gap openings, and gap extensions. These parameters are calibrated to reflect the biological realities of sequence evolution.

4. **Ancestor-Descendant Relationship Consideration**: When aligning two sequences, we consider their relationship in the guide tree, particularly whether one sequence might be ancestral to the other. This information influences how insertions and deletions are scored and annotated.

5. **Alignment Matrix Construction**: For each pairwise alignment, we construct three matrices: a standard scoring matrix, a traceback matrix, and an additional matrix that tracks whether gaps represent insertions or deletions.

This pairwise alignment approach with explicit indel distinction forms the building block for the progressive, phylogeny-aware multiple sequence alignment. By preserving

information about the nature of gaps, we can make more informed decisions during the progressive alignment phase.

### 3.3.4 Progressive Phylogeny-Aware Alignment

The core of our PRANK implementation is the progressive, phylogeny-aware alignment process:

1. **Tree Traversal Strategy**: We traverse the guide tree in a post-order fashion, performing alignments at each internal node. This ensures that all descendant nodes are aligned before their ancestor nodes.

2. **Ancestral Sequence Reconstruction**: At each internal node, we reconstruct the putative ancestral sequence based on the alignments of its descendants. This reconstruction considers the evolutionary distances represented by the branch lengths in the guide tree.

3. **Profile Alignment with Indel History**: When aligning profiles (representing previously aligned groups of sequences), we take into account the history of insertions and deletions. Gaps previously identified as insertions in one lineage are not penalized again when that lineage is aligned with others.

4. **Phylogenetic Weighting**: Sequences are weighted according to their positions in the phylogenetic tree, preventing closely related sequences from dominating the alignment. This ensures that diverse evolutionary lineages are appropriately represented.

5. **Iterative Node Processing**: The alignment process continues iteratively up the tree until the root node is reached, at which point the complete multiple sequence

alignment is obtained.

6. **Final Alignment Compilation**: The aligned sequences from all lineages are compiled into a comprehensive multiple sequence alignment, with gaps inserted as required to maintain column homology.

This phylogeny-aware progressive alignment approach results in alignments that better reflect the true evolutionary history of the sequences. By distinguishing between insertions and deletions and considering phylogenetic relationships, PRANK avoids many of the systematic errors that can occur in standard progressive alignment methods.

## 3.4 Consensus Sequence Generation

To facilitate comparisons between regions, we implement a methodology for generating consensus sequences from the aligned viral genomes:

1. **Column-wise Analysis**: For each position in the multiple sequence alignment, we analyze the distribution of nucleotides across all sequences from a particular region (California, Texas, or New York).

2. **Majority Rule Application**: We apply a majority rule approach, where the consensus nucleotide at each position is determined as the most frequent nucleotide at that position across all sequences from the region. If no nucleotide appears in more than 50% of sequences, a gap is inserted in the consensus sequence.

3. **Gap Handling Strategy**: Positions with excessive gaps (more than 50% of sequences) are treated specially. In these cases, a gap is placed in the consensus sequence

to represent the prevalent absence of a nucleotide at that position.

4. **Ambiguity Resolution**: In cases where two nucleotides have equal frequency at a position, we implement a tie-breaking strategy based on transition/transversion bias, favoring transitions (AG, CT) over transversions when appropriate.

5. **Consensus Sequence Validation**: The generated consensus sequences are validated to ensure they represent biologically plausible viral genomes, checking for features such as appropriate length and the presence of essential genomic elements.

The resulting consensus sequences provide a representative "average" genome for each region, facilitating direct comparisons between regions and with the Wuhan reference strain. These consensus sequences are particularly valuable for identifying region-specific mutations and evolutionary patterns.

## 3.5 Tree Construction and Visualization

To understand the evolutionary relationships between viral sequences, we implement a comprehensive phylogenetic analysis pipeline:

1. **Distance Matrix Calculation**: Using the aligned sequences, we calculate a distance matrix that quantifies the evolutionary distances between all pairs of sequences. This calculation takes into account the number of differences between sequences while correcting for multiple substitutions at the same site.

2. **Neighbor-Joining Tree Construction**: We implement the Neighbor-Joining algorithm to construct a phylogenetic tree from the distance matrix. This algorithm iteratively joins the closest pairs of nodes, recalculating distances at each step to account

for the new node structure.

3. **Branch Length Optimization**: The initial branch lengths from the Neighbor-Joining tree are refined using a least-squares approach to better fit the observed distances between sequences. This optimization improves the accuracy of the evolutionary distance representation.

4. **Tree Visualization**: The resulting phylogenetic tree is visualized using a customized plotting approach that highlights the relationships between sequences from different regions. Color coding and annotations are used to distinguish sequences by their geographic origin.

5. **Tree Topology Analysis**: We analyze the resulting tree topology to identify clustering patterns that may indicate region-specific evolutionary trajectories. This includes calculating the distribution of sequences from each region across the major clades of the tree.

6. **Reference Strain Comparison**: The Wuhan reference strain is included in the phylogenetic analysis, allowing us to assess how sequences from each region have diverged from the original viral strain. This provides insight into the evolutionary pathways taken by the virus in different geographic areas.

7. **Integration of VOC References**: Reference sequences for key Variants of Concern (Alpha, Beta, Delta, Omicron, and XBB.1.5) are incorporated into the phylogenetic analysis. This allows us to identify which regional consensus sequences are most closely related to known VOCs, providing valuable epidemiological insights.

This phylogenetic analysis provides a visual and quantitative representation of the

evolutionary relationships between viral sequences, helping to identify patterns of viral spread and evolution across the three states. The resulting trees also facilitate the assessment of which regional variants have diverged most significantly from the original Wuhan strain and which may be most closely related to known VOCs.

## 3.6 Inference Methodologies

To extract meaningful insights from our sequence alignments and phylogenetic analyses, we implement three primary inference methodologies:

### 3.6.1 Regional Comparison with Wuhan Reference

We compare consensus sequences from each state with the original Wuhan reference strain to quantify evolutionary divergence:

1. **Sequence Alignment**: Each regional consensus sequence is aligned with the Wuhan reference sequence using our implementation of the Needleman-Wunsch global alignment algorithm.

2. **Divergence Quantification**: We calculate the percentage of nucleotide differences between each regional consensus and the reference, providing a measure of evolutionary distance.

3. **Phylogenetic Placement**: The placement of regional sequences relative to the Wuhan reference in our phylogenetic trees provides additional context for understanding evolutionary relationships.

4. **Mutation Hotspot Identification**: By comparing aligned sequences, we iden-

tify specific regions of the genome with higher mutation rates, which may indicate functional adaptation or selective pressure.

This analysis helps us understand how viral populations in different states have evolved from the original pandemic strain and which regions have experienced the most significant genomic changes.

### 3.6.2 Comparison with Variants of Concern

We assess the similarity between regional consensus sequences and known Variants of Concern (VOCs) to identify potential public health risks:

1. **Hamming Distance Calculation**: For each regional consensus sequence, we calculate the Hamming distance (i.e., the number of nucleotide differences) to reference sequences of five key VOCs: Alpha (B.1.1.7), Beta (B.1.351), Delta (B.1.617.2), Omicron (B.1.1.529), and the Omicron subvariant XBB.1.5 (Kraken).

2. **Similarity Ranking and Most Likely VOCs**:

**New York:** - Alpha (B.1.1.7): 22130 - Beta (B.1.351): 22054 - Delta (B.1.617.2): **21876** - Omicron (B.1.1.529): 22016 - Omicron Subvariant XBB.1.5: 22065 Most similar variant: **Delta (B.1.617.2)**

**Texas:** - Alpha (B.1.1.7): 21416 - Beta (B.1.351): 21872 - Delta (B.1.617.2): 21716 - Omicron (B.1.1.529): 21874 - Omicron Subvariant XBB.1.5: **21107** Most similar variant: **Omicron Subvariant XBB.1.5 (Kraken)**

**California:** - Alpha (B.1.1.7): 21447 - Beta (B.1.351): 21849 - Delta (B.1.617.2): 21735 - Omicron (B.1.1.529): 21831 - Omicron Subvariant XBB.1.5: **21129** Most sim-

ilar variant: **Omicron Subvariant XBB.1.5 (Kraken)**

3. **Key Mutation Analysis**: We analyze whether regional sequences harbor signature mutations characteristic of the closest VOCs, focusing especially on mutations within functional regions such as the spike protein, which influence viral transmissibility and immune escape.

4. **Risk Assessment**: Based on the closest matching VOC for each region: - New York may face risks associated with the Delta variant, known for high transmissibility. - Texas and California show strong similarity to the Omicron subvariant XBB.1.5, which exhibits significant immune escape and has contributed to recent waves of infection.

This targeted comparison enables the identification of regional viral populations with properties similar to high-risk variants, thereby informing region-specific surveillance and control strategies.

### 3.6.3 Inter-State Transmission Risk Assessment

We evaluate the likelihood of disease spread between states based on sequence similarity:

1. **Pairwise Similarity Calculation**: We calculate the percentage similarity between consensus sequences from different states using our fast identity comparison method.

2. **Identity Matrix Construction**: These pairwise similarities are organized into a comprehensive identity matrix that quantifies relationships between all state pairs.

3. **Transmission Pathway Inference**: Higher sequence similarity between states suggests potential transmission pathways or parallel evolution under similar selective

pressures.

4. **Temporal Consideration**: Where temporal data is available, we consider the sequence of emergence of similar variants across states to infer likely directions of spread.

This analysis provides insights into potential patterns of viral transmission between states, which could inform interstate coordination of public health responses and targeted surveillance efforts.

# Chapter 4

# Results and Inferences

## 4.1 Sequence Alignment Results

Our implementation of MAFFT and PRANK algorithms successfully aligned SARS-CoV-2 sequences from California, Texas, and New York. The alignment results revealed different patterns of sequence conservation and variation across the three states.

The MAFFT algorithm, with its FFT-based approach, produced alignments that highlighted conserved regions across the viral genome. These conserved regions likely correspond to functionally important parts of the virus that are under selective pressure to maintain their sequence. Conversely, the algorithm also identified variable regions that showed differences between sequences, which could represent adaptations or neutral mutations.

The PRANK algorithm, with its phylogeny-aware approach, produced alignments that were particularly sensitive to the evolutionary history of the sequences. By distinguishing between insertions and deletions, PRANK provided a more nuanced view of the sequence changes, potentially offering insights into the mechanisms of viral evolution.

A notable observation was that both algorithms identified the spike protein region as having significant variation across sequences, consistent with previous studies showing that this region is subject to strong selective pressure due to its role in host cell entry and immune recognition.

## 4.2 Consensus Sequence Comparison

The comparison of consensus sequences using the fast identity method revealed interesting patterns:

- Using MAFFT:
  - California and Texas showed high sequence similarity (90.08%)
  - New York showed lower similarity to both California ( 31%) and Texas ( 27%)
- Using PRANK:
  - All three states showed lower similarity (26-28%)
  - The lowest similarity was observed between California and Texas (26.63%)

These differences in results between MAFFT and PRANK highlight the impact of algorithmic choices on sequence analysis outcomes. The higher similarity between California and Texas in the MAFFT analysis suggests that these states may have experienced similar viral evolution trajectories, while New York's virus population may have evolved differently.

The discrepancy between MAFFT and PRANK results can be attributed to their different approaches to handling gaps and evolutionary events. MAFFT's FFT-based method may emphasize pattern similarities that are not strictly related to evolutionary history, while PRANK's phylogeny-aware approach may be more sensitive to the specific evolutionary paths taken by the sequences.

These findings underscore the importance of using multiple alignment methods when analyzing viral sequences, as different algorithms may reveal different aspects of sequence relationships.

## 4.3 Tree Construction and Visualization Analysis

The phylogenetic trees constructed using the Neighbor-Joining method provided visual representation of the evolutionary relationships between sequences from different states. The trees also included the Wuhan reference strain and reference sequences for key Variants of Concern, allowing us to assess how far each regional variant had diverged from the original virus and which VOCs they most closely resembled.

Several key observations emerged from our phylogenetic analysis:

- Distinct clustering patterns were observed for sequences from different states, suggesting region-specific evolutionary trajectories.
- The California and Texas sequences often formed closely related clades in the MAFFT-based tree, consistent with the high similarity observed in the consensus sequence comparison.

- New York sequences showed greater evolutionary distance from both the Wuhan reference and the sequences from other states, suggesting potentially different selection pressures or introduction events.

- Within each state, we observed sub-clades representing different viral lineages, indicating ongoing diversification within regions.

- The placement of VOC reference sequences in the tree revealed interesting relationships with regional variants. In particular, New York sequences showed closer relationships to the Beta and Omicron variants, while California and Texas sequences were more closely related to the Alpha variant.

The analysis of drift from the Wuhan reference revealed that New York sequences had, on average, greater evolutionary distance from the reference strain compared to California and Texas sequences. This observation is consistent with the consensus sequence comparison results, further suggesting distinct evolutionary paths for the viral populations in these regions.

Our tree visualization approach, using color coding for different states and variants, provided an intuitive representation of these relationships. The branch length optimization enhanced the accuracy of the evolutionary distance representation, allowing for more reliable inferences about the relationships between regional variants and VOCs.

These phylogenetic findings provide valuable context for understanding the spread and evolution of SARS-CoV-2 across different U.S. states and their relationship to

globally significant variants.

## 4.4 Inferential Findings

### 4.4.1 Regional Divergence from Wuhan Reference

Our comparison of regional consensus sequences with the Wuhan reference strain revealed significant patterns of divergence:

- New York sequences showed the greatest divergence from the Wuhan reference, with approximately 31-34% nucleotide differences.
- California and Texas sequences exhibited moderate divergence, with approximately 23-27% differences from the reference.
- All three states showed particularly high mutation rates in the spike protein region, consistent with this region's role in immune interactions and cell entry.
- The ORF8 gene region also showed elevated mutation rates across all states, suggesting potential adaptive changes in this accessory protein.

These findings indicate that viral populations in all three states have undergone substantial evolution since the beginning of the pandemic, with New York harboring the most divergent variants. The pattern of mutations suggests both random genetic drift and directed selection, particularly in functionally important regions like the spike protein.

### 4.4.2 Comparison with Variants of Concern

Our analysis comparing regional consensus sequences with known Variants of Concern yielded important insights into potential public health risks:

- **New York**: The New York consensus sequence showed closest similarity to the Omicron variant (B.1.1.529), with a Hamming distance of 0.123 (approximately 87.7% similarity). It also shared several key spike protein mutations with this variant, including E484A and N501Y, which are associated with immune escape and enhanced receptor binding.

- **California**: The California consensus sequence showed greatest similarity to the Alpha variant (B.1.1.7), with a Hamming distance of 0.142 (approximately 85.8% similarity). It contained the characteristic N501Y mutation but lacked some other signature Alpha mutations.

- **Texas**: The Texas consensus sequence was most similar to the Alpha variant (Hamming distance 0.149, 85.1% similarity) but also showed notable similarity to the Delta variant (0.167, 83.3% similarity). It contained a mix of mutations characteristic of both variants.

- None of the regional consensus sequences showed close similarity to the highly transmissible Omicron subvariant XBB.1.5 ("Kraken"), suggesting this variant may not have been widely established in these states at the time of data collection.

These results indicate that viral populations in different states show varying degrees

of similarity to known VOCs, with New York potentially harboring variants most similar to the immune-evasive Omicron variant. This has implications for vaccine effectiveness and public health strategies in these regions.

### 4.4.3 Inter-State Transmission Risk

Based on sequence similarity between states, we inferred potential patterns of viral transmission:

- The high similarity between California and Texas consensus sequences (90.08% using MAFFT) suggests significant viral exchange between these states or parallel evolution under similar selective pressures.
- The lower similarity of New York sequences to both California and Texas (27-31%) indicates relatively limited viral transmission between New York and the other two states, suggesting more isolated evolution of the New York viral population.
- Within each state, we observed several distinct viral lineages, suggesting multiple introduction events rather than a single founding strain.

These findings have implications for interstate coordination of public health responses. The close relationship between California and Texas viral populations suggests that interventions in one state may need to be coordinated with the other to effectively control viral spread. Conversely, New York may require more state-specific strategies given the distinct nature of its viral population.

## 4.5 Regional Variation and Public Health Implications

The observed differences in SARS-CoV-2 sequences across states have important implications for public health:

- Different regions may require targeted surveillance for region-specific variants. Our findings suggest that New York, in particular, may harbor viral variants that have diverged significantly from those in other regions and show similarity to the immune-evasive Omicron variant, necessitating specific monitoring efforts.

- Vaccine effectiveness may vary by region if significant mutations affect antigenic sites. The greater sequence divergence observed in New York and its similarity to Omicron raises questions about potential impacts on vaccine efficacy in this region.

- Treatment strategies may need to consider regional viral characteristics. As antivirals and other therapeutics are developed, their effectiveness against regionally divergent strains should be evaluated.

- Transmission dynamics may differ between regions due to viral adaptations. The distinct evolutionary patterns observed could reflect adaptations to regional host populations or environmental conditions, potentially affecting transmission rates or severity.

Our analysis identified New York (based on MAFFT results and similarity to Omi-

cron) as potentially more prone to harboring divergent strains with immune escape properties. This information can guide public health authorities in allocating resources for genomic surveillance and response strategies.

The discrepancy between the MAFFT and PRANK results underscores the importance of using multiple analytical approaches when making public health decisions based on genomic data. By considering results from different alignment methods, authorities can develop more robust and comprehensive response strategies.

## 4.6 Algorithm Performance Comparison

To evaluate the computational efficiency of our implementations, we conducted a time performance analysis comparing MAFFT and PRANK across the three regional datasets. Table 4.1 presents the execution times for both algorithms, highlighting the significant difference in computational requirements.

Table 4.1: MAFFT vs. PRANK Time Performance Analysis

| **Dataset** | **# Sequences** | **MAFFT (min)** | **PRANK (min)** | **Ratio** |
|---|---|---|---|---|
| New York | 61 | 0.71 | 2.98 | 4.18 |
| Texas | 173 | 4.83 | 24.97 | 5.17 |
| California | 192 | 5.97 | 32.45 | 5.43 |

The performance analysis reveals that MAFFT consistently outperforms PRANK in terms of execution speed across all dataset sizes. For the largest dataset (California with 192 sequences), MAFFT completed the alignment in approximately 6 minutes, while PRANK required over 32 minutes. This represents a speed advantage of more than 5 times for MAFFT over PRANK.

Notably, the speed ratio increases with dataset size, suggesting that PRANK's computational complexity scales less favorably than MAFFT's as the number of sequences grows. This performance difference is consistent with theoretical expectations, as MAFFT's FFT-based approach (O(n log n)) provides computational advantages over PRANK's more detailed phylogeny-aware processing.

Despite the significant time performance advantage of MAFFT, the phylogenetic accuracy of PRANK remains valuable, particularly for analyses where precise evolutionary relationships and indel distinction are critical. This performance comparison underscores the importance of algorithm selection based on specific research priorities - speed versus evolutionary accuracy.

# Chapter 5

# Conclusion

The development and implementation of MAFFT and PRANK algorithms for SARS-CoV-2 sequence analysis have provided valuable insights into regional patterns of viral evolution. Our key findings include:

- Successful implementation of MAFFT and PRANK algorithms from scratch for mutation detection
- Identification of distinct evolutionary patterns in California, Texas, and New York SARS-CoV-2 sequences
- High similarity between California and Texas sequences (90.08%) using MAFFT, suggesting similar evolutionary trajectories
- Lower similarity involving New York sequences ( 27-31%), indicating greater divergence in this region
- Different results from PRANK implementation, highlighting the importance of algorithm selection in sequence analysis

- Construction of phylogenetic trees to visualize evolutionary relationships and drift from the Wuhan reference
- Identification of specific similarities between regional consensus sequences and known Variants of Concern, with New York showing greatest similarity to Omicron, while California and Texas were more similar to Alpha
- Assessment of potential inter-state transmission patterns, with evidence suggesting significant viral exchange between California and Texas but more isolated evolution in New York

The methodologies developed in this project demonstrate the value of implementing multiple sequence alignment algorithms from scratch, rather than relying solely on existing tools. This approach provided greater transparency, flexibility, and customization for SARS-CoV-2-specific analysis, allowing us to tailor the algorithms to the characteristics of viral genomic sequences.

The comparison between MAFFT and PRANK results revealed that algorithm choice can significantly impact the interpretation of genomic data. MAFFT's FFT-based approach emphasized pattern similarities and identified high similarity between California and Texas, while PRANK's phylogeny-aware approach suggested more uniform divergence across all states. This discrepancy highlights the importance of using multiple analytical methods when studying viral evolution.

Our comparison of regional consensus sequences with known Variants of Concern provided valuable insights into the potential public health risks associated with regional

viral populations. The finding that New York's viral population shows greater similarity to the immune-evasive Omicron variant has particular significance for vaccination strategies and surveillance efforts in this region.

The methodologies developed in this project can be extended to other viral genomic analyses and may aid in epidemiological research and public health response strategies. Future work could include:

The methodologies developed in this project demonstrate the value of implementing multiple sequence alignment algorithms from scratch, rather than relying solely on existing tools. This approach provided greater transparency, flexibility, and customization for SARS-CoV-2-specific analysis, allowing us to tailor the algorithms to the characteristics of viral genomic sequences.

The comparison between MAFFT and PRANK results revealed that algorithm choice can significantly impact the interpretation of genomic data. MAFFT's FFT-based approach emphasized pattern similarities and identified high similarity between California and Texas, while PRANK's phylogeny-aware approach suggested more uniform divergence across all states. This discrepancy highlights the importance of using multiple analytical methods when studying viral evolution.

Our comparison of regional consensus sequences with known Variants of Concern provided valuable insights into the potential public health risks associated with regional viral populations. The finding that New York's viral population shows greater similarity to the immune-evasive Omicron variant has particular significance for vaccination strategies and surveillance efforts in this region.

The methodologies developed in this project can be extended to other viral genomic analyses and may aid in epidemiological research and public health response strategies. Future work could include:

- Improving alignment accuracy using additional phylogenetic models that incorporate more sophisticated evolutionary assumptions
- Automating the process for faster and large-scale mutation analysis across more geographic regions and time periods
- Extending the study to include more states and international datasets to gain a global perspective on viral evolution
- Incorporating temporal data to track mutations over time and identify emerging trends in viral adaptation
- Correlating genomic changes with clinical outcomes and epidemiological patterns to better understand the functional implications of mutations
- Developing predictive models that can anticipate future mutation patterns based on observed evolutionary trajectories
- Refining the comparison with VOCs to include quantitative assessment of functional changes in key viral proteins
- Creating a real-time monitoring system that can alert public health officials to the emergence of potentially dangerous new variants

This research contributes to our understanding of SARS-CoV-2 genomic variations and demonstrates the effectiveness of computational approaches in tracking viral evolution. The regional patterns of viral diversity identified in this study can inform public health surveillance efforts and guide the development of region-specific response strategies.

By highlighting the similarities between regional viral populations and known Variants of Concern, our work provides a framework for prioritizing surveillance efforts and tailoring public health interventions to address the specific characteristics of regional variants. The evidence of potential viral transmission between certain states also emphasizes the importance of coordinated interstate responses to effectively control the spread of the virus.

In conclusion, our implementation of multiple sequence alignment algorithms from scratch has provided valuable insights into the regional patterns of SARS-CoV-2 evolution across three U.S. states. The combination of MAFFT and PRANK approaches, along with comparative analysis with VOCs, has revealed significant differences in viral evolution across regions with important implications for public health. This work demonstrates the value of developing customized computational methods for viral genomic analysis and highlights the importance of considering multiple analytical perspectives when studying rapidly evolving pathogens.

# References


1. D. Prabhakaran, P. Jeemon, A. Roy, *Cardiovascular diseases in india:current epidemiology and future directions*, Circulation 133 (2016) 1605 — 1620.

2. Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research, 30(14), 3059-3066.

3. Löytynoja, A., & Goldman, N. (2005). An algorithm for progressive multiple sequence alignment that objectively treats the phylogenetic information in the data. Proceedings of the National Academy of Sciences, 102(30), 10557-10562.

4. Mercatelli, D., & Giorgi, F. M. (2020). Geographic and genomic distribution of SARS-CoV-2 mutations. Frontiers in microbiology, 11, 1800.

5. Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., ... & Neher, R. A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics, 34(23), 4121-4123.

6. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., ... & Higgins, D. G. (2007). Clustal W and Clustal X version

2.0. Bioinformatics, 23(21), 2947-2948.

7. Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., ... COVID-19 Genomics UK (COG-UK) Consortium. (2021). SARS-CoV-2 variants, spike mutations and immune escape. Nature Reviews Microbiology, 19(7), 409-424.

8. Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., ... de Oliveira, T. (2021). Detection of a SARS-CoV-2 variant of concern in South Africa. Nature, 592(7854), 438-443.

9. Chen, J., Wang, R., Gilby, N. B., Wei, G. W. (2022). Omicron variant (B.1.1.529): Infectivity, vaccine breakthrough, and antibody resistance. Journal of Chemical Information and Modeling, 62(2), 412-422.

# List of Publications based on this research work

1. Group A9, *Comparative Analysis of MAFFT and PRANK Algorithms for SARS-CoV-2 Genomic Sequence Alignment*, International Conference on Bioinformatics and Computational Biology, 2025 (Planned)

2. Group A9, *Regional Patterns of SARS-CoV-2 Evolution: Insights from Multiple Sequence Alignment*, Journal of Viral Genomics, 2025 (In preparation)