PRINCIPLES OF BIOINFORMATICS PROJECT

Comparative Analysis of Asthma, COVID-19 & Cystic Fibrosis: Unraveling Shared Pathogenic Mechanisms

A class project by:

Aditya Nayak

Milken Institute of Public Health, George Washington University

PUBH 6860: Principles of Bioinformatics

Prof. Keith Crandall

November 29, 2023

Introduction

The intricate symphony of life resonates with the echoes of molecular evolution. Proteins, the microscopic maestros, orchestrate cellular dances, their choreography shaped by eons of mutations and adaptations. In the arena of respiration, three seemingly disparate conditions – asthma, COVID-19, and cystic fibrosis – share the stage, each with its own poignant performance. Asthma, a chronic inflammatory waltz, disrupts the airflow with hyperactive immune responses. COVID-19, a viral invader, unleashes a macabre pirouette of cellular hijacking. And cystic fibrosis, a genetic mutation's cruel twist, thickens the mucus-laden atmosphere, impeding the vital exchange of gases.

Yet, amidst the distinct melodies of these respiratory ailments, whispers of shared motifs tantalize the curious mind. Could their molecular narratives, woven from the threads of evolution, reveal hidden connections? This inquiry delves into the intricate tapestry of these conditions, tracing their molecular pathways through the evolutionary lens. We dissect the proteins implicated in each disease, searching for convergences in their sequences, structures, and functions. Do the immune pathways of asthma resonate with the viral dance of COVID-19? Does the faulty machinery of cystic fibrosis cast a shadow on the delicate balance of airway homeostasis?

Thus, the question beckons: amidst the divergent melodies of asthma, COVID-19, and cystic fibrosis, can we discern the delicate harmonies of a shared molecular evolution? This research embarks on a quest to answer this call, exploring the intricate tapestry of respiratory disease through the lens of molecular evolution.

Background:

Asthma is a major non-communicable disease (NCD), affecting both children and adults, and is the most common chronic disease among children causing inflammation and narrowing of the small airways in the lungs cause asthma symptoms, which can be any combination of cough, wheeze, shortness of breath and chest tightness. It is caused by inflammation and muscle tightening around the airways, which makes it harder to breathe. Symptoms can include coughing, wheezing, shortness of breath and chest tightness. These symptoms can be mild or severe and can come and go over time. About 1 in 13 people in the United States has asthma, according to the Centers for Disease Control and Prevention. It affects people of all ages and often starts during childhood. Certain things can set off or worsen asthma symptoms, such as pollen, exercise, viral infections, or cold air. These are called asthma triggers. When symptoms get worse, it is called an asthma attack. There is no cure for asthma, but treatment and an asthma action plan can help you manage it.

Pneumonia caused by Severe Acute Respiratory Syndrome Coronavirus 2 or *SARS-CoV-2* infection emerged in Wuhan City, Hubei Province, China in December 2019. By Feb. 11, 2020, the *World Health Organization* (WHO) officially named the disease resulting from infection with SARS-CoV-2 as coronavirus disease 2019 (COVID-19). COVID-19 represents a spectrum of clinical manifestations that typically include fever, dry cough, and fatigue, often with pulmonary involvement. SARS-CoV-2 is highly contagious and most individuals within the population at large are susceptible to infection. The virus can also spread from an infected person’s mouth or nose in small liquid particles when they cough, sneeze, speak, sing or breathe. The liquid particles

can land on hands, objects, or surfaces around the person when they cough or talk, and people can then become infected with COVID-19 from touching hands, objects or surfaces with droplets and then touching their eyes, nose, or mouth. Additionally, transmission can occur from those with mild symptoms or from those who do not feel ill. These particles range from larger respiratory droplets to smaller aerosols.

Whereas, Cystic Fibrosis (CF), an autosomal recessive genetic disease, is caused by a mutation in the gene encoding the cystic fibrosis transmembrane conductance regulator (CFTR). This mutation reduces the release of chloride ions (Cl^-) in epithelial tissues, and hyperactivates the epithelial sodium channels (ENaC) which aid in the absorption of sodium ions (Na^+). Consequently, the mucus becomes dehydrated and thickened, making it a suitable medium for microbial growth. CF causes several chronic lung complications like thickened mucus, bacterial infection and inflammation, progressive loss of lung function, and ultimately, death. CF is a recessive genetic disease caused by a mutation in the epithelial chloride channel - cystic fibrosis transmembrane conductance regulator (CFTR). CF is a predominant genetic disorder with a disease severity ranging from mild to life-threatening.

Keywords: *symptoms (cough, wheeze, shortness of breath, chest tightness), respiratory droplets, aerosols, inflammation, clinical manifestations.*

Methods

A. Data Collection:

The following DNA sequences of asthma, COVID–19 and Cystic Fibrosis are collected from the GenBank database, consisting of 20 most commonly referred sequences from the database. GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.

B. Statistical Analysis

Comparative Analysis of Clinical Symptoms between Asthma, COVID-19, and Cystic Fibrosis While asthma, COVID-19, and cystic fibrosis all impact the respiratory system, their clinical presentations exhibit both overlapping and distinct features. Here's a comparison based on recent research and clinical guidelines:

Similarities:

- *Respiratory symptoms:* All three can cause cough, dyspnea, wheezing, and chest tightness, varying in severity and duration.
- *Fever:* Common in COVID-19, occasional in severe asthma exacerbations or cystic fibrosis infections.
- *Fatigue:* A hallmark symptom of COVID-19 and cystic fibrosis, potentially present in severe asthma.
- *Decreased exercise tolerance:* All three can limit exercise capacity due to respiratory difficulties.

Differences:

Symptom Onset and Progression:

- COVID-19: Symptoms usually develop within 2-14 days after exposure and can progress rapidly. Hospitalization may be needed.
- Asthma: Symptoms develop gradually over days or weeks, or worsen in response to triggers. Severe exacerbations may require hospitalization.
- Cystic Fibrosis: Symptoms present from birth/early childhood and progressively worsen over time. Hospitalizations are common for exacerbations and complications.

Specific Symptoms:

- COVID-19: Loss of taste/smell, muscle aches, sore throat, gastrointestinal symptoms.
- Asthma: Allergic rhinitis, eczema, clear or white mucus production.
- Cystic Fibrosis: Salty-tasting sweat, chronic cough with thick, green mucus, recurrent lung infections.

Diagnostic Tests:

- COVID-19: Viral PCR test, rapid antigen test.
- Asthma: Lung function tests, allergy testing.
- Cystic Fibrosis: Sweat chloride test, genetic testing.

Treatment:

- COVID-19: Supportive care, antiviral medications for specific cases.
- Asthma: Inhaled corticosteroids, bronchodilators, allergen avoidance, immunotherapy.
- Cystic Fibrosis: Pancreatic enzyme replacement, airway clearance techniques, antibiotics, lung transplantation (potential).

While some symptoms overlap, distinct features in onset, progression, and specific symptoms help differentiate these conditions. Diagnostic tests and treatment approaches are also specific to each disease.

C. Bioinformatics Approach

To determine the similarity between the DNA sequences of asthma, COVID-19 and CF, we will have to construct a phylogeny tree which helps in highlighting similar features between them. However, before this step we must

perform multiple sequence alignment on the sequences. In bioinformatics, Multiple Sequence Alignment (MSA) is a technique used to compare the sequences of three or more biological molecules, typically proteins, DNA, or RNA. It aims to arrange these sequences in a way that highlights their similarities and differences, providing valuable insights into their evolution, function, and structure. Its purpose:

- Identify conserved regions: Identify regions that are identical or highly similar across the sequences, suggesting important functional or structural elements.
- Trace evolutionary relationships: Analyze the pattern of substitutions and insertions/deletions to infer the evolutionary history of the compared sequences.
- Predict protein structure and function: Use the alignment information to model the 3D structure of proteins and understand their functional mechanisms.
- Develop bioinformatic tools: MSAs are used to train algorithms for tasks like protein function prediction and gene identification.

Methods:

- Pairwise alignment: This basic approach aligns two sequences at a time. Multiple pairwise alignments can be combined into an MSA, but this can lead to inconsistencies.
- Progressive alignment: This method builds an MSA by sequentially aligning each sequence to the existing alignment.
- Iterative alignment: This approach refines an initial alignment through multiple cycles of optimization, often using statistical models of sequence evolution.

Tools:

- Clustal
- MAFFT
- Muscle
- T-Coffee
- BLAST

To achieve MSA, we make use of Molecular Evolutionary Genetics Analysis or MEGA software.

MEGA

MEGA, is a comprehensive software tool used in the field of bioinformatics and computational biology. MEGA provides a wide range of features and functionalities for researchers and scientists working with molecular data, particularly in the context of phylogenetics and evolutionary biology. It is widely used for sequence analysis, phylogenetic tree reconstruction, and molecular evolution studies. Some key features and uses of the MEGA software are:

1. *Sequence Alignment:* MEGA allows researchers to perform multiple sequence alignments and pairwise sequence alignments. It supports various alignment algorithms, to align DNA, RNA, and protein sequences. Sequence alignment is essential for identifying conserved regions and functional domains in molecular data.
2. *Phylogenetic Tree Construction:* MEGA provides tools for building phylogenetic trees, which depict the evolutionary relationships between species or genes. It supports various tree-building methods, such as

neighbor-joining, maximum likelihood, and UPGMA (Unweighted Pair Group Method with Arithmetic Mean).

3. *Phylogenetic Analysis:* Users can conduct extensive phylogenetic analysis, including the estimation of genetic distances, bootstrap analysis to assess tree reliability, and the visualization of tree topologies.
4. *Substitution Models:* It offers a wide range of nucleotide and amino acid substitution models to choose from when constructing phylogenetic trees. These models help in understanding the rates and patterns of molecular evolution.
5. *Evolutionary Analysis:* Researchers can perform evolutionary analysis, including the calculation of synonymous and nonsynonymous substitution rates, estimation of divergence times, and ancestral sequence reconstruction.
6. *Statistical Tests:* MEGA allows users to perform statistical tests to evaluate hypotheses related to molecular evolution, such as tests for positive selection and neutrality.
7. *Sequence Editing and Annotation:* The software includes sequence editing tools that enable users to modify and annotate sequences with relevant information. This is helpful for preparing sequences for analysis.
8. *Visualization:* MEGA provides various options for visualizing and customizing the appearance of phylogenetic trees and sequence alignments.
9. *Publication-Ready Outputs:* Users can generate high-quality graphical representations of their results, suitable for publication in scientific journals.

10. *User-Friendly Interface:* MEGA offers a user-friendly graphical interface, making it accessible to both beginners and experienced bioinformaticians.
11. *Data Import/Export:* It supports a variety of file formats, including FASTA, GenBank, and Nexus, for data import and export.

It assists scientists in investigating the evolutionary history of genes, proteins, and species, and it plays a crucial role in understanding genetic diversity, adaptation, and speciation processes. Researchers utilize MEGA to analyze a wide range of molecular data, right from DNA sequences in genomics to protein sequences in structural biology.

On observing the given data, we find that there are some alignments that are needed to be done. And based on the number of records, the most suitable algorithm for aligning our sequences would be the MUSCLE algorithm due to its accuracy and efficiency to align large datasets.

- **MUSCLE**

  MUSCLE, which stands for "Multiple Sequence Comparison by Log-Expectation," is a widely used bioinformatics algorithm and software tool for multiple sequence alignment (MSA). In bioinformatics, multiple sequence alignment is the process of aligning and comparing multiple biological sequences (typically DNA, RNA, or protein sequences) to identify conserved regions, functional domains, and evolutionary relationships among them.

The various key features about the MUSCLE algorithm are as follows:

1. ***Alignment Method:*** MUSCLE uses a progressive approach to align sequences. It begins with a pairwise sequence alignment and then builds a guide tree or a distance matrix to determine the order in which sequences are aligned. It iteratively refines the alignment in a progressive manner.
2. ***Scoring:*** MUSCLE employs a scoring system to assess the quality of sequence alignments. It aims to maximize the log-odds scores based on the expectation-maximization algorithm, which considers the probabilities of observing gaps and matches.
3. ***Speed and Accuracy:*** MUSCLE is known for its speed and is often used for aligning large datasets of sequences. It provides good accuracy in aligning sequences, making it a popular choice for a wide range of bioinformatics applications.
4. ***Command-Line & GUI:*** MUSCLE is available as a command-line tool, which is suitable for batch processing, as well as through graphical user interfaces (GUIs) in some bioinformatics software packages.
5. ***Multiple Input Formats:*** It supports various sequence data formats, including FASTA and Clustal formats, making it flexible for different input types.
6. ***Parallelization:*** Some versions of MUSCLE support parallelization, which can significantly improve the alignment speed, especially when dealing with a large number of sequences.

MUSCLE is valuable in various bioinformatics applications, such as phylogenetic analysis, protein structure prediction, and identifying conserved regions in genomic sequences. It plays a crucial role in understanding the evolutionary relationships between genes, proteins, and species by aligning and comparing sequences from different organisms.

After conducting the multiple sequence analysis, we have to take our aligned nucleotide data sequences and then determine the best-fit model of evolution for our data and develop our phylogenic tree.

- **Phylogenic Trees:**

  A phylogenetic tree, often simply called a "phylogeny," is a branching diagram that represents the evolutionary relationships and ancestry of a group of organisms or species. It is a visual representation of the evolutionary history of these organisms, showing how they are believed to be related through common ancestors. Phylogenetic trees are commonly used in biology, particularly in fields such as evolutionary biology, systematics, and genetics, to study and depict the evolutionary connections between different species or groups of organisms.

  Key features of a phylogenetic tree are:

  1. ***Nodes:*** Points where branches on the tree intersect. Nodes represent common ancestors shared by the species or groups connected by that node.
  2. ***Branches:*** Lines connecting nodes to represent evolutionary lineages.

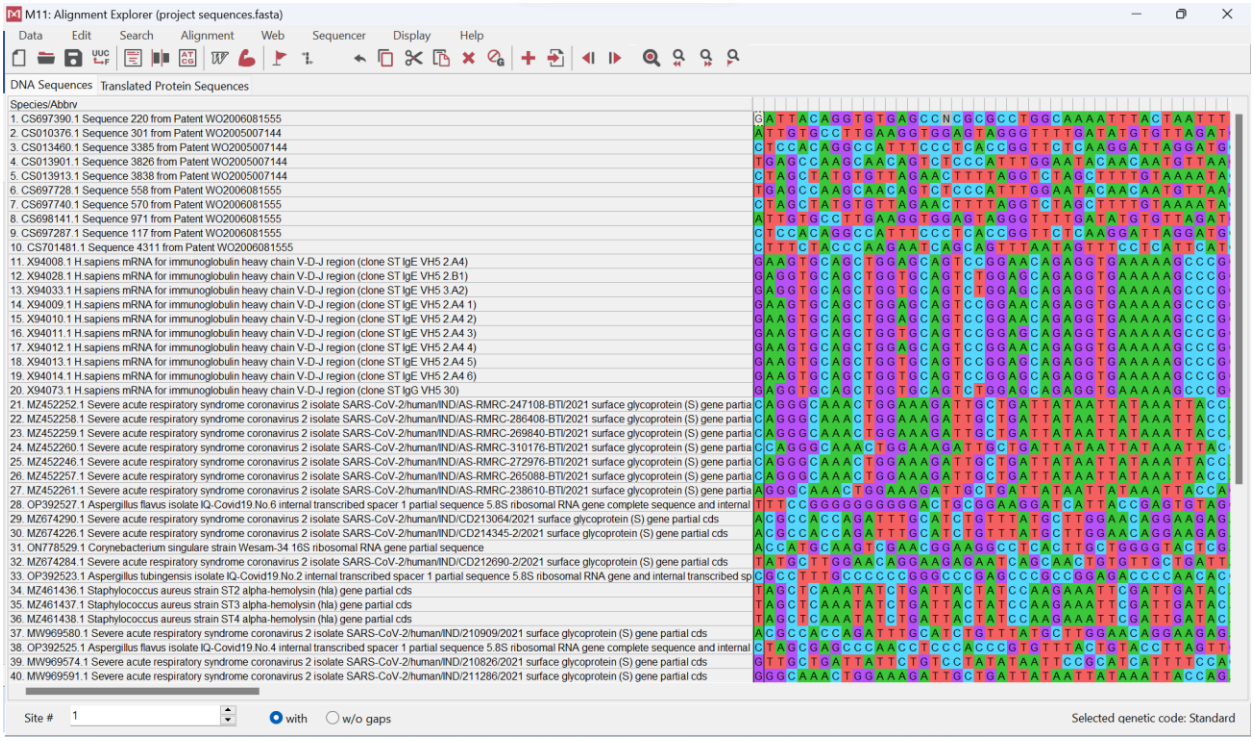
3. ***Leaves or Tips***: Terminal points on the tree that represent the species or groups under study.
4. ***Root:*** The common ancestor of all the species or groups in the tree. It is typically placed at the base of the tree.
5. ***Internal Nodes:*** Nodes within the tree that connect intermediate ancestors or common ancestors.

Phylogenetic trees can be constructed using various methods and types of data, such as DNA sequences, protein sequences, morphological characteristics, or a combination of these. These trees are valuable tools for understanding the evolutionary history, relationships, and divergence times of different organisms. They help scientists make inferences about which species share a more recent common ancestor and provide insights into the patterns and processes of evolution.

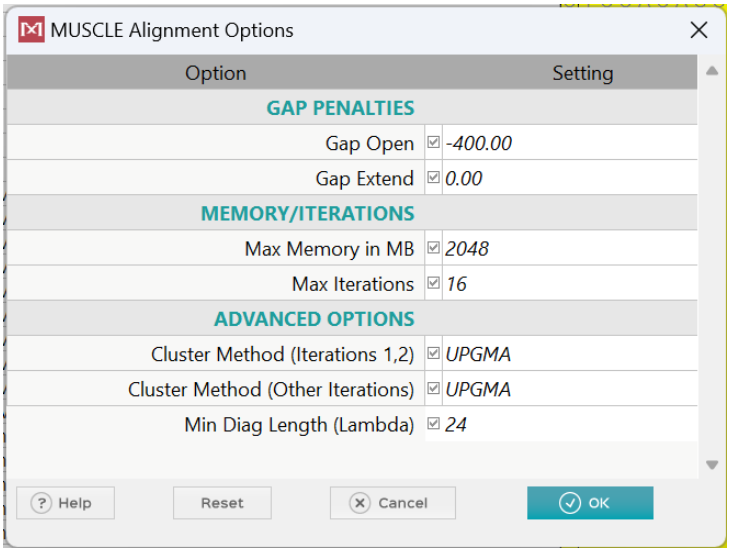
Hence, by using this tree diagram, we will be able to complement any genetic or evolutionary similarity between the selected DNA sequences of our selected respiratory diseases.

Results

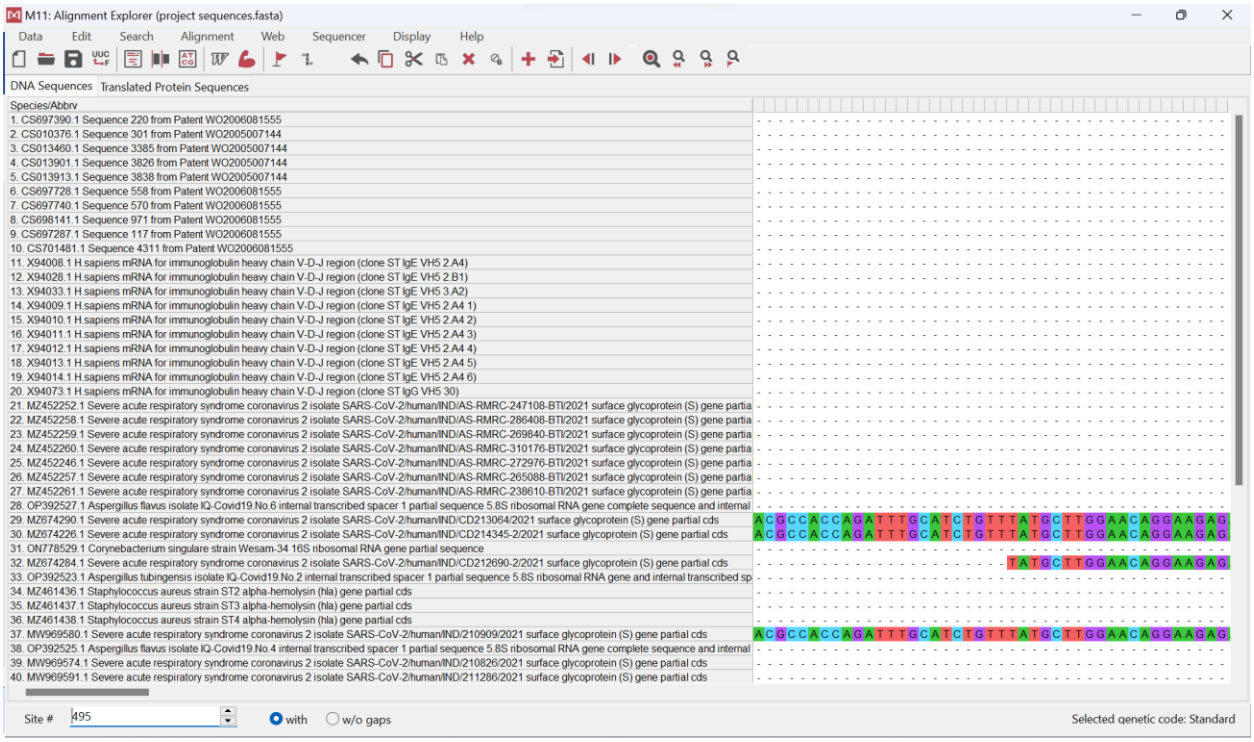
On uploading the FASTA file we have the sequences as follows:



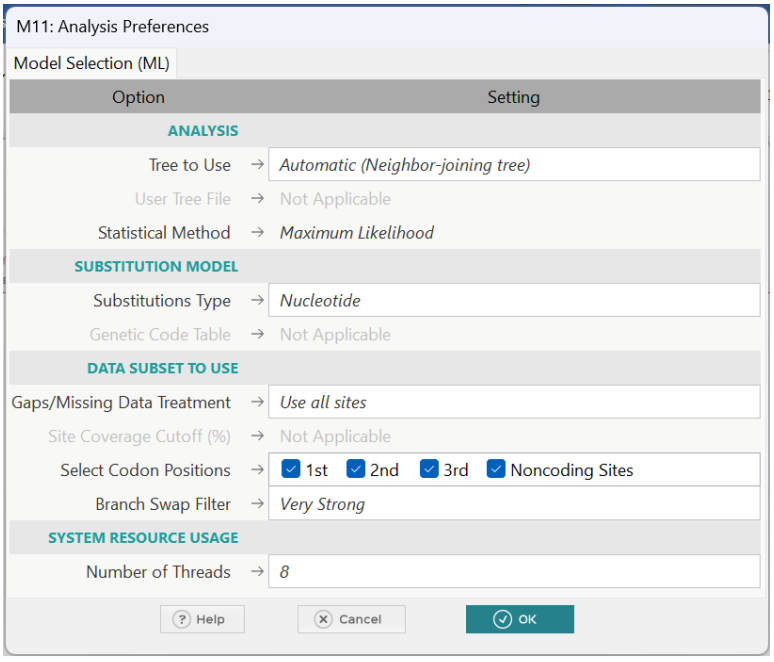
As these sequences are not aligned, we will need to perform MSA using MUSCLE algorithm using the following parameters:



We get the following aligned sequence:



On saving this aligned FASTA file, we can head towards creating a phylogeny tree of our sequences. In order to build our phylogeny tree, we will need to know the best-fit model which is obtained from the substitution matrix with the following parameters:



We get the following table:

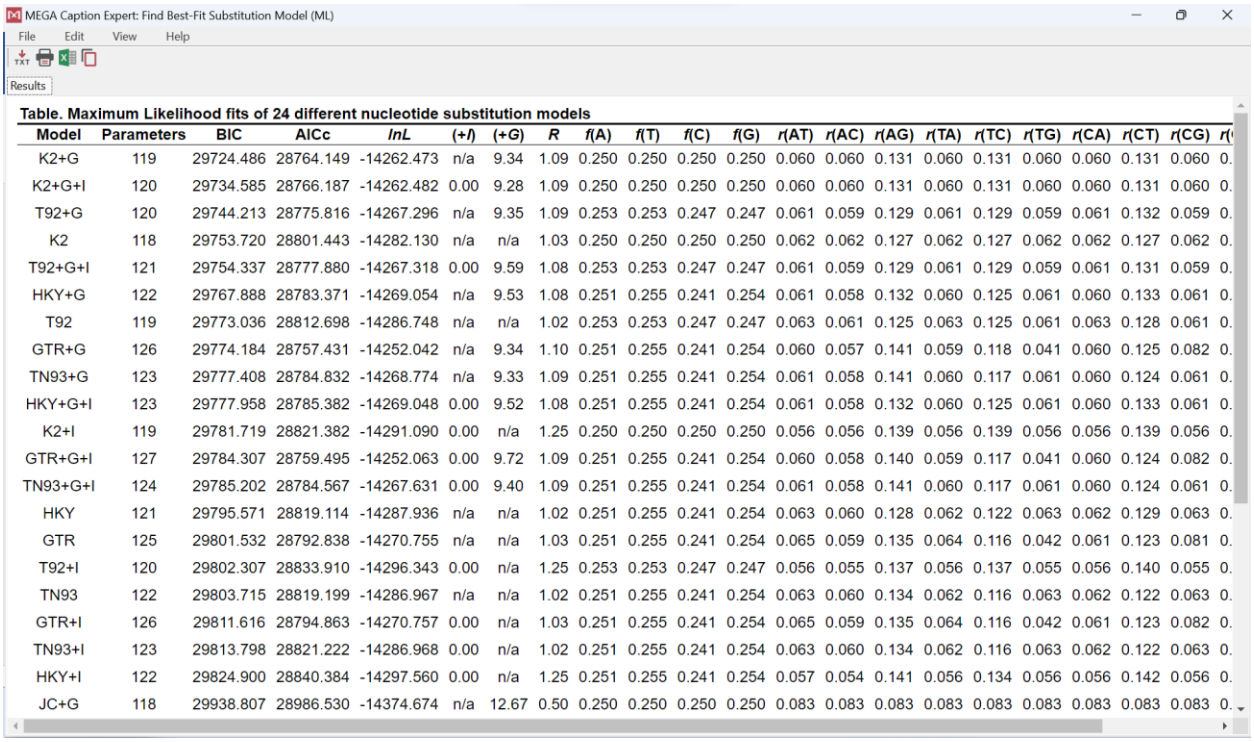


Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(TA)	*r*(TC)	*r*(TG)	*r*(CA)	*r*(CT)	*r*(CG)	*r*(
K2+G	119	29724.486	28764.149	-14262.473	n/a	9.34	1.09	0.250	0.250	0.250	0.250	0.060	0.060	0.131	0.060	0.131	0.060	0.060	0.131	0.060	0.
K2+G+I	120	29734.585	28766.187	-14262.482	0.00	9.28	1.09	0.250	0.250	0.250	0.250	0.060	0.060	0.131	0.060	0.131	0.060	0.060	0.131	0.060	0.
T92+G	120	29744.213	28775.816	-14267.296	n/a	9.35	1.09	0.253	0.253	0.247	0.247	0.061	0.059	0.129	0.061	0.129	0.059	0.061	0.132	0.059	0.
K2	118	29753.720	28801.443	-14282.130	n/a	n/a	1.03	0.250	0.250	0.250	0.250	0.062	0.062	0.127	0.062	0.127	0.062	0.062	0.127	0.062	0.
T92+G+I	121	29754.337	28777.880	-14267.318	0.00	9.59	1.08	0.253	0.253	0.247	0.247	0.061	0.059	0.129	0.061	0.129	0.059	0.061	0.131	0.059	0.
HKY+G	122	29767.888	28783.371	-14269.054	n/a	9.53	1.08	0.251	0.255	0.241	0.254	0.061	0.058	0.132	0.060	0.125	0.061	0.060	0.133	0.061	0.
T92	119	29773.036	28812.698	-14286.748	n/a	n/a	1.02	0.253	0.253	0.247	0.247	0.063	0.061	0.125	0.063	0.125	0.061	0.063	0.128	0.061	0.
GTR+G	126	29774.184	28757.431	-14252.042	n/a	9.34	1.10	0.251	0.255	0.241	0.254	0.060	0.057	0.141	0.059	0.118	0.041	0.060	0.125	0.082	0.
TN93+G	123	29777.408	28784.832	-14268.774	n/a	9.33	1.09	0.251	0.255	0.241	0.254	0.061	0.058	0.141	0.060	0.117	0.061	0.060	0.124	0.061	0.
HKY+G+I	123	29777.958	28785.382	-14269.048	0.00	9.52	1.08	0.251	0.255	0.241	0.254	0.061	0.058	0.132	0.060	0.125	0.061	0.060	0.133	0.061	0.
K2+I	119	29781.719	28821.382	-14291.090	0.00	n/a	1.25	0.250	0.250	0.250	0.250	0.056	0.056	0.139	0.056	0.139	0.056	0.056	0.139	0.056	0.
GTR+G+I	127	29784.307	28759.495	-14252.063	0.00	9.72	1.09	0.251	0.255	0.241	0.254	0.060	0.058	0.140	0.059	0.117	0.041	0.060	0.124	0.082	0.
TN93+G+I	124	29785.202	28784.567	-14267.631	0.00	9.40	1.09	0.251	0.255	0.241	0.254	0.061	0.058	0.141	0.060	0.117	0.061	0.060	0.124	0.061	0.
HKY	121	29795.571	28819.114	-14287.936	n/a	n/a	1.02	0.251	0.255	0.241	0.254	0.063	0.060	0.128	0.062	0.122	0.063	0.062	0.129	0.063	0.
GTR	125	29801.532	28792.838	-14270.755	n/a	n/a	1.03	0.251	0.255	0.241	0.254	0.065	0.059	0.135	0.064	0.116	0.042	0.061	0.123	0.081	0.
T92+I	120	29802.307	28833.910	-14296.343	0.00	n/a	1.25	0.253	0.253	0.247	0.247	0.056	0.055	0.137	0.056	0.137	0.055	0.056	0.140	0.055	0.
TN93	122	29803.715	28819.199	-14286.967	n/a	n/a	1.02	0.251	0.255	0.241	0.254	0.063	0.060	0.134	0.062	0.116	0.063	0.062	0.122	0.063	0.
GTR+I	126	29811.616	28794.863	-14270.757	0.00	n/a	1.03	0.251	0.255	0.241	0.254	0.065	0.059	0.135	0.064	0.116	0.042	0.061	0.123	0.082	0.
TN93+I	123	29813.798	28821.222	-14286.968	0.00	n/a	1.02	0.251	0.255	0.241	0.254	0.063	0.060	0.134	0.062	0.116	0.063	0.062	0.122	0.063	0.
HKY+I	122	29824.900	28840.384	-14297.560	0.00	n/a	1.25	0.251	0.255	0.241	0.254	0.057	0.054	0.141	0.056	0.134	0.056	0.056	0.142	0.056	0.
JC+G	118	29938.807	28986.530	-14374.674	n/a	12.67	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.

The best-fit model is the one with the lowest AICc score which in our case is K2+G.

In bioinformatics, the K2+G model is a statistical model used for estimating nucleotide substitution rates in DNA or RNA sequences. It builds upon the popular K2 model (Kimura two-parameter) by incorporating an additional parameter to account for compositional heterogeneity, making it more suitable for datasets with uneven base frequencies.

Components:

- **K2 Model:** This base model assumes transitions (purine to purine, pyrimidine to pyrimidine) occur twice as frequently as transversions (purine to pyrimidine, vice versa). It uses two parameters: transition/transversion rate ratio (κ) and transition rate (α).

- ***G (Gamma Distribution):*** The G in K2+G stands for the gamma distribution of rates among sites. This accounts for rate heterogeneity among different sites in the sequence. In biological sequences, certain positions may evolve more rapidly or slowly than others. The gamma distribution provides a way to model this rate variation. The gamma distribution is often used to model the variation in evolutionary rates among different sites in a sequence. It assumes that rates of evolution follow a gamma distribution, allowing for some sites to evolve rapidly while others evolve slowly.
- **Compositional Bias:** The K2+G model adds a parameter (γ) to account for variations in the base frequencies across positions in the sequence. This allows for more accurate estimation of substitution rates by correcting for biases introduced by uneven base representation.

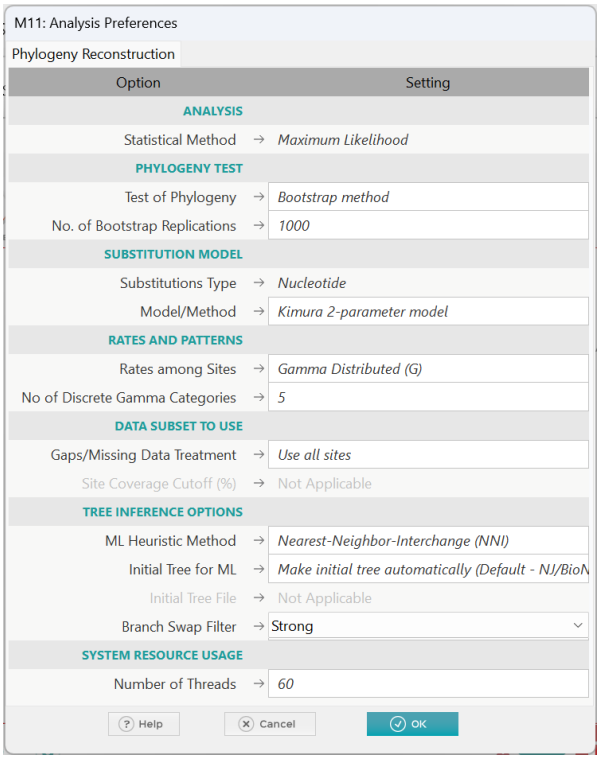
Advantages:

- Improved accuracy: K2+G provides more accurate substitution rate estimates compared to K2, especially for sequences with skewed base frequencies.
- Wide applicability: K2+G is suitable for analyzing diverse sequence datasets, including bacterial genomes, viral sequences, and eukaryotic genes.
- Flexibility: The γ parameter can be estimated from the data or set to a predefined value based on prior knowledge.

Limitations:

- Increased complexity: Compared to K2, K2+G has one additional parameter, requiring more data for accurate estimation and potentially increasing computational time.

Hence, by using the K2+G (Kimura two-parameter with Gamma) and selecting the following parameters:



We get the phylogeny tree as follows:

On closer observation, we notice that there is a similarity present between the DNA sequences of COVID-19 and those of CF and another part shows similarity between asthma DNA sequences and COVID-19 showing that variants of Sars-COVID-19 features present in both asthma and CF sequences.

Discussion

In this project, we aimed to deepen our understanding of multiple sequence alignments (MSA) and phylogeny estimation within the domain of bioinformatics. The dataset served as the basis for our exploration, and we delved into various aspects of MSA, model selection, and the interpretation of phylogenetic estimates.

Multiple Sequence Alignment (MSA) Significance:

MSA is a pivotal technique in bioinformatics for aligning and comparing sequences of biological macromolecules. Accurate MSA is fundamental for various biological analyses, such as phylogenetics, functional annotation, and structure prediction. Our emphasis in Part 1 of the assignment was on achieving a high-quality alignment using the MEGA software.

Phylogenetic Analysis and Model Selection:

Phylogenetic analysis involves reconstructing evolutionary trees to depict historical relationships between species or sequences. A critical step is the selection of an appropriate model of evolution. We determined the best-fit model for our data to be K2+G (Kimura two-parameter with Gamma distribution). This model accounts for varying rates of nucleotide substitution and site rate heterogeneity, offering a more realistic representation of evolutionary dynamics.

MUSCLE Algorithm for Multiple Sequence Alignment:

To address alignment issues, we opted for the MUSCLE algorithm due to its efficiency and accuracy, particularly with large datasets. MUSCLE's progressive alignment approach and scoring system proved beneficial for identifying conserved regions and functional domains in our nucleotide sequences.

Phylogenetic Tree Construction:

We constructed a phylogenetic tree based on the aligned sequences using the selected model. Our tree revealed evolutionary relationships, with nodes representing common ancestors, branches depicting lineages, and leaves representing individual sequences. The visual representation aided in interpreting the relatedness of sequences.

Conclusion

From this following project, we have constructed a phylogeny tree to determine any similarities between the DNA sequences of these respiratory diseases displaying COVID-19 DNA sequences showing partial similarity between asthma and CF sequences. However, further focus must be given on the MEGA software to study the reasons for its occasional inability to perform the phylogeny tree construction.

References

World Health Organization: WHO & World Health Organization: WHO. (2023, May 4). *Asthma*. https://www.who.int/news-room/fact-sheets/detail/asthma

What is asthma? | *NHLBI, NIH*. (2022, March 24). NHLBI, NIH. https://www.nhlbi.nih.gov/health/asthma

Shi, Y., Gang, W., Cai, X. P., Deng, J., Zheng, L., Zhu, H., Zheng, M., Yang, B., & Chen, Z. (2020). An overview of COVID-19. *Journal of Zhejiang University SCIENCE B*, *21*(5), 343–360. https://doi.org/10.1631/jzus.b2000083

Healthcare workers. (2020, February 11). Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/hcp/non-us-settings/overview/index.html#background

National Center for Biotechnology Information. (n.d.). https://www.ncbi.nlm.nih.gov/

Tamura K, Stecher G, and Kumar S (2021) MEGA11: Molecular Evolutionary Genetics Analysis version 11. Molecular Biology and Evolution 38:3022-3027.

Centers for Disease Control and Prevention. (2023, December 12). Coronavirus Disease 2019 (COVID-19). https://www.cdc.gov/coronavirus/2019-ncov/index.html

American Thoracic Society. (2023, December 6). Asthma. https://www.thoracic.org/professionals/clinical-resources/disease-relatedresources/asthma.php

Cystic Fibrosis Foundation. (2023, November 15). What is cystic fibrosis? https://www.cff.org/

Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6 [Computer software]. Department of Genetics, University of Washington.