Department of Artificial Intelligence

22BIO201: Intelligence of Biological Systems - 1, Fall 2024

Project Report

GENOMIC ANALYSIS & CHAOS GAME REPRESENTATION

Team Members:

Abhishek Pandey: CB.SC.U4AIE23205
Adith Krishna: CB.SC.U4AIE23202
Aditya Santosh: CB.SC.U4AIE23207
Arjun Gopal: CB.SC.U4AIE23271

Supervised By:

Kalaivani. S. S
Assistant Professors
Department of Artificial Intelligence
Amrita Vishwa Vidyapeetham

Date of Submission: 15/11/2024

Signature of the Project Supervisor:

Amrita Vishwa Vidyapeetham

Coimbatore

CERTIFICATE

This is to certify that we, the student of Amrita School of Artificial Intelligence, has completed the "***Genomic Analysis"*** as part of their ***Chaos Game Representation in*** Introduction to Python.

The project work was carried out under the guidance and supervision of Kalaivani. S. S, Assistant Professor, Amrita School of Artificial Intelligence, Coimbatore. To the best of our knowledge this work has not formed the basis for the aware of any degree/diploma/ associate ship/fellowship/ or a similar award to any candidate in any University.

Date:

Kalaivani. S. S

Amrita School of Artificial Intelligence

Amrita Vishwa Vidyapeetham

Coimbatore

Acknowledgement

We would like to make this an opportunity to transfuse our appreciation to everyone who was behind the successful completion of this design. First and foremost, we would like to thank "Department of Artificial Intelligence" for accepting our project.

Kindest regard goes to **Prof. Soman KP**, Head of the Department of Artificial Intelligence for allowing us to accomplish the design project.

We would like to convey appreciation to our supervisor**. Kalaivani. S. S** for her encouragement and motivation to expand our vision regarding proper process designing. We are very thankful for her help and supervision. This task would have been little success without their proper guidance and support.

Last but not the least, we are thankful to our group members and friends of our department for their friendly support given to accomplish this project successfully and our family members for always providing the best guidance possible.

Abstract

The goal of this project is to investigate the genetic relationships and possible evolutionary lineage between SARS-CoV-2 and Bat Beta coronavirus RaTG13 using a comparative genomic analysis. Similarities between SARS-CoV-2, the virus that caused the COVID-19 pandemic, and coronaviruses in bats, especially RaTG13, may provide information about its zoonotic origin. The study evaluates the structure and function of the SARS-CoV-2 genome using a variety of bioinformatics tools, including as GC skew, clump detection, K-Mer frequency analysis, and Chaos Game Representation (CGR).

K-Mer analysis identifies conserved sequences essential to the virus's operation or often occurring sequence patterns that could be hotspots for mutation. While GC skew reveals nucleotide biases, such as the replication origin and terminus, which are important for comprehending viral replication, cluster analysis draws attention to regions of high-frequency repeats. Through this approach, the project establishes a bioinformatics framework that can be used for studying other zoonotic viruses with pandemic potential, contributing to foundational research in virology. Findings from this study may inform future strategies for pathogen detection, prevention, and treatment by offering a deeper understanding of viral evolution and genetic makeup.

1. Introduction

1.1 General

An important family of viruses known as coronaviruses can cause respiratory diseases in both humans and animals. Some of these viruses can spread from animals to humans, a phenomenon known as zoonotic transmission. Among these, the virus that caused the COVID-19 pandemic, SARS-CoV-2, has genetic resemblances to the Bat Beta coronavirus RaTG13, which raises concerns over its zoonotic evolution and origin. Determining SARS-CoV-2's evolutionary trajectory, possible mutations, and adaptive mechanisms—all of which have consequences for global health preparedness and response tactics—requires an understanding of these genetic connections.

K-Mer frequency, clump analysis, GC skew, and Chaos Game Representation (CGR) are some of the genomic analytic methods used in this effort to uncover the genetic characteristics of SARS-CoV-2 and its resemblances to Bat Beta coronavirus RaTG13. By dividing a genome into substrings of length k, a technique known as K-Mer analysis, we might identify recurrent sequence motifs that can point to areas of evolutionary importance. Researchers can find conserved genetic patterns between SARS-CoV-2 and RaTG13 by comparing their most common K-Mers. These patterns may indicate ancestral sequences or shared functional elements.

This comparison is further enhanced by cluster analysis, which finds repeating K-Mers in particular genomic areas. These "clumps" indicate places with higher activity or structural significance in viral replication by highlighting times when genetic sequences repeat often in close proximity. GC skew, which quantifies the asymmetry between guanine (G) and cytosine (C) nucleotides throughout the genome, is another effective tool in genomic analysis. Because it draws attention to nucleotide bias that might affect transcription and replication, this metric is helpful in locating important genomic landmarks like the terminus and the origin of replication (oriC).

Lastly, Chaos Game Representation (CGR) provides a fractal-based visualization of DNA sequences, mapping each nucleotide to specific coordinates. Through CGR, we can observe the structural organization of SARS-CoV-2's genome, creating a graphical fingerprint that helps detect structural similarities and differences with RaTG13.

1.2 Scope of Project

- ***Comparative Genomic Analysis:*** Use K-Mer frequency, clump analysis, GC skew, and CGR to compare the genomic structures of SARS-CoV-2 and Bat Beta coronavirus RaTG13, identifying genetic similarities and differences.
- ***Evolutionary Lineage Identification:*** Trace the evolutionary connections between SARS-CoV-2 and related coronaviruses, focusing on understanding zoonotic origins and genetic adaptations.
- ***Detection of Genetic Hotspots:*** Identify high-frequency K-mers and clump regions within the SARS-CoV-2 genome that may reveal mutations, conserved sequences, and functional elements.
- ***Replication Origin and Terminus Mapping:*** Utilize GC skew analysis to locate the origin and terminus of replication, aiding in the understanding of viral replication mechanisms.
- ***Graphical Representation of Genome:*** Employ Chaos Game Representation (CGR) to create a visual map of SARS-CoV-2's nucleotide patterns, aiding in the structural analysis of the genome.
- ***Application to Other Viral Pathogens:*** Establish a bioinformatics framework that can be applied to study the genomes of other viruses with pandemic potential, supporting future pathogen research.

2. Overview

This project investigates the genetic relationship between SARS-CoV-2, the virus responsible for COVID-19, and Bat Betacoronavirus RaTG13 to understand potential evolutionary connections and zoonotic origins. Given the genetic similarities between these two coronaviruses, this study aims to provide a comprehensive analysis of their genomes to identify common features, mutations, and evolutionary markers that may contribute to cross-species transmission.

Using bioinformatics techniques like K-mer analysis, GC skew, and Chaos Game Representation (CGR), the study delves into structural patterns within the genomes, identifying nucleotide motifs, clumps, and unique sequences that characterize each virus. The K-mer analysis focuses on short nucleotide sequences that appear frequently in the genomes, highlighting conserved elements and mutation-prone regions. GC skew analysis aids in identifying replication origins and the terminus within the genomes, providing insights into replication mechanisms crucial to viral survival. CGR offers a visual representation of the genomes, mapping nucleotide distribution patterns and revealing unique fractal structures associated with viral adaptability.

This analysis enhances our understanding of SARS-CoV-2's evolutionary path and its genetic proximity to Bat Betacoronavirus RaTG13, with a 96% similarity in certain regions. The findings contribute valuable insights into the zoonotic potential of coronaviruses, informing both future research on viral evolution and the development of prevention and treatment strategies. The project establishes a framework for comparative genomic studies of emerging zoonotic viruses, which is critical for advancing pathogen detection, monitoring, and control efforts.

3. Literature Review

Chaos game representation and its applications in bioinformatics - Chaos game representation (CGR), a milestone in graphical bioinformatics, has become a powerful tool regarding alignment-free sequence comparison and feature encoding for machine learning. The algorithm maps a sequence to 2-dimensional space, while an extension of the CGR, the so-called frequency matrix representation (FCGR), transforms sequences of different lengths into equal-sized images or matrices. The CGR is a generalized Markov chain and includes various properties, which allow a unique representation of a sequence. Therefore, it has a broad spectrum of applications in bioinformatics, such as sequence comparison and phylogenetic analysis and as an encoding of sequences for machine learning - *Hannah Franziska Löchel, Dominik Heider [1]. Chaos Game Representation* - The chaos game representation (CGR) is an interesting method to visualize one-dimensional sequences. In this paper, we show how to construct a chaos game representation. The applications mentioned here are biological, in which CGR was able to uncover patterns in DNA or proteins that were previously unknown. We also show how CGR might be introduced in the classroom, either in a modelling course or in a dynamical systems course. Some sequences that are tested are taken from the Online Encyclopaedia of Integer Sequences, and others are taken from sequences that arose mainly from a course in experimental mathematics - *Eunice Y. S. Chan, Robert M. Corless [2]. Use of 3D chaos game representation to quantify DNA sequence similarity with applications for hierarchical clustering* - A 3D chaos game is shown to be a useful way for encoding DNA sequences. Since matching subsequence's in DNA converge in space in 3D chaos game encoding, a DNA sequences 3D chaos game representation can be used to compare DNA sequences without prior alignment and without truncating or padding any of the sequences. Two proposed methods inspired by shape-similarity comparison techniques show that this form of encoding can perform as well as alignment-based techniques for building phylogenetic trees.

The first method uses the volume overlap of intersecting spheres and the second uses shape signatures by summarizing the coordinates, oriented angles, and oriented distances of the 3D chaos game trajectory. The methods are tested using: (1) the first exon of the beta-globin gene for 11 species, (2) mitochondrial DNA from four groups of primates, and (3) a set of synthetic DNA sequences. Simulations show that the proposed methods produce distances that reflect the number of mutation events; additionally, on average, distances resulting from deletion mutations are comparable to those produced by substitution mutations - *Stephanie Young, Jerome Gilles [3]. Chaos Game Representation and its applications in Bioinformatics* - Chaos game representation (CGR), a milestone in graphical bioinformatics, has become a powerful tool regarding alignment-free sequence comparison and feature encoding for machine learning. The algorithm maps a sequence to 2-dimensional space, while an extension of the CGR, the so-called frequency matrix representation (FCGR), transforms sequences of different lengths into equal-sized images or matrices. The CGR is a generalized Markov chain and includes various properties, which allow a unique representation of a sequence. Therefore, it has a broad spectrum of applications in bioinformatics, such as sequence comparison and phylogenetic analysis and as an encoding of sequences for machine learning. This review introduces the construction of CGRs and FCGRs, their applications on DNA and proteins, and gives an overview of recent applications and progress in bioinformatics - *Hannah F Löchel, Dominik Heider [4]. Chaos game representation of gene structure* - This paper presents a new method for representing DNA sequences. It permits the representation and investigation of patterns in sequences, visually revealing previously unknown structures. Based on a technique from chaotic dynamics, the method produces a picture of a gene sequence which displays both local and global patterns. The pictures have a complex structure which varies depending on the sequence. The method is termed Chaos Game Representation (CGR). CGR raises a new set of questions about the structure of DNA sequences, and is a new tool for investigating gene structure - *H. Joel Jeffrey Nucleic Acids Research, Volume 18, Issue 8, 25 April 1990, Pages 2163–2170 [5].*

Characteristics of SARS-CoV-2 and COVID-19 - severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a highly transmissible and pathogenic coronavirus that emerged in late 2019 and has caused a pandemic of acute respiratory disease, named 'coronavirus disease 2019' (COVID-19), which threatens human health and public safety. In this Review, we describe the basic virology of SARS-CoV-2, including genomic characteristics and receptor use, highlighting its key difference from previously known coronaviruses. We summarize current knowledge of clinical, epidemiological and pathological features of COVID-19, as well as recent progress in animal models and antiviral treatment approaches for SARS-CoV-2 infection. We also discuss the potential wildlife hosts and zoonotic origin of this emerging virus in detail - *Ben Hu, Hua Guo, Peng Zhou & Zheng-Li Shi [6]. SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and Furin-cleavage effects - Coronaviruses infect a range of mammalian and avian species1. SARS-CoV-2, the agent of* *the COVID-19 pandemic2,3, belongs to the Sarbecovirus subgenus of beta coronaviruses, members of which mostly infect bats4,5. Hence, bat coronaviruses were identified as a likely evolutionary precursor of SARS-CoV-2, and the bat virus RaTG13 was identified as the closest known relative of SARS-CoV-2. It is not known how SARS-CoV-2 evolved to infect humans, but two mechanisms have been hypothesized: selection in an animal host before zoonotic transfer (possibly via an intermediate host), or natural selection in humans following direct zoonotic transmission from bats.*

The S protein of SARS-CoV-2 mediates attachment of the virus to cell-surface receptors and fusion between virus and cell membranes1. The receptor for SARS-CoV-2, like that for SARS-CoV9,10, is the human cell-surface-membrane protein angiotensin-converting enzyme 2 (ACE2). - Antoni G. Wrobel, Donald J. Benton, Pengqi Xu, Chloë Roustan, Stephen R. Martin, Peter B. Rosenthal, John J. Skehel & Steven J. Gamblin [7]. Binding and molecular basis of the bat coronavirus RaTG13 virus to ACE2 in humans and other species - Emerging and re-emerging pathogens threaten global public health and cause tremendous economic losses (Gao, 2018). The new coronavirus (CoV) severe acute respiratory syndrome CoV 2 (SARS-CoV-2) was detected and then isolated in early 2020 (Tan et al., 2020; The 2019-nCoV Outbreak Joint Field Epidemiology Investigation Team and Li, 2020; Wang et al., 2020a; Zhu et al., 2020). SARS-CoV-2 later spread worldwide, causing a global pandemic. As of March 5, 2021, the World Health Organization (WHO; https://covid19.who.int/) has recorded more than 114 million confirmed cases of CoV disease 2019 (COVID-19) globally and more than 2.5 million related deaths *[8].*

4. Methodology:

4.1 Data Collection and Sources

This study utilizes high-quality genomic data to investigate sequence similarities and structural characteristics between the SARS-CoV-2 genome and a similar bat coronavirus genome (RaTG13) by analyzing their k-mer distribution and Chaos Game Representation (CGR) patterns. The dataset consists of nucleotide sequences sourced from reputable genomic databases, ensuring accuracy and consistency in the information gathered for k-mer and CGR analysis. Key features collected from the dataset include:

1) Nucleotide Sequences: Each genome sequence is comprised of four nucleotide bases (adenine, thymine, cytosine, and guanine), whose distributions are essential for CGR and k-mer pattern analysis. Complete sequences are required for generating CGR plots and identifying patterns in k-mer occurrences.

2) Genome Length: The total length of each genome provides critical context for k-mer analysis, as the frequency and clumping of k-mers depend on the sequence length. Variations in length can impact the visual and quantitative comparison of CGR plots, which is accounted for in analytical approaches.

3) Mutation Regions: Certain regions in the SARS-CoV-2 genome contain mutations that differentiate it from other similar viruses. These mutation hotspots are tracked as they can correlate with shifts in clumping patterns or distinctive features in the CGR plots, potentially indicating evolutionary divergences between the genomes.

4) Organism Metadata: Information regarding the organism and genomic region aids in contextualizing patterns observed in CGR and k-mer analysis, providing insights into evolutionary relationships and functional similarities between the viral genomes.

4.2 Chaos Game Representation (CGR)

Chaos Game Representation (CGR) is a mathematical technique used to transform sequences of nucleotides into a two-dimensional plot that visually represents the distribution of k-mers, allowing for the detection of recurring patterns and clumping within the genome. The steps to create and analyze CGR plots are as follows:

1) Initial Positioning and Coordinate Mapping:

In CGR, each nucleotide (A, T, C, G) is assigned a unique position within a square: A at (0,0), T at (0,1), C at (1,0), and G at (1,1). The starting point of the CGR plot is set at the center of this unit square (0.5, 0.5), from which each successive point in the sequence is plotted.

2) Iterative Position Calculation:

For each nucleotide in the sequence, the current position moves halfway towards the corner assigned to that nucleotide, creating a fractal pattern characteristic of the genome sequence. For instance, if the current point is at (x, y) and the next nucleotide is A, the new point will be located at halfway between the center and the A corner. This iterative process continues for the entire length of the sequence, generating a visual distribution of all nucleotide combinations.

3) k-mer Representation in CGR:

The CGR plot inherently visualizes k-mer patterns where k indicates the sequence length of repeated units. Shorter k-mers, such as 1-mers or 2-mers, occupy broader regions of the plot, while higher-order k-mers (e.g., 4-mers or 5-mers) concentrate in specific regions. The density and clustering of points in certain regions of the CGR plot reflect the frequency of specific k-mer sequences, allowing researchers to visually identify repetitive motifs or clumping within the genome.

4) Resolution Setting and Analysis:

Adjusting the resolution of the CGR plot can reveal or obscure finer details. Higher resolutions capture subtle clumping patterns and are useful for in-depth analysis of longer k-mers, while lower resolutions can simplify the visualization, highlighting general sequence trends. This study set the CGR matrix resolution to optimize computational efficiency and ensure distinct visualization of common k-mers.

5) Visual Interpretation of CGR Patterns:

CGR plots generate unique fractal patterns that can be visually analyzed to identify nucleotide distribution and clumping behavior across the genome. Dense clusters in the CGR plot suggest regions of repeated nucleotide sequences, while sparse areas indicate less common k-mers. For example, a concentration of points near one corner could signify frequent occurrences of k-mers rich in a particular nucleotide, aiding in identifying structural or functional motifs in the genome.

4.3 K-mer Clumping Analysis

The k-mer clumping analysis focuses on identifying patterns in which specific sequences of length k occur repeatedly in close proximity within the genome. This approach helps reveal structural motifs and repetitive elements that may contribute to the genomic organization and function.

1) K-mer Extraction and Frequency Calculation:

The genome sequence is processed to extract every possible k-mer of a specified length (for example, k=4 or k=5), and the frequency of each unique k-mer is calculated. This frequency provides a measure of how often each k-mer appears in the sequence, with higher counts indicating more common k-mers that may play a role in genetic stability or function.

2) Clumping Detection with Sliding Window Approach:

To detect clumping, a sliding window of fixed length is moved along the sequence. Within each window, the occurrence of each k-mer is recorded. K-mers that appear multiple times within a single window are classified as "clumps," indicating localized repetition in the sequence. This step reveals areas where the genome may exhibit higher levels of structural redundancy, potentially correlating with functional or regulatory elements.

3) Frequency Thresholding and Analysis:

To focus on the most significant clumping patterns, a frequency threshold is applied, and only k-mers exceeding this threshold are considered. For instance, if a 4-mer must appear at least five times within a particular genomic segment to be considered a clump, k-mers below this frequency are disregarded. The resulting high-frequency k-mers, or clumps, provide insight into prominent sequence motifs that could have biological relevance, particularly in viral genomes where short repeated sequences may influence replication or infectivity.

4) Identification of High-Frequency Clumps:

After thresholding, the most frequent k-mers are identified and analyzed for their positions and spatial distribution. The spatial proximity of these high-frequency clumps within the genome highlights areas that may be functionally important or structurally conserved. The identification of these recurring motifs offers clues to regions within the genome that may play roles in genetic stability or viral adaptation, supporting hypotheses regarding the evolutionary similarities or differences between SARS-CoV-2 and RaTG13.

4.4 CGR-Based Similarity Comparison

To quantitatively assess the similarity between the CGR patterns of SARS-CoV-2 and RaTG13, we used matrix-based correlation analysis. Each CGR plot was converted into a matrix representation, where matrix values correspond to the density of points (or frequency of k-mers) at each position within the CGR plot.

1) CGR Matrix Conversion:

The CGR plot for each genome sequence is converted into a density matrix, where each cell value represents the number of times points were mapped to that cell in the CGR plot. This matrix serves as a numerical representation of the CGR, capturing the structural and k-mer distribution features of the genome sequence.

2) Matrix Correlation Calculation:

A Pearson correlation coefficient is calculated between the CGR matrices of SARS-CoV-2 and RaTG13, providing a similarity score that quantifies the extent to which the k-mer distribution patterns align between the two genomes. The correlation score ranges from -1 (indicating complete dissimilarity) to 1 (indicating perfect similarity). Higher correlation scores suggest that the genomes have similar structural and sequence characteristics, while lower scores indicate more distinct k-mer distributions.

3) Threshold Analysis for Genomic Similarity:

To assess the significance of the similarity score, a correlation threshold is applied, typically based on known values for closely related genomes. If the similarity score exceeds this threshold, it suggests substantial structural resemblance in the sequence distribution of k-mers, highlighting conserved regions that may contribute to shared evolutionary or functional traits.

4) Visualization of CGR Similarity:

Comparative CGR plots are generated to visualize similarities and differences in k-mer distribution, highlighting conserved motifs and distinct regions. Visual inspection of CGR plots allows further insight into sequence organization and structural motifs, offering an additional layer of understanding in the genomic analysis.

4.5 Similarity Analysis

To quantify similarity between CGR matrices of SARS-CoV-2 and other genomes, the CGR plots were converted into matrix forms. The matrices were compared using correlation analysis, giving a similarity percentage that represents the genetic alignment between two sequences. Formula Used:

$$\text{Similarity Percentage} = \text{Correlation Coefficient} \times 100$$

4.6 Evaluation Metrics

The quality and significance of the CGR and k-mer clumping analyses were evaluated through correlation metrics and frequency distribution comparisons:

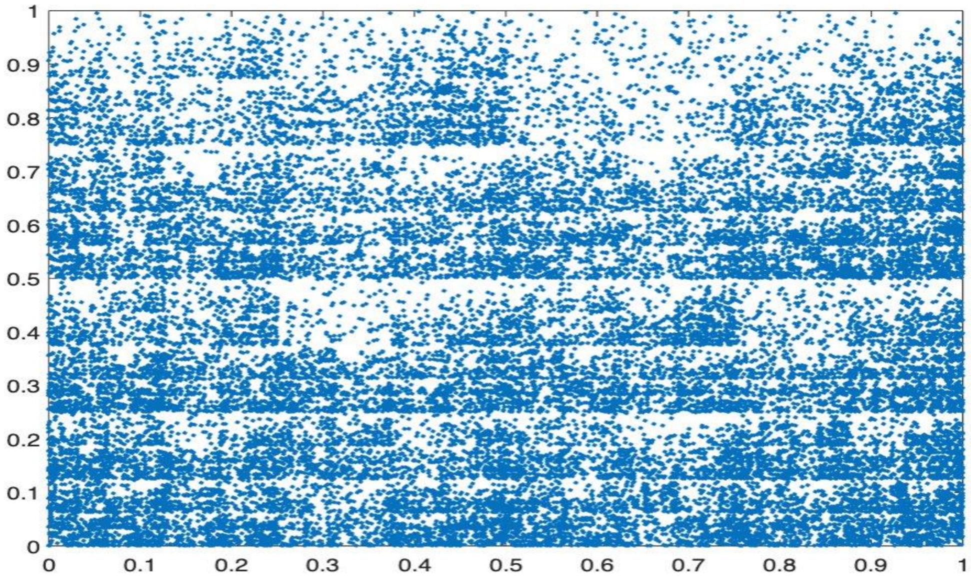
1) Pearson Correlation Coefficient: Used to quantify CGR matrix similarity, providing an objective measure of structural resemblance between SARS-CoV-2 and RaTG13.

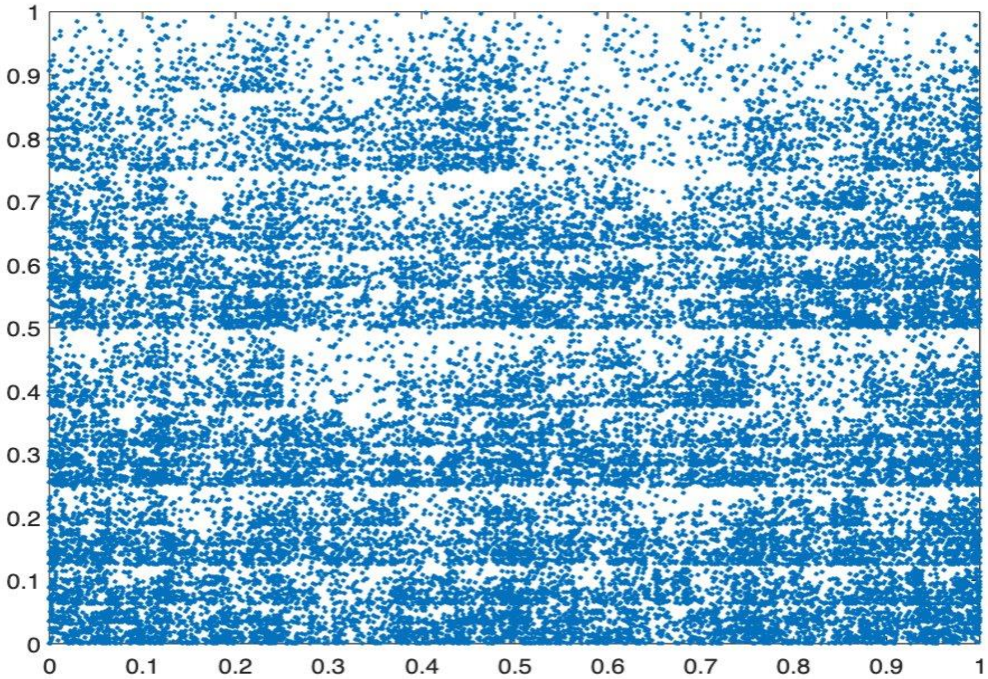
2) Top-K Frequency Analysis: The distribution of high-frequency k-mers was analyzed, offering insights into shared or unique clumping patterns, which reveal potential functional or evolutionary relationships.

This methodological approach combines CGR visualization, k-mer clumping, and correlation analysis to investigate sequence patterns and structural motifs, supporting a comprehensive genomic comparison between SARS-CoV-2 and RaTG13.

5. Results :

```
The CGR matrix for SARS cov-2 genetic sequence
```





6. Future Work

- ***Expanded Comparative Analysis with Other Coronaviruses:*** A more thorough understanding of the evolutionary tree of SARS-CoV-2 may be obtained by expanding the analysis to include coronaviruses other than RaTG13, such as other bat coronaviruses and strains originating from pangolins. This would make it easier to pinpoint the precise genetic alterations that permit zoonotic transmission.

- ***Machine Learning for Predicting Mutations***: Applying machine learning models to the genomic data may improve the ability to anticipate evolutionary trends and mutations in SARS-CoV-2 and related viruses. Additionally, predictive models may aid in locating genomic areas that might alter as a result of host interactions or environmental stressors.
- ***Functional Studies of Important Genetic Regions***: To confirm the biological significance of the conserved K-mers, clump regions, and replication origin locations found in our analysis, researchers should carry out laboratory-based investigations.
- ***Real-Time Monitoring and Surveillance:*** Developing bioinformatics tools that leverage insights from this analysis could improve real-time monitoring of emerging coronavirus strains. Such tools could be integrated into global surveillance systems to detect new mutations associated with increased transmissibility or virulence.

7. Conclusion

This project provides a detailed comparative genomic analysis of SARS-CoV-2 and Bat Betacoronavirus RaTG13, uncovering significant genetic similarities that suggest a shared evolutionary lineage. Through bioinformatics tools like K-mer analysis, GC skew, and Chaos Game Representation (CGR), the study reveals conserved sequence patterns, structural motifs, and key genomic regions that may play critical roles in the viruses' replication and transmission. The high genomic similarity—up to 96% in specific regions—between SARS-CoV-2 and RaTG13 underscores the potential for zoonotic crossover, highlighting RaTG13 as a possible precursor or close relative of SARS-CoV-2.

This comparative approach not only advances our understanding of SARS-CoV-2's genetic makeup but also offers a methodological framework for analysing other viruses with pandemic potential. By identifying patterns and structural markers associated with viral replication and evolution, the study contributes to broader research efforts in virology and epidemiology. The insights gained could inform ongoing strategies for monitoring, preventing, and treating viral outbreaks, reinforcing the need for continued research into zoonotic viruses to better anticipate and mitigate future public health threats.

8. References

[1] Löchel, H. F., & Heider, D. (2021). Chaos game representation and its applications in bioinformatics. *Genes & Diseases, 8*(4), 505–515. Retrieved from https://www.sciencedirect.com/science/article/pii/S2001037021004736

[2] Chan, E. Y. S., & Corless, R. M. (2020). Chaos game representation. *arXiv preprint*. arXiv:2012.09638. Retrieved from https://arxiv.org/abs/2012.09638

[3] Young, S., & Gilles, J. (2024). Use of 3D chaos game representation to quantify DNA sequence similarity with applications for hierarchical clustering. *arXiv preprint*. arXiv:2411.05266. Retrieved from https://arxiv.org/abs/2411.05266

[4] Löchel, H. F., & Heider, D. (2021). Chaos game representation and its applications in bioinformatics. Retrieved from https://www.researchgate.net/publication/356117663_Chaos_Game_Representation_and_its_applications_in_Bioinformatics

[5] Jeffrey, H. J. (1990). Chaos game representation of gene structure. *Nucleic Acids Research, 18*(8), 2163–2170. Retrieved from https://academic.oup.com/nar/article/18/8/2163/2383530?login=false

[6] Hu, B., Guo, H., Zhou, P., & Shi, Z. L. (2021). Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology, 19*(3), 141–154. Retrieved from https://www.nature.com/articles/s41579-020-00459-7

[7] Wrobel, A. G., Benton, D. J., Xu, P., Roustan, C., Martin, S. R., Rosenthal, P. B., Skehel, J. J., & Gamblin, S. J. (2021). SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and Furin-cleavage effects. *Nature Structural & Molecular Biology, 28*(8), 762–769. Retrieved from https://www.nature.com/articles/s41594-020-0468-7

[8] Tan, W., Wang, X., Li, W., Zhu, N., & The 2019-nCoV Outbreak Joint Field Epidemiology Investigation Team. (2021). Binding and molecular basis of the bat coronavirus RaTG13 virus to ACE2 in humans and other species. *PLOS Pathogens, 17*(5), e1009539. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC8142884/