Some Bioinformatics Studies on SARS-CoV-2

by

Sangita Mitra

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

The ongoing COVID-19 pandemic is impacting the lives of billions of people worldwide as well as the medical and socioeconomic systems. The genomic variability of this virus makes it capable of being prevalent in humans around the world for a long time and migrating from one place to another. It requires a detailed study to understand the trend of SARS-CoV-2 as well as its molecular epidemiology, evolutionary models, and phylogenetic analysis. In this dissertation, we perform several bioinformatics studies on coronaviruses and SARS-CoV-2, focusing on their evolution. The time series analysis on the spike proteins, membrane proteins, and envelope proteins mutations of SARS-CoV-2 are performed to understand how they evolve over time. The spike proteins play a vital role in binding with the human ACE2 receptor. The implication, co-occurrence, and recurrence of spike mutations are investigated. D614G mutation increases infection, and we found in implication analysis that 98% of the time, if D614G mutation occurs, 28 other mutations occur in spike proteins. We got several recurrent mutation pairs in spike proteins that appeared periodically. The relationship of SARS-CoV-2 with two previous outbreaks such as SARS-CoV and MERS-CoV in terms of time series of mutations in spike proteins is analyzed. The mutation rate of six variants of interest and variants of concerns is analyzed to understand the number of mutation change over time. We observed that the COVID-19 pandemic follows some time-series patterns and thus applied the forecasting to predict the upcoming mutations. In this perspective, a prominent long-short term memory network (LSTM) like encoder-decoder LSTM model is applied to predict nucleotide mutations and spike proteins mutations at certain positions of SARS-CoV-2. We propose two bootstrapping techniques as statistical tests to evaluate the model's performance in general and predict each mutation site. The statistical tests show that our model is highly robust in prediction on most sites despite missing data. The results show that the forecasting is more confident in some biologically significant sites than others insignificant sites.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor, Dr. Guohui Lin for his guidance, advice, encouragement and support that have contributed tremendously to my Master's study. Next, I would like to thank all dissertation committee members for their valuable time and effort to read the dissertation. I would also like to thank IST stuff for providing technical support and troubleshooting technical problems. I am very grateful to everyone who has contributed in one way or anther to the completion of my thesis. I sincerely appreciate all their help and support. Finally, I want to give my appreciation to my family for their unconditional love and support. My deepest thanks go to my mother and my sisters for their continuous encouragement, endless love and unwavering support.

Contents

List of Tables

List of Figures

Glossary

ACE2: Angiotensin-Converting Enzyme 2

CFR: Case Fatality Rate

COVID-19: Coronavirus Disease 2019

E: Envelope protein

ER: Endoplasmic Reticulum

LSTM: Long Short Term Memory Network

M: Membrane protein

MAE: Mean Absolute Error

MERS-CoV: Middle East Respiratory Syndrome Coronavirus

MSA: Multiple Sequence Alignment

MSE: Mean Square Error

N: Nucleocapsid protein

NTD: N-Terminal Domain

NSP: Non-structural Protein

ORF: Open Reading Frame

RBD: Receptor Binding Domain

RBM: Receptor Binding Motif

RMSE: Root Mean Square Error

RNN: Recurrent Neural Network

S: Spike protein

SARS-CoV: Severe Acute Respiratory Syndrome Coronavirus

SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2

TMPRSS2: Transmembrane Serine Protease 2

Chapter 1

Introduction

1.1 Coronaviruses Background

Coronaviruses (CoVs) are enveloped, single strand, non-segmented positive-sense RNA viruses which genome size varies in between 26,000 and 32,000 bases [34]. This large group of viruses (order *Nidovirales*, family *Coronaviridae*, subfamily *Coronavirinae*) includes four genera *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus*, and *Deltacoronavirus* [175]. The term 'coronavirus' comes from the fact that under electronic microscopy, the virus looks like a blob surrounded by crown-like spikes, which is called *corona* in Latin [147]. The whole genome contains six to eleven open reading frames (ORFs) [12]. The first ORF encodes 16 non-structural proteins (nsps), which represents approximately 67% of the entire genome, while the remaining ORFs encode accessory proteins and structural proteins [145]. There are sixteen non-structural proteins (nsp1 to nsp16) at the 5' end, and four structural proteins, namely nucleocapsid (N), envelope (E), membrane (M), and spike (S) [60], [78] at the 3' end. The 5 primer and 3 primer end represents the direction of DNA. It refers to the number of carbon atom in a DNA sugar backbone. The envelope of all coronaviruses includes membrane, envelope and spike proteins shown in Figure 1.1. The membrane protein attaches the nucleocapsid protein and enhances viral assembly and budding, the envelope protein is important in viral morphogenesis, release, and pathogenesis, and the spike protein helps the virus attack target cells by contributing to homotrimeric spikes that identify the cell receptor [54], [68], [159].

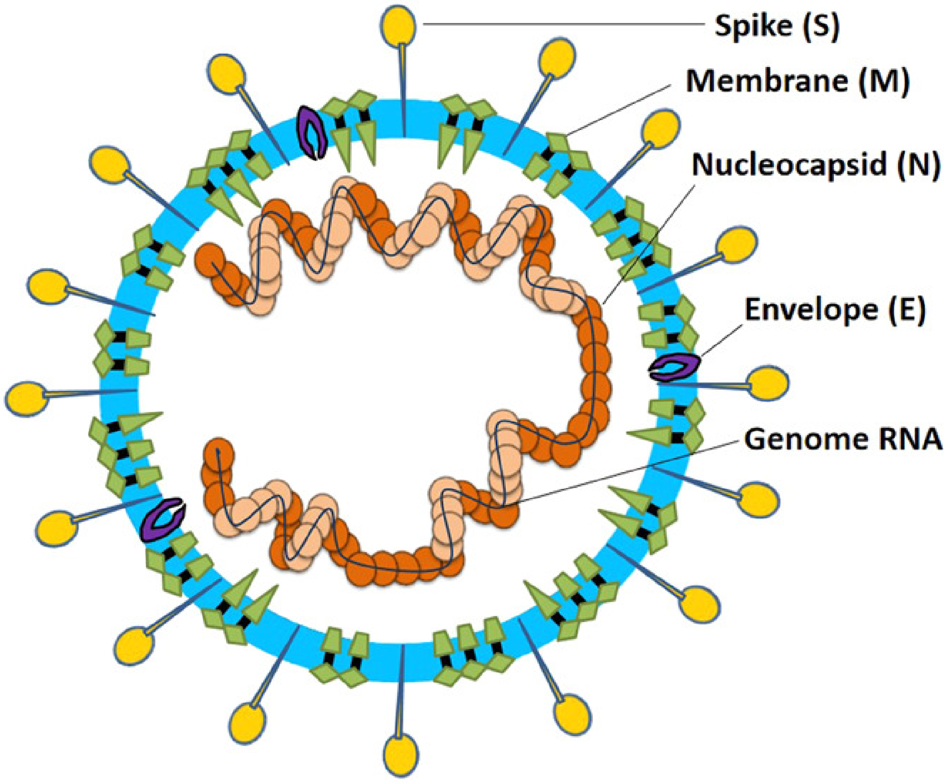


Figure 1.1: Viron structure of coronavirus. [79].

1.2 Motivations and Thesis Objective

The disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is called "coronavirus disease 2019" (abbreviated "COVID-19"). On January 30, 2020, the World Health Organization (WHO) declared the COVID-19 pandemic as a "public health emergency of international concern external icon" (PHEIC). According to WHO, currently (as of August 26, 2021) there have been more than 213 million confirmed cases reported, which have caused more than 4.4 million deaths across the world. More than 4.9 billion vaccine doses have been administrated globally throughout this time. The COVID-19 mortality rate is currently estimated to be close to 6% globally. Among them, older people with health complications are the most vulnerable [9]. Infection with SARS-CoV-2 can be symptomatic with respiratory illness ranging from mild to severe disease and death, or asymptomatic with no symptoms in people [113], [167]. Typical symptoms of SARS-CoV-2 are fever, dry cough, sore throat, chest pain, problems breathing, loss of taste and smell [146], [180], [188]. This virus can be transmitted through direct, indirect, or close contact with infected people while an infected person coughs, sneezes, talks, or sings [86]. Although different protective measures have been taken, including quarantine and isolation of cases and contacts, hand-washing practice, wearing masks, limited public gatherings, etc., to stop the pandemic, we need to reduce the morbidity and mortality of COVID-19 through the vaccine and effective drug treatment [77]. In addition, to understand the implication of the virus, it is important to understand the molecular epidemiology, evolutionary models, and the phylogenetic analysis of SARS-CoV-2 to estimate genetic variability and the evolutionary rate.

SARS-CoV-2 emerged in December 2019, but it's still capable of attacking human cells. This virus is evolving and migrating from one place to another around the world. Following the COVID-19 pandemic, many researchers and scientists are working relentlessly to understand the evolution of the SARS-CoV-2 virus. The complete genome sequence of SARS-COV-2 was published on January 10, 2020 [179]. Previous research discussed common human coronaviruses and their symptoms, death rate, and case fatality rate. We reexamine the evolutionary relationship by performing phylogenetic analysis and multiple sequence alignment to understand how SARS-CoV-2 is related to other coronaviruses. Literature mentioned some possible intermediate hosts of SARS-CoV-2 such as Pangolin, Turtle, and Snake. We performed multiple sequence alignment of the spike protein of pangolin, SARS-CoV-2, and bat coronavirus to confirm the hypothesis of intermediate hosts. The current vaccine and drug development to fight against COVID-19 are targeting spike protein as it plays an important role in virus entry into human cells. Previous research had performed the COVID-19 predictions with a deep learning model to provide a forecast on transmission rate, confirmed case, death case, and recovery case [7], [100]. It is important to analyze the mutations of different proteins of SARS-CoV-2 to understand the pattern of mutations. SARS-CoV-2 produced new variants, and it's important to understand the mutations trend of these variants. The goal of this research is to present a complete picture of the

COVID-19 pandemic. The motivation of these studies is to understanding the coronavirus, evolution, migration, co-relation of mutation and functionality, and time series analysis of mutations. The main objective of this thesis is to perform numerical analysis on mutations of SARS-CoV-2 and map back the result to their biological functionality to identify the important mutations which can be useful for the vaccine and drug development.

In this dissertation, we perform several bioinformatics studies on the mutation change of SARS-CoV-2 to mitigate the gap of current research. For designing proper vaccines and drugs for COVID-19, it is vital to understand the mutation change. We perform the time series analysis of the envelope, membrane, and spike proteins to understand how mutations change with time. The co-occurrence, recurrence, and implication analysis on spike protein, membrane protein, and envelope protein are performed to find the correlation between mutation pairs. The biological function of these mutations helps to understand the evolution of the virus better. We also perform a time series analysis on the spike protein of SARS-CoV and MERS-CoV. This analysis helps to understand the difference in mutation change in SARS-CoV-2 from previous outbreaks. The mutation change with time is also examined for multiple variants. This analysis helps identify which mutations are unique, which mutations are common, and how the number of mutations changes for different variants. We employ a deep learning model with Long Short Term Memory (LSTM) network to predict the upcoming nucleotide and spike mutations at specific positions with the time series forecasting. Here, we predict which mutations could survive in the future in SARS-CoV-2. We develop our research website to present findings from our analysis. We build a database of coronavirus sequences by aggregating data from different virus repositories with additional functionalities.

1.3 Thesis Outline

The dissertation is organized into seven chapters. Chapter 2 contains the description of dataset collection and preparation for different experiments and the overview of our research website. The method of data preparation for different experiments such as mutation analysis of proteins, mutation analysis of variants, and mutation prediction are discussed in detail. We explain the methods and tools for our website and database development. The website serves different functionalities such as performing multiple sequence alignment, phylogenetic analysis with downloaded sequences from a database, data visualization. Chapter 3 reviews the background information on human coronaviruses and the evolutionary relationship of SARS-related coronaviruses. We review previous research on intermediate hosts, vaccine and drugs development, and time series forecasting and present our findings on them. We observe that current vaccine and drug development are targeting the spike protein to develop antibody and immune response. That's why we analyze the spike protein indepth in the following chapters. Chapter 4 presents mutation analysis of spike, membrane, and envelope proteins in terms of time series analysis, implication analysis, co-mutation, and recurred mutation analysis. We find several functional recurrent co-mutation pairs in spike proteins which are defying mutations for different variants of SARS-CoV-2. In Chapter 5, a time series analysis of six variants of SARS-CoV-2 is investigated. We find several new mutations in those variants besides their defying mutations. We analyze the trend of new and unique mutations per day for those variants. In Chapter 6, we present the methods and results of a prominent Encoder-Decoder LSTM network model with time series forecasting for predicting SARS-CoV-2 nucleotide and spike mutations. We propose two statistical tests with bootstrapping to validate the model's performance and each predicted mutation site. We observe that the model provides confident prediction in biologically significant mutation sites than random mutation sites. Finally, in Chapter 7, we summarize our major contributions and discuss some future directions of this research.

Chapter 2

Data Processing and Website Development

In this chapter, we describe the data collection and preparations for different experiments. We prepared 15 datasets in total for various purposes. We present the methods of multiple sequence alignment (MSA) and phylogenetic analysis techniques. We also review our website development process in this chapter. We develop an aggregated database of coronavirus sequences collected from NCBI and GISAID. The website's main objective is to provide a common platform to search sequences, download sequences, and perform MSA and phylogenetic analysis. Users can also check interactive visualizations on our website.

2.1 Dataset Preparation

All coronavirus sequences used in our research were retrieved from NCBI Virus Repository [14] and GISAID [141]. For phylognetic analysis, dataset 1 is created with 40 whole-genome sequences of coronaviruses that contains 8 alpha coronaviruses, 28 beta coronaviruses, and 4 delta coronaviruses strains. All these coronaviruses are responsible for infecting various mammals. This dataset is used to create a phylogenetic tree to understand the evolutionary relations of different coronaviruses. For the phylogenetic analysis of spike protein of coronaviruses, dataset 2 is created with 30 coronavirus strains which contain 10 SARS-CoV strains, 5 Bat SARS related coronavirus strains, 14

SARS-CoV-2 strains, and one pangolin coronavirus strain. This dataset is a subset of dataset 1 as we got the spike protein of 30 coronavirus strains among 40 coronaviruses in dataset 1. For local pairwise alignment, we used closely related coronavirus sequences from the NCBI in our study (Accession ID: SARS-CoV-2: NC_045512.2, SARS-CoV: AY278741.1, MERS-CoV: NC_019843.3, Bat-SL-RaTG13: MN996532, Bat-SL-CoVZC45: MG772933.1, Bat-SL-CoVZXC21: MG772934.1). For multiple sequence alignment (MSA) of membrane protein of SARS-related coronaviruses, 12 coronaviruses membrane protein sequences are reported in dataset 3. For MSA of envelope proteins of SARS related coronaviruses, we created dataset 4 with SARS coronavirus strains and bat coronavirus strains.

For mutation analysis, with a sequence length of 1273, 93613 SARS-CoV-2 spike protein sequences are downloaded from December 2019 to April 2021 form NCBI. Redundant sequences with 100% sequence similarity were removed using Jalview version 2 [166]. All sequences containing unknown amino acid (X) were also removed, and finally, we get 6453 unique spike protein sequences, and all mutations of these sequences are presented in dataset 5. The unique mutations of spike protein sequences from dataset 5 are reported in dataset 6. The individual frequency of each mutation is reported in dataset 7. Dataset 3,4, and 5 are used for time series analysis, implication analysis, co-occurrence and recurrence analysis of spike protein. To perform the time series analysis of spike protein mutations of SARS-CoV and MERS-CoV, we downloaded all available spike protein sequences from NCBI. We followed the same procedure to remove redundant and incomplete sequences and reported unique mutations of SARS-CoV and MERS-CoV in dataset 8 and dataset 9 respectively.

For mutation analysis on membrane protein, we have collected all available membrane protein sequences of SARS-CoV-2 from NCBI from December 2019 to June 2021 with a sequence length of 222. After removing all identical sequences and sequences with unknown amino acids 530 unique membrane protein sequences are found considering YP_009724393 as a reference sequence. All mutations of the membrane protein of SARS-CoV-2 are presented in the dataset 10. The 1st occurrences of each unique mutations are listed on dataset

11.

For mutation analysis on envelope protein, all envelope proteins of SARS-CoV-2 from NCBI are downloaded with a sequence length of 75 from December 2019 to June 2021. After removing all identical sequences with 100% sequence similarity and sequences with unknown amino acids, we got 184 unique envelope protein sequences. All mutations of the envelope protein of SARS-CoV-2 are presented in dataset 12. The 1st occurrences of each mutations in envelope protein are listed on dataset 13.

For mutations forecasting using the LSTM, we pepare two datasets, one for nucleotide mutation prediction and other for spike mutation prediction. The genome sequences are collected from December 2019 to December 2020. The spike sequences are collected from from December 2019 to April 2021. For training the LSTM model, the datasets are then transformed into the one-hot matrix. If a mutation site is present in a particular sequence, then for that sequence collection date, the corresponding cell value will be 1, otherwise 0. To perform time series analysis, 123 nucleotide mutation sites and 103 spike mutation sites are selected in dataset 14 and dataset 15 respectively. Both datasets are then split into a training dataset (70%) on which the model is trained and a testing dataset (30%) on which the model's performance is tested.

2.2 Research Website

We have dedicated a website Coronavirus Evolution to organize our work on coronaviruses where our main focus is SARS-CoV-2. This website contains a large number of nucleotide and protein sequences of SARS-CoV-2 and other coronaviruses from all over the world. We collect whole genome sequences and annotations of SARS-CoV-2 from two available sources, NCBI and GISAID. Nucleotide and protein sequences that have been released till July 20, 2020, in NCBI have displayed on COVID-19 Nucleotide Sequence Data and COVID-19 Protein Sequence Data webpage. Anyone can download this data in FASTA format and perform multiple sequence alignments or phylogenetic analyses by using three Multiple Sequence Alignment (MSA) tools (CLUSTALW, MUSCLE, and MAFFT) that are attached on the webpage. In addition, All Protein Sequence Data contains more than 1,000,000 protein sequences of different species from 1982 to 2020 from the NCBI. All SARS-CoV-2 sequences data till June 30, 2020, from GISAID are also added on the GISAID COVID-19 data webpage. We have also added reference sequences of different coronaviruses in Reference Sequence Data which have been collected from the NCBI. Also, Phylogeny of different proteins of COVID-19 contains the phylogenetic analysis of envelope, spike, nucleocapsid, orf3a, orf6, orf7a, orf7b, orf8, and orf10 protein sequences of SARS-CoV-2, which gives an idea about how these proteins are changing in different regions with time. The Data Visualization page provides a clear view of the current situation of the COVID-19 pandemic in terms of collected strains data around the world.

The website has been upgraded with new functionalities and features to improve the performance as well as the user's experience. The website has been implemented using mainly ReactJs, HTML, CSS, and Javascript for the frontend and Nodemon, MongoDB for the backend. The new website can be accessed at: `https://coronavirus-evolution-website.netlify.app/`. With the help of technology and frameworks, the website's performance has been increased significantly by 35-50% faster than the older version using pure HTML, CSS, JavaScript. At the beginning of the front page is a short description of the research, goal, and team member. With the help of Bootstrap, the website can be implemented in a shorter time and more user friendly. Below that, the front page has been organized, and it could be divided into three main categories, including the nucleotide/protein information page, protein structure page, and visualization page. The current website retrieves all the information of nucleotides and protein sequences from the NCBI and GISAID database. This page also records the previous search sequence for reference and provides an option to download all the selected sequences in FASTA format for further usage. It uses RESTful API to retrieve the sequences from the NCBI server when the users search for a sequence that is more time-efficient.

2.3 Multiple Sequences Alignment and Phylogenetic Analysis

Multiple sequence alignment (MSA) refers to the process of aligning evolutionary related sequences while taking into consideration mutation, insertion, deletion of residues [21]. The MSA of complete genome sequences was performed using MAFFT version 7 [61] with default parameters. When the sequences are highly conserved, we use Multialign [24] as the MSA program, which is a fast and easy tool which represents the MSA as the high consensus residue (red), low consensus residue (blue) and neutral residue (black). The local pairwise alignments of nucleotide and protein sequences are performed using the EMBOSS Water program [80], [128]. The visualization of MSA is done using Jalview version 2.11.1.0 [166], BioEdit version 7.2 [47] sequence alignment editors. The phylogenetic analysis of alpha, beta, and delta coronaviruses is performed using the Neighbour-joining (NJ) methods in MAFFT and validated with 100 bootstraps replications [31] and Poisson substitution model [39]. The phylogenetic tree is visualized using the MEGAX Software[70]. The main advantage of NJ approach is that it is faster than the least squares, maximum parsimony and maximum likelihood methods. NJ is more computationally efficient for analyzing large data sets and for bootstrapping compared to maximum likelihood [135]. That's why we have used a NJ approach to create the phylogenetic tree from alpha, beta, and delta coronavirus strains.

2.4 Conclusions

In this chapter, we present the data preparation scheme and the website development process. We explain the methodology for dataset preparation which we will use in the following chapters. We develop a database of nucleotide and protein sequences of different coronaviruses by collecting data from the NCBI and GISAID virus repository. Some of the functionality of our website are downloading selected sequences, sorting sequences by different features, performing MSA and phylogenetic analysis using online tools, visualizing phylogenetic analysis and time series analysis of different proteins of SARS-CoV-2. Different bioinformatics tools used for multiple sequence alignment and phylogenetic ananlysis in our research.

Chapter 3

Background and Preliminary Analysis

In this chapter, we review the background information of coronavirus and SARS-CoV-2 focusing on their evolution. We analyze the evolutionary relationship of SARS-CoV-2 with other SARS-related coronavirus in terms of their sequence similarity, sequence identity, and phylogenetic analysis. We explore some possible intermediate hosts of SARS-CoV-2. The current vaccine and drug development are targeting the spike protein as it helps the virus to attach with the host cell receptor and the binding affinity of ACE2 and spike protein elevates the human-to-human transmissions. We review previous research on time series forecasting of COVID-19 transmission rate, death rate, confirmed cases, and active cases using machine learning models.

3.1 General Information

Human coronaviruses (HCoVs) are divided into two types: alpha coronaviruses (such as HCoV-229E and HCoV-NL63) and beta coronaviruses (such as HCoV-HKU1, HCoV-OC43, SARS-CoV, MERS-CoV, and SARS-CoV-2) [85]. Besides human coronaviruses, alpha coronaviruses also contain strains from some avian species like porcine epidemic diarrhea virus (PEDV), transmissible gastroenteritis virus (TGEV), mink coronavirus, and turkey coronavirus [94]. The delta coronaviruses are derived from avians and contain sequences of bulbul coronavirus HKU11, sparrow coronavirus, magpie robin coronavirus, and night heron coronavirus HKU19 [158]. A number of human coronaviruses cause major public health issue that has massive human mortality rate. Common symptoms of human coronaviruses are mild to moderate upper-respiratory tract illnesses like the common cold [139]. In the case of Severe Acute Respiratory Syndrome coronavirus (SARS-CoV) and the Middle East Respiratory Syndrome coronavirus (MERS-CoV) [184], the symptoms of headache, cold, fever was observed with severe respiratory illness. SARS-CoV first emerged in Guangdong, China, in November 2002, and MERS-CoV was first detected in the UAE in 2012 [114], [194]. The symptoms of SARS-CoV may include shivering, diarrhea, headache, and fever with respiratory illness in extreme cases [102]. In 2003, more than 8000 cases of SARS-CoV were reported in 26 countries [45]. Since 2004, there are no documented cases of SARS-CoV around the world [55]. The case fatality rate (CFR) is the ratio of the number of dead people to the number of all diagnosed people with the disease over a particular period [1], [191]. The SARS-CoV outbreak in 2003 caused 774 deaths with 8098 confirmed cases worldwide, resulting in a CFR of 9.6% [92]. The outbreak of MERS-CoV reported 791 deaths with 2229 laboratory-confirmed cases globally between April 2012 and October 2018, resulting in a CFR of 35.5% [122]. The CFR for COVID-19 varies from 0.1% to 14.8% in different countries [124], [173].

3.2 Specific Information

3.2.1 SARS-CoV-2 Sequences Similarity

The genome of SARS-CoV-2 is around 29.8 kb nucleotides long, and the genome order is (5' to 3') ORF1ab, spike protein, ORF3a protein, envelope protein, membrane protein, ORF6 protein, ORF7a protein, ORF7b protein, ORF8 protein, nucleocapsid phosphoprotein, and ORF10 protein [63], [130], [169], [171], [193]. The ORF1ab polyprotein breaks down into the following subsections: leader protein, nsp2, nsp3, nsp4, 3C-like proteinase, nsp6, nsp7, nsp8, nsp9, nsp10, RNA-dependent RNA polymerase (RdRp), helicase, 3'-to 5' exonuclease, endoRNAse, 2'-o-ribose methyltransferase, and nsp11 [157], [174].

SARS-CoV-2 sequences are more than 99.9% identical at the whole-genome level despite being collected from different regions around the world (GISAID accession numbers EPI_ISL_402127–402130) [123], [195]. SARS-CoV-2 virus shares about 79% and 50% identity at the whole genome level with SARS-CoV and MERS-CoV respectively [66], [74], [176]. SARS-CoV-2 is 96.2% identical at the whole-genome level to a bat coronavirus, Bat-SL-RatG13, which was previously detected in Rhinolophus Affinis from Yunnan province in 2013 [95]. The sequence similarity gives the measure of resemblance or likeness between two sequences, while the sequence identity gives the number of matching characters between two different sequences.

The phylogenetic analysis Figure 3.1 is performed using sequences of dataset 1 by using a Neighbour-Joining tree with 100 bootstraps using the MAFFT. All SARS-CoV-2 sequences are conserved together in the beta coronaviruses group. All SARS-CoV strains are also conserved together. They are responsible for human infections and close to bat coronaviruses in the phylogenetic tree. All SARS-CoV-2 sequences are conserved together with 99.9% sequence similarity. Bat-SL-RatG13 is the closest coronavirus to SARS-CoV-2 in this tree, and it shows 96% similarity at the whole genome level. Again, SARS-CoV strains are conserved altogether with around 79% sequence similarity. The MERS-CoV is the furthest one from SARS-CoV-2 sequences in the phylogenetic tree, which shows around 50% sequence similarity. Bat-SL-RaTG13 is the most closely related to SARS-CoV-2 strains in the phylogenetic analysis along with the other two closely related Zhoushan bat coronaviruses BAT-SL-CoVZC45 and BAT-SL-CoVZXC21. Bat-SL-CoVZC45 shows almost 88% sequence identity with SARS-CoV-2, while Bat-SL-CoVZXC21 shows 87.6% identity in the pairwise alignment [148].

The spike protein of SARS-CoV-2 exhibits 97.4% sequence identity and 98.4% sequence similarity with Bat-SL-RaTG13, which presents on the same branch with SARS-CoV-2 strains in Figure 3.2 indicates with a brown arrow. The phylogenetic analysis is performed using sequences from dataset 2 using the neighbor-joining tree approach. SARS-CoV and other SARS-related bat coronaviruses are distinct from SARS-CoV-2 in the phylogenetic tree, and they

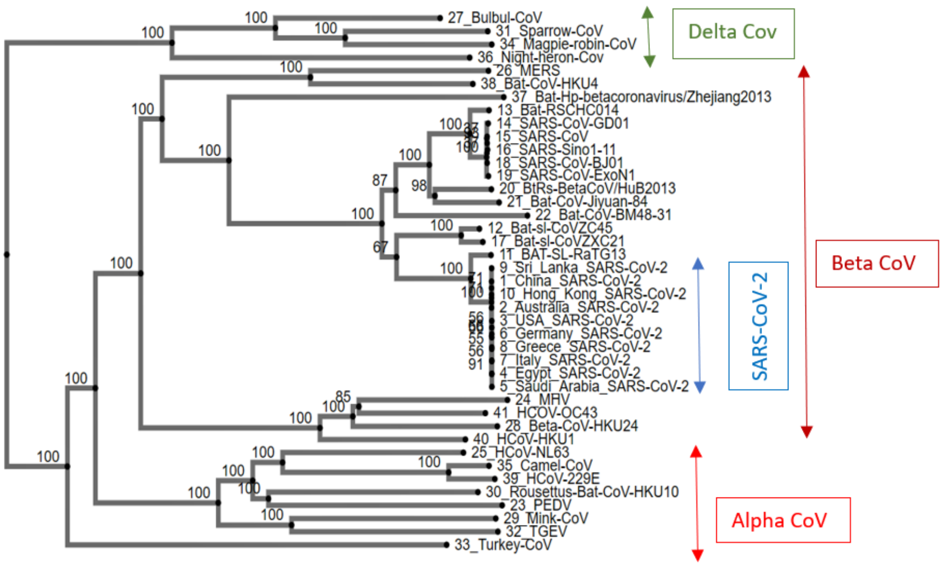


Figure 3.1: Phylogenetic tree of coronaviruses based on nucleotide sequences of complete genomes.

share less than 80% sequence identity [195].

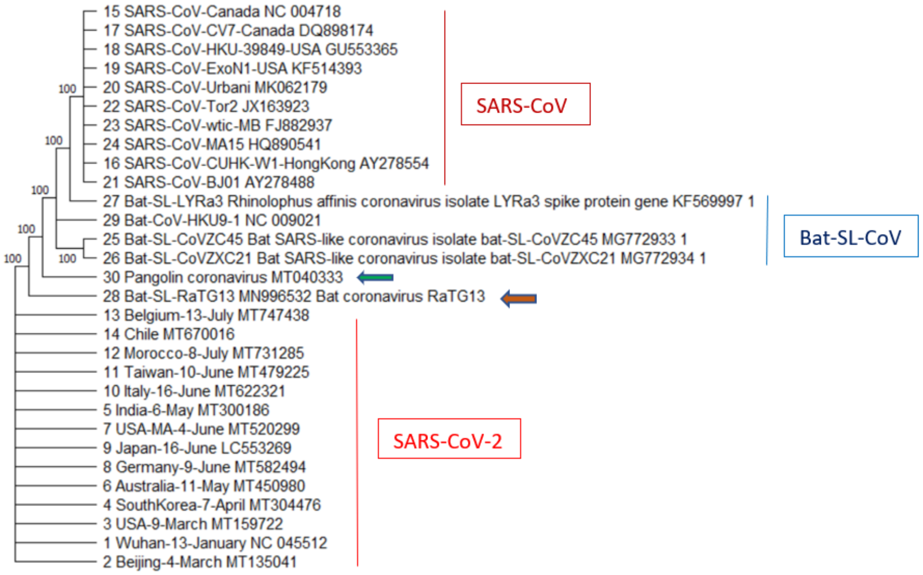

15 SARS-CoV-Canada NC 004718
17 SARS-CoV-CV7-Canada DQ898174
18 SARS-CoV-HKU-39849-USA GU553365
19 SARS-CoV-ExoN1-USA KF514393
20 SARS-CoV-Urbani MK062179
22 SARS-CoV-Tor2 JX163923
23 SARS-CoV-wtic-MB FJ882937
24 SARS-CoV-MA15 HQ890541
16 SARS-CoV-CUHK-W1-HongKong AY278554
21 SARS-CoV-BJ01 AY278488
27 Bat-SL-LYRa3 Rhinolophus affinis coronavirus isolate LYRa3 spike protein gene KF569997 1
29 Bat-CoV-HKU9-1 NC 009021
25 Bat-SL-CoVZC45 Bat SARS-like coronavirus isolate bat-SL-CoVZC45 MG772933 1
26 Bat-SL-CoVZXC21 Bat SARS-like coronavirus isolate bat-SL-CoVZXC21 MG772934 1
30 Pangolin coronavirus MT040333
28 Bat-SL-RaTG13 MN996532 Bat coronavirus RaTG13
13 Belgium-13-July MT747438
14 Chile MT670016
12 Morocco-8-July MT731285
11 Taiwan-10-June MT479225
10 Italy-16-June MT622321
5 India-6-May MT300186
7 USA-MA-4-June MT520299
9 Japan-16-June LC553269
8 Germany-9-June MT582494
6 Australia-11-May MT450980
4 SouthKorea-7-April MT304476
3 USA-9-March MT159722
1 Wuhan-13-January NC 045512
2 Beijing-4-March MT135041
100
100
100
100
100
100
SARS-CoV
Bat-SL-CoV
SARS-CoV-2

Figure 3.2: Phylogenetic tree of spike proteins of SARS-related Bat coronaviruses.

The membrane protein of SARS-CoV-2 shows high similarity (98% identity) with the bat coronaviruses like Bat-SL-RaTG13, Bat-SL-CoVZXC21, and BAT-SL-ZC45 with a few differences in amino acid positions. SARS-CoV-2 shows a 90.2% sequence identity and 96.4% sequence similarity with SARS-CoV, where 14 amino acid differences are observed. The multiple sequence alignment of the membrane protein of dataset 3 in Figure 3.3 contains five bat coronavirus strains, seven SARS-CoV strains, and one SARS-Cov-2 membrane protein sequence. The insertion of an S residue in the 4th position of SARS-CoV-2 is unique, which isn't observed in other SARS-related bat coronaviruses. The MSA shows that membrane proteins of SARS-CoV-ExoN1, SARS-CoV-GD01, SARS-CoV-BJ01 are identical with SARS-CoV membrane protein sequences. BtRS-Beta CoV and SARS-CoV-Sino1 membrane protein sequences show high sequence similarity with two mutations with SARS-CoV, L27I, and F28C, respectively. Other bat coronaviruses also preserve high sequence similarity with SARS-CoV membrane proteins. The MSA is performed using MAFFT and visualized using the Multialign software version 5.4.1 [24].

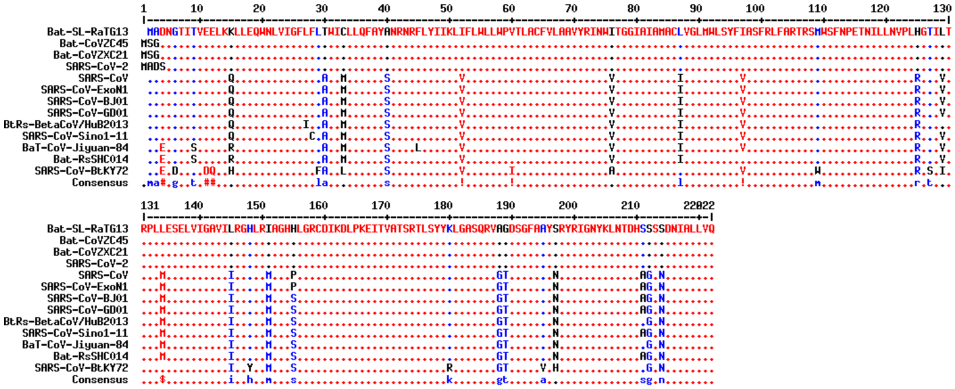
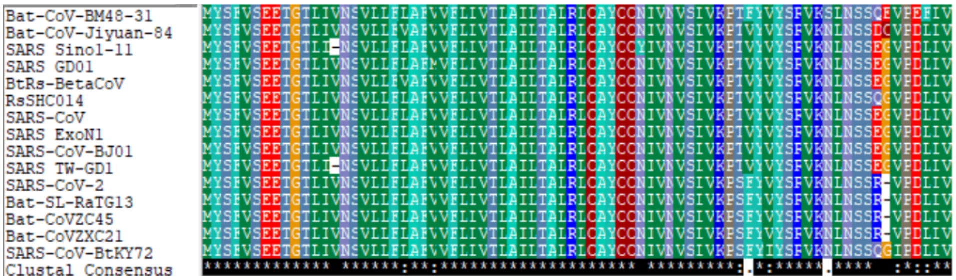


Figure 3.3: MSA of the membrane proteins of SARS-related CoVs.

The hydrophobic envelope protein of SARS-CoV-2 is 75 amino acids long. The SARS-CoV-2 E proteins show high identity with SARS-CoV E protein. The SARS-CoV-2 possesses four mutations with SARS-CoV in terms of the envelope protein, which was responsible for infecting humans, bats, and other mammals. Those mutations are T55S, V56F, E69R, and deletion of G residue at the 70th position in envelope proteins of SARS-CoV-2. SARS-CoV-2 shows 100% similarity in terms of the E protein with two Zhoushan bat coronaviruses BAT-SL-ZC45 and BAT-SL-ZXC21, which are genetically significantly different. The MSA of the envelope protein of SARS-related coronaviruses from dataset 4 shows that the envelope protein of SARS-CoV-2 is highly identical with its closely related coronaviruses. The MSA is performed using the MAFFT sequence alignment program [30] and visualized using BioEdit version 7.2[47].



```
Bat-CoV-BM48-31      MYSFVSEETGTLIVNSVLLFLAFVVFLIVTLAILTALRLCAYCCNIVNVSIVKPTFYVYSFVKSLNSSQEVPEFIV
Bat-CoV-Jiyuan-84    MYSFVSEETGTLIVNSVLLFVAFVVFLIVTLAILTALRLCAYCCNIVNVSIVKPTVYVYSFVKNLNSSDCVPDLIV
SARS Sino1-11        MYSFVSEETGTLI-NSVLLFLAFVVFLIVTLAILTALRLCAYCCYIVNVSIVKPTVYVYSFVKNLNSSEGVPDLIV
SARS GD01            MYSFVSEETGTLIVNSVLLFLAFMVFLIVTLAILTALRLCAYCCNIVNVSIVKPTVYVYSFVKNLNSSEGVPDLIV
BtRs-BetaCoV         MYSFVSEETGTLIVNSVLLFVAFVVFLIVTLAILTALRLCAYCCNIVNVSIVKPTVYVYSFVKNLNSSEGVPDLIV
RsSHC014             MYSFVSEETGTLIVNSVLLFLAFVVFLIVTLAILTALRLCAYCCNIVNVSIVKPTVYVYSFVKNLNSSQGVPDLIV
SARS-CoV             MYSFVSEETGTLIVNSVLLFLAFVVFLIVTLAILTALRLCAYCCNIVNVSIVKPTVYVYSFVKNLNSSEGVPDLIV
SARS ExoN1           MYSFVSEETGTLIVNSVLLFLAFVVFLIVTLAILTALRLCAYCCNIVNVSIVKPTVYVYSFVKNLNSSEGVPDLIV
SARS-CoV-BJ01        MYSFVSEETGTLIVNSVLLFLAFVVFLIVTLAILTALRLCAYCCNIVNVSIVKPTVYVYSFVKNLNSSEGVPDLIV
SARS TW-GD1          MYSFVSEETGTLI-NSVLLFLAFVVFLIVTLAILTALRLCAYCCNIVNVSIVKPTVYVYSFVKNLNSSEGVPDLIV
SARS-CoV-2           MYSFVSEETGTLIVNSVLLFLAFVVFLIVTLAILTALRLCAYCCNIVNVSIVKPSFYVYSFVKNLNSSR-VPDLIV
Bat-SL-RaTG13        MYSFVSEETGTLIVNSVLLFLAFVVFLIVTLAILTALRLCAYCCNIVNVSIVKPSFYVYSFVKNLNSSR-VPDLIV
Bat-CoVZC45          MYSFVSEETGTLIVNSVLLFLAFVVFLIVTLAILTALRLCAYCCNIVNVSIVKPSFYVYSFVKNLNSSR-VPDLIV
Bat-CoVZXC21         MYSFVSEETGTLIVNSVLLFLAFVVFLIVTLAILTALRLCAYCCNIVNVSIVKPSFYVYSFVKNLNSSR-VPDLIV
SARS-CoV-BtKY72      MYSFVSEETGTLIVNSVLLFLAFVVFLIVTLAILTALRLCAYCCNIVNVSIVKPSFYIYSFVKNLNSSQGIPDLIV
Clustal Consensus    ************* ******:**:******************** *********:.*:*****.****  :*::**
```

Figure 3.4: MSA of the envelope proteins of SARS-related bat coronaviruses.

3.2.2 Binding Affinity of ACE2 to Spike Protein

The receptor-binding domain (RBD) and the angiotensin-converting enzyme-2 (ACE2) receptor play an important role in entering into target cells [13], [116], [163]. For inter-species transmission, five residues in the SARS-CoV RBD domain are responsible - Y442, L472, N479, D480, and T487 [81]. These residues have been modified into L455, F486, Q493, and N501 in SARS-CoV-2. K479N and S487T mutations play a vital role in binding the SARS-CoV-2 RBD to the human ACE2 receptor [56]. In SARS-CoV-2 RBD, N479 and T487 of SARS-CoV correspond to Q493 and N501, respectively, which might be responsible for more interaction with the human receptor than SARS-CoV [116]. In addition, the Q493 of the spike protein of SARS-CoV-2 forms a potential hydrogen bond with Q35 of ACE2, which is responsible for strong binding affinity in SARS-CoV-2. In the case of bat coronavirus, the residue corresponding to Q493 is a tyrosine, which is unlikely to create a bond with Q35 of the ACE2 receptor [138]. This indicates that the ACE2 receptor of the SARS-CoV-2 possibly makes SARS-CoV-2 more robust and more stable than other close related bat coronaviruses like RaTG13 [170].

3.2.3 Intermediate Host of SARS-CoV-2

Apart from BAT-SL-RaTG13, the pangolin coronaviruses show high sequence similarity (nearly 96%) with SARS-CoV-2 [168] in spike protein. The pangolin coronavirus in *Manis Javanica* shows a 100% sequence similarity in the envelope protein, and 98% sequence similarity in the membrane protein with SARS-CoV-2 [11]. At the whole genome level, pangolin coronaviruses exhibit higher sequence identity than Bat-SL-RaTG13 (91.02% and 90.55% respectively) with SARS-CoV-2 [192]. Even the sequence similarity of ACE2 in RBD shows higher sequence similarity between pangolins and humans (nearly 85%) rather than in between Bat-SL-RaTG13 and humans (nearly 81%) [73]. Therefore, pangolin is considered as a potential intermediate host of SARS-CoV-2 [48], [82], [87]. Besides this, turtle acts as another virus reservoir that carries a large number of viruses, including ranavirus (RV), nidovirus (NV), papillomavirus (PV), soft-shelled turtle iridovirus (STIV), soft-shelled turtle systemic septicemia spherical virus (STSSSV), tortoise picornavirus (ToPV), and trionyx sinensis hemorrhagic syndrome virus (TSHSV) [57], [96], [111], [189]. There is a possibility that Bat-SL-RaTG13 and other SARS-related coronaviruses infect turtles and might transmit to humans after evolution [89]. Another analysis suggests that besides pangolins and turtles, the snake is a potential intermediate host for SARS-CoV-2, considering its most similar codon usage bias with snake compared to other animals [59].

Our analysis agrees with previous analyses on intermediate hosts of SARS-CoV-2 and justifies that pangolin might be considered as an intermediate host of the 2019 novel coronavirus. It's intimately present with all SARS-CoV-2 strains in the phylogenetic tree of Figure 3.2, which is indicated with a green arrow. With 30 insertions of amino acids, the spike protein of pangolin coronaviruses (*Manis Javanica*) (Accession ID: QIA48632) will be identical with SARS-CoV-2, which is shown in Figure 3.5 with the red arrow. As pangolins and bats both share similar ecological environments, share similar food habits, and both animals are nocturnal, it supports that Pangolin-CoV has a high possibility of being an intermediate host of SARS-CoV-2. Further analysis should be performed to identify all potential intermediate hosts of SARS-CoV-2.

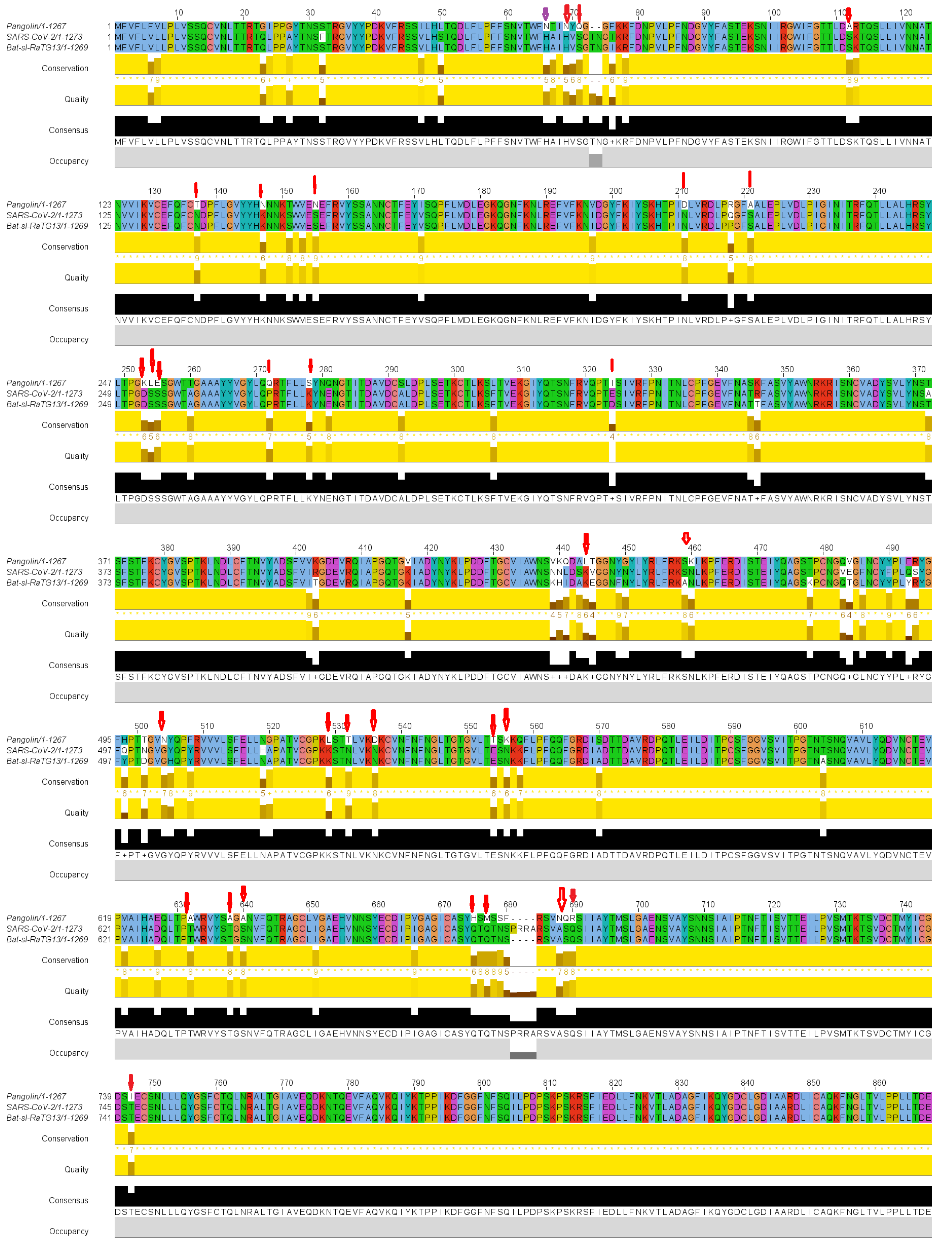


Figure 3.5: MSA of spike proteins of SARS-CoV-2 with pangolin coronaviruses.

3.3 Vaccine and Drug Development

The spike protein plays a vital role since it is immunogenic, and antibodies targeting it can neutralize the virus [71], [185]. The envelope protein is also an attractive vaccine target since the deletion of envelope proteins from coronaviruses shows a mucosal immune response [27]. The envelope protein of SARS-CoV-2 is clustered with known coronavirus envelope protein sequences in the phylogenetic analysis, which supports that by mutating the envelope protein, an E-based vaccine may represent an alternate candidate for SARS-CoV-2 vaccines [125]. At this moment, there are lots of vaccines, and drug development is ongoing[154]. A set of B cell and T cell epitopes derived from the spike and nucleocapsid proteins in the immunogenic structural proteins of SARS-CoV shows an identical map with the SARS-CoV-2 proteins [3]. In these identified epitopes among the available SARS-CoV-2 sequences, no mutation has been observed, suggesting that immune targeting of these epitopes may potentially offer protection against this novel virus. As vaccine candidates are identified, testing the vaccine development on animal models is essential to confirm its effectiveness.

3.3.1 COVID-19 Vaccine Design

As spike protein helps the virus to attach with the human cell receptor and ACE2, it is a prime target for vaccine design. Antibodies targeting spike protein of SARS-CoV-2 can neutralize the infection caused by the virus [88]. Therefore, many vaccines are developed by targeting antigens focusing on spike protein. The number of current COVID-19 vaccine candidates around the world had exceeded 100 [42], [77]. Most of the COVID-19 vaccine developments platforms fall into the following categories: Replicating and non-replicating vectors, Virus-like particles (VLPs), DNA platform, RNA platform, Inactivated, Recombinant protein-based vaccine, and live attenuated vaccines.

As of August 3, 2021, 105 vaccine candidates are active in pipelines, of which 21 vaccine candidates are clinically approved. 76 vaccine candidates are in the clinical phase, and 8 candidates are still in pre-clinical phase. List of authorized vaccine candidates summarized in Table 3.1. The data is collected from COVID-19 Vaccine Tracker.

Vaccine Name	Vaccine Type
Pfizer (BNT162b2)	mRNA-based vaccine
Moderna(mRNA-1273)	mRNA-based vaccine
AstraZeneca (AZD1222)	Adenovirus vaccine
Sputnik V	Recombinant adenovirus vaccine
Sputnik Light	Recombinant adenovirus vaccine (rAd26)
Janssen (JNJ-78436735)	Non-replicating viral vector
CoronaVac	Inactivated vaccine
BBIBP-CorV	Inactivated vaccine
EpiVacCorona	Peptide vaccine
Convidicea (PakVac, Ad5-nCoV)	Recombinant adenovirus vaccine (rAd5)
Covaxin (BBV152)	Inactivated vaccine
WIBP-CorV	Inactivated vaccine
CoviVac	Inactivated vaccine
ZF2001 (ZIFIVAX)	Recombinant vaccine
QazVac (QazCovid-in)	Inactivated vaccine
COVIran Barekat	Inactivated vaccine
Abdala (CIGB 66)	Protein subunit vaccine
Soberana 02	Conjugate vaccine
MVC-COV1901	Protein subunit vaccine

Table 3.1: List of authorized vaccine candidates.

mRNA vaccines

The mRNA is directly injected into the host's cell, where it undergoes cytoplasmic translation. The mRNA molecules contained within lipid nanoparticles express the spike protein of SARS-CoV-2 that facilitates the entry of mRNA into the host cells [35]. When the spike protein exposes inside the host cell, it will induce antibody responses to fight against SARS-CoV-2. There are two types of mRNA-based vaccines: non-amplifying mRNA-based vaccines and self-amplifying mRNA-based vaccines [129]. The self-amplifying mRNA-based vaccine technology is capable of swift and cost-effective vaccines production. It can induce antigen-specific T and B cell immune responses. The mRNA-1273 vaccine is a nucleoside-modified messenger RNA (mRNA)-based vaccine developed by Moderna and NIAID. It encodes the information from the spike protein and directly administers the information to a human. The preliminary findings reported that it generates a significant immunogenic response and induces binding affinity responses to both full-length spike protein and receptor-binding domain in all participants after the first vaccination. The immune responses increase with time and increasing the dose of the vaccine. Moderna and Pfizer are two approved mRNA vaccines, and both vaccines show an acceptable efficacy rate against COVID-19 [19], [99].

Adenovirus vaccines

The University of Oxford and AstraZeneca produces ChAdOx1 nCoV-19 of AZD1222 vaccine by using non replicating simian adenovirus vector ChAdOx1, which contains full-length structural spike protein of SARS-CoV-2. The vaccine encodes the spike proteins by using a replication-deficient ChAd (Chimpanzee Adenoviruses) viral vector isolate Y25, which has been tested in preclinical and clinical trials of influenza A, Ebola, and MERS-CoV [29]. When the vaccine is injected, it starts to produce spike proteins inside human cells. After the vaccine is injected, the adenoviruses bounce into human cells and attach to proteins on their surface. Then the cell wraps the virus in a bubble and pulls it inside. Once inside, the adenovirus escapes from the bubble and pushes its DNA into the nucleus. The body's immune system then reacts and produces antibodies and activates T-cells to destroy cells with the spike protein. Later, when the person is infected with SARS-CoV-2, antibodies and T-cells are triggered to fight against the virus. The preliminary report released by the research group shows that two separate 0.5 mL vaccine doses were safe and tolerated [69]. The second dose of vaccine is given after 4 to 12 weeks of getting the first dose. It takes around two weeks after the second dose of vaccine to develop significant protection against COVID-19. ChAdOx1 nCoV-19 showed an acceptable safety profile, sufficient immunogenicity, and reduced reactogenicity in preventing COVID-19 disease beginning two weeks after the second dose. Another approved adenovirus vaccine is the Johnson & Johnson (Janssen) vaccine. It delivers the virus DNA to human cells to make the spike protein which helps the immune system to create the antibody to protect from COVID-19 infection [90]. It uses a different adenovirus than AstraZeneca, which is called adenovirus 26. For this vaccine, one dose is safe, and after 14 days of vaccination, the immune system develops antibodies to fight against coronavirus spike protein. Both the mRNA vaccines and adenovirus vaccines are effective in producing antibodies against COVID-19.

3.3.2 COVID-19 Drug Design

Several drugs are trialed for the treatment of COVID-19 [72], [106]. Efforts to develop therapeutic drugs will focus on preventing the virus from replicating in the host cell [84]. Currently, Remdesivir shows excellent potential for treating COVID-19, which is approved by the WHO on February 24 [16]. This drug binds to the viral RNA-dependent RNA polymerase and inhibits viral replication, is commonly used as an emergency treatment for COVID-19 [83]. Another promising antiviral drug is Umifenovir, which targets ACE2 interaction with the spike protein and hinders the fusion of the membrane and viral envelope [25], [131]. Lopinavir and Ritonavir are HIV protease inhibitors that exhibit an in-vitro activity against other novel coronaviruses by inhibiting 3-chymotrypsin-like protease [15], [110]. These drugs are not recommended by the NIH and IDSA for the treatment of COVID-19 except in the context of clinical trials [43], [182]. Clinical trials with the nucleotide inhibitor drug Ribavirin, which inhibits viral RNA-dependent RNA polymerase, have also shown promising outcomes against COVID-19 [64]. A database of 78 commonly used antiviral drugs, including those currently on the market and undergoing clinical trials for SARS-CoV-2 was published [172]. Camostat Mesylate can also prevent coronavirus from entering the host cell by blocking the performance of the transmembrane serine protease TMPRSS2 enzyme [53]. It has been tested in the treatment of chronic pancreatitis in the clinic. Therefore it is being considered as a potential therapy for COVID-19 treatment [62]. Chloroquine and Hydroxychloroquine, which were previously used for the treatment of malaria, show promising results for the treatment of COVID-19 Pneumonia [144]. Chloroquine's antiviral and anti-inflammatory characteristics may explain its significant efficacy in treating COVID-19 pneumonia patients by elevating endosomal pH, which is essential for virus/cell fusion [178]. The drug is recommended by the National Health Commission of the People's Republic of China for the treatment of pneumonia caused by COVID-19 [36]. Besides these drugs, antibody-based treatment has the potential to fight against COVID-19 by neutralizing the capability of the virus to infect healthy cells by targeting the spike protein [103]. A clinical trial of 20 severe COVID-19 cases found that tocilizumab, an antibody-based treatment, reduces fever and lung lesion opacity while also increasing the percentage of recovered lymphocytes in blood cells [104]. As the ACE2 receptors help for virus entry of SARS-CoV-2 in human cells, there is a chance that the angiotensin receptor blocker drugs may exacerbate the course of COVID-19 [44]. However, wide clinical experiments should be considered to confirm the efficacy and safety of these drugs.

3.4 Time Series Forecasting on COVID-19

The number of total confirmed COVID-19 cases is increasing with time, and more SARS-CoV-2 genomic sequences have become available. Many researchers are using this data to forecast COVID-19 transmission using various machine learning algorithms [4], [67]. Modeling spatio-temporal sequences with recurrent neural networks (RNN) has shown promising results for time series forecasting [136]. Previous studies implemented the LSTM algorithm to forecast confirmed cases of COVID-19 in Canada and China and achieved good performance [22], [181]. LSTM was first introduces in 1997 [49]. The spread of COVID-19 in India was predicted using a variety of LSTM models [20]. To forecast the percentage of active cases per population in different countries, a comparison of six alternative time-series approaches is discussed [120]. To anticipate future COVID cases and mortality, transfer learning to LSTM network models is used [37]. Other machine learning techniques, such as XGBoost, ARIMA, and the random forest, were utilized to create models that predicted diagnosis, criticality, mortality, and survival of COVID-19 patients [160], [177]. LSTM models were also applied to predict the genome mutation rate of this virus [121]. We believe that LSTM models would be well suited to forecast mutations of COVID-19 and thus employ it to provide time-series forecasts on nucleotide mutations and spike mutations.

3.5 Conclusions

In this chapter, we present related works on different human coronaviruses, the sequence similarity of SARS-CoV-2 with other coronaviruses, intermediate hosts of SARS-CoV-2, the current vaccine and drug development, and the time series forecasting on COVID-19. We perform the multiple sequence alignment and phylogenetic analysis to understand the evolutionary relationship of SARS-related coronaviruses. Our analysis suggests that SARS-CoV-2 is closely related with bat coronaviruses such as Bat-SL-RaTG13, BAT-SL-ZC45, and BAT-SL-ZXC21 in different proteins and whole genome levels. Our analysis supports that Pangoloin coronavirus is an intermediate host of SARS-CoV-2. Although the mechanism is different, both m-RNA and adenovirus vaccines target the spike proteins and produce antibody and immune responses to fight against COVID-19. Both vaccines are now approved and show an acceptable efficacy rate. As the LSTM network shows promising results and works better with long time series data, that's why we use this model for mutation prediction of SARS-CoV-2.

Chapter 4

Mutation Analysis of Spike, Membrane and Envelope Proteins

The objective of this chapter is to present methods and results of different mutation analyses of spike protein, membrane protein, and envelope protein of SARS-CoV-2. For these proteins, we perform the time series analysis, implication analysis, and recurred co-mutation analysis. The time series analysis of mutations shows the trend of mutation for different proteins. The implication analysis identifies mutations which occurrences depend on other mutations. We perform the co-mutation and recurrent co-mutation analysis of proteins to identify the circular mutation pairs which occurred periodically. We identify several functional mutations in spike proteins which can be targeted for the vaccine and drug design.

4.1 Protein Description

4.1.1 Spike Proteins

The spike protein (S) of coronaviruses is critical for the binding with host cell receptors, which facilitates the entry into host cells that elevate the human-to-human transmission rate [93], [118]. The spike protein of SARS-CoV-2 is 1273 amino acids long, which is divided into two domains. The S1 domain contains receptor binding, and the S2 domain contains downstream membrane fusion [97]. The receptor-binding domain (RBD) directly binds the angiotensin-converting enzyme 2 (ACE2) receptor to engage the virus with its target cells [56], [58], [81], [151]. The S1/S2 junction of SARS-CoV-2 is processed by a furin-like protease in the virus's producer cell, while transmembrane serine protease 2 (TMPRSS2) processes the S1/S2 junction at the cell surface or the target cells in SARS-CoV [38], [98], [142]. Analyzing the function and structure of the spike protein is very important in understanding immunogenicity, viral tropism, and the pathogenesis of the virus [187]. The mutations in the receptor-binding domain and S1/S2 regions are essential in the spike proteins as they influence the binding with the human ACE2 receptor [165]. It introduces a new cleavage site, which enhances the host cell entry and is responsible for human-to-human transmission [10]. The receptor-binding domain consists of 333-527 amino acid residues in SARS-CoV-2 spike proteins [150]. In the receptor-binding domain, two important mutations occur - G476S and V483A. The B cell epitope prediction shows that G476S and D614G mutations are present in antigenic peptides "STEIYQAGS" and "NQVAVLYQDVNCTEVPVAIHADQ" respectively, which are antigenic determinants, and that's why these mutations play a vital role in the vaccine and antibody design [134].

We performed the MSA of the spike proteins of SARS-CoV-2, SARS-CoV, SARS-related bat coronaviruses, and pangolin coronaviruses from dataset 2 in Figure 4.1. It exhibits that the receptor-binding domain (RBD) region in the N terminal domain (NTD) is very well conserved, indicated with a black rectangular box. It also contains two important mutation changes between SARS-CoV-2 and SARS-CoV (N479 to Q493 and T487 to N501) mentioned in Section 3.2.2. The receptor-binding motif (RBM) region in the C terminal domain shows more variations illustrated with a red rectangular box. It also has two deletion regions of 5 and 14 residues (position: 453-457 and 481-494) in Bat-SL-CoV-ZC45 and Bat-SL-CoV-ZXC21 in the receptor-binding domain despite their high sequence identity in spike protein, which supports the transmission mechanism difference of these closely related coronaviruses. The S1/S2 cleavage site at R667, indicated with the green arrow, is conserved in all analyzed sequences, followed by a unique insertion of 4 amino acids

"PRRA" (position: 681-684) in SARS-CoV-2. The S2 cleavage site at R797 in SARS-CoV-2 is also well conserved in all other sequences.

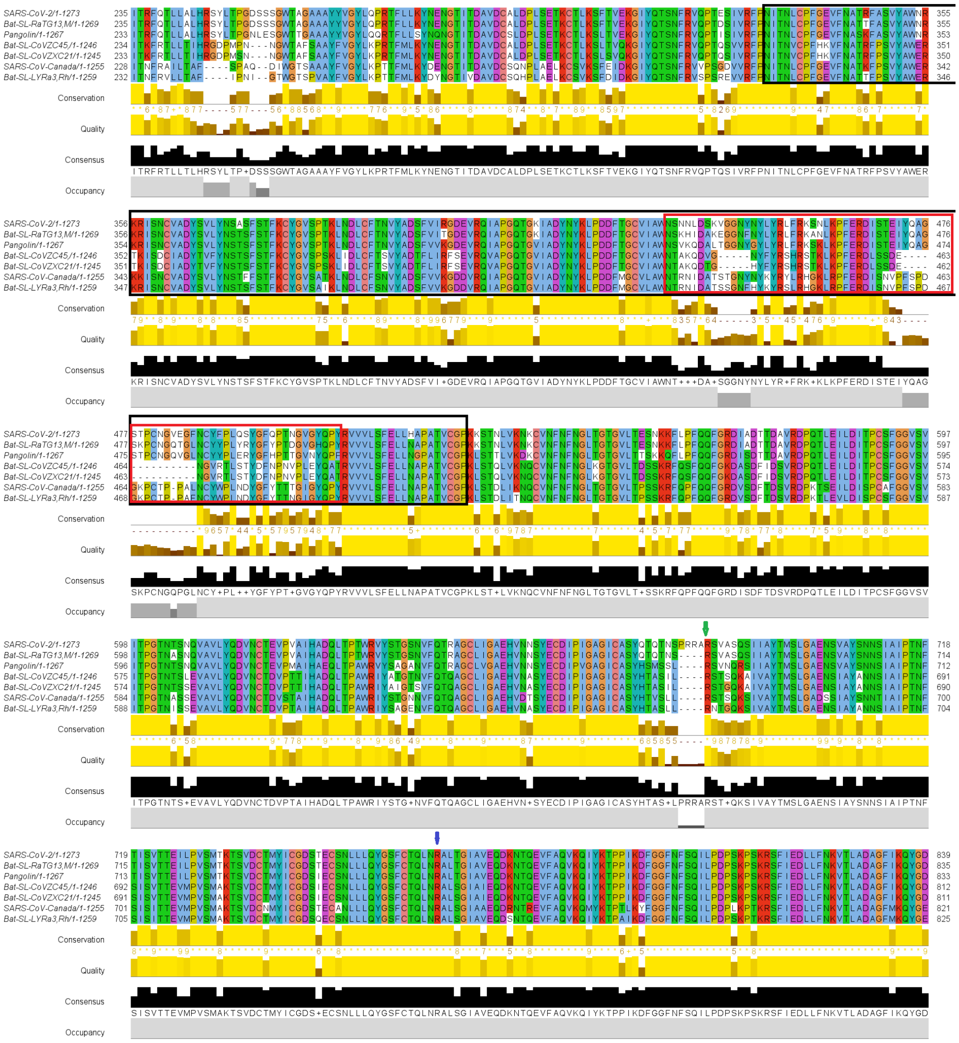


Figure 4.1: MSA of RBD of SARS-related coronaviruses.

4.1.2 Membrane Proteins

The membrane glycoprotein of coronavirus is a type III transmembrane glycoprotein which is the most abundant structural protein. Transmembrane proteins act as gateways, allowing specific materials to pass through the membrane. Based on the position of the N-terminal and C-terminal on different sides of the lipid bilayer, transmembrane glycoprotein can be classified into four types - I, II, III, and IV. In type III transmembrane protein, the N-terminal domains are targeted to the endoplasmic reticulum (ER). Membrane protein plays a vital role in CoV assembly and budding via interactions with the envelope, spike, and nucleocapsid proteins [153]. The M protein of coronavirus is approximately 230 amino acids long and is composed of three parts: a short N-terminal domain, three transmembrane domains, and a carboxy-terminal domain [156]. It is recognized by N-linked glycosylation in alpha and delta coronaviruses, and O-linked glycosylation in the beta coronaviruses [115]. The M glycoprotein extends the membrane bilayer, which oriented the virion with a short NH2-terminal domain and a long COOH terminus [109]. It helps to bind with Nucleocapsid protein which completes the viral assembly inside the virion by stabilizing the N protein-RNA complex [8]. It also determines the shape of the viral envelope and cooperates with S protein which influences the virus budding process [6].

4.1.3 Envelope Proteins

The envelope (E) protein is the smallest structural protein in coronaviruses, which is responsible for creating a viral envelope with the help of the viral membrane protein [132]. The E protein in coronaviruses plays a vital role in the virus life cycle, which is responsible for the infection, replication, assembly, and budding [137]. It has three critical domains: (N)-terminus, transmembrane domain (TMD), and (C)-terminus [137]. E protein affects host immune responses by activating NLRP3 inflammasome in the transmembrane domain and by binding function in C-terminal domain [112]. The transmembrane domain creates a pentameric ion channel, but the C-terminal domain shows no well-defined structure in envelope protein [101].

4.2 Methods

4.2.1 Time Series Analysis of Mutations

For time series analysis, we first downloaded protein spike, membrane and envelope protein sequences from the NCBI virus repository. Then we removed redundant and incomplete protein sequences using Jalview tool. Thus we got unique protein sequences. Then we performed MSA of those sequences by considering Wuhan-Hu-1 as reference sequence. From there we get mutations of each protein sequences using Python scripts. Finally, we plot the mutation positions by collection date of sequences to see the trend of mutations of different proteins.

4.2.2 Implication Analysis of Mutations

The implication analysis shows the probability of occurrence of one mutation given another mutation's occurrence. The implication analysis of spike protein is performed for the 30 most frequent mutation sites from dataset 7, which occurred more than 100 times. The highest occurred mutation site is D614G. From December 2019 to April 2021, mutations L5F, P681H, T95I, L452R, D253G, E484K S13I, and W152C occurred more than 500 times. For all 435 mutation pairs, we calculate how many sequences mutation A occurs, mutation B occurs, mutation A and B co-occur, and none of the mutations occur. Then the co-mutation rate is calculated from the relative frequency. To understand the implication of mutation pairs, the conditional probability of occurrence of mutation A given mutation B and of occurrence of mutation B given A are calculated using equation (4.1) and (4.2).

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{4.1}$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{4.2}$$

4.2.3 Co-Mutation and Recurred Mutations Analysis

The co-mutation and recurrence mutations analyze the occurrence of mutations with time to identify mutation pairs in spike protein that happened periodically. We checked how many months each mutation occurs using data from dataset 5. For 16 months (Jan 2020 - Apr 2021), the frequency of each mutation for each month is calculated by dividing the number of every amino acid mutation for each month by the total number of mutations in that month.

The frequency matrix is used to calculate the pairwise correlation coefficient to determine the correlation between mutation pairs. The correlation coefficient determines the strength of association between data points [2]. In this analysis, Pearson Correlation Coefficient is calculated for all mutation pairs using SciPy's Pearson function. We measured the Phi coefficient for the same mutation pairs to measure the degree of association between mutation pairs. The phi coefficient of mutation pairs A and B is calculated using the equation (4.3).

	Mutation A	**Not Mutation A**	**Total**
Mutation B	a	b	e
Not Mutation B	c	d	f
Total	g	h	n

Table 4.1: Phi Coefficient matrix.

$$\phi = \frac{a * d - b * c}{\sqrt{efgh}} \tag{4.3}$$

The values of the Pearson correlation coefficient and the Phi coefficient range from +1 to -1, with 1 indicating a strong positive association, -1 indicating a strong negative relationship, and 0 indicating no relationship at all. In this analysis, a mutation pair will be identified as recurred mutation if the value of Pearson correlation coefficient and Phi coefficient are greater than 0.7.

4.3 Results and Discussions

4.3.1 Time Series Analysis of Mutations

Spike Protein

For time series analysis on spike protein, the MSA of sequences from dataset 5 is performed using MAFFT [61] using Wuhan-Hu-1, China (NCBI accession code: YP_009724390.1) [174] as the reference sequence. The MSA is visualized here using an online MSA visualization tool Jalview version 2.11.1.0. In total, spike protein has mutated 27,736 positions throughout this time. The time series of spike protein mutations in Figure 4.2 was created using Tableau workbook. It shows the amino acid at the 614 positions has mutated throughout the time period, which is responsible for increasing infectivity in human cells. The mutation D614G in the vicinity of the S1/S2 region is present in more than 96% unique spike protein sequences. In the N-terminus domain, receptor-binding domain, and receptor-binding motif regions, we observed several mutations continuously mutated throughout this time. Some other positions which mutated continuously are 141 and 6, which are in the N-terminus domain.

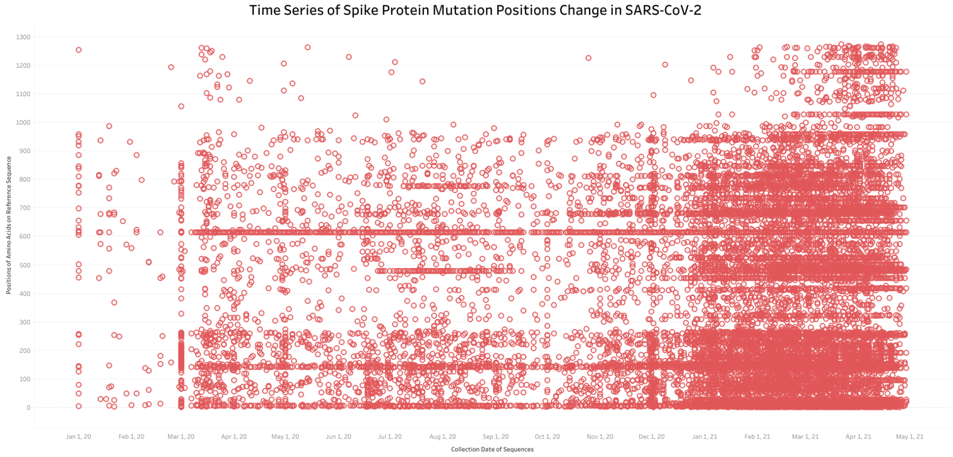


Figure 4.2: Time series of spike proteins mutations of SARS-CoV-2.

The time series of unique mutations of spike proteins from dataset 6 is plotted in Figure 4.3 which shows that the protein has mutated throughout the time period. The number of novel mutations was increasing until September 2020, and then the number went down; however, from January 2021, the number went up again, and till April 2021, the mutations are still present in the spike protein. In the visualization, we see that until March 2020, there was a very low number of mutations. Between March and July 2020, we observed many mutations that mostly occurred in the N-terminus domain, receptor-binding domain, and receptor-binding motif. After July, we see a decrease in the number of unique mutations until Dec 2020, and the increasing trend continues until April 2021. In the receptor-binding domain, two important mutations occur - G476S and V483A. Most of the unique mutations occurred in Australia and the USA.

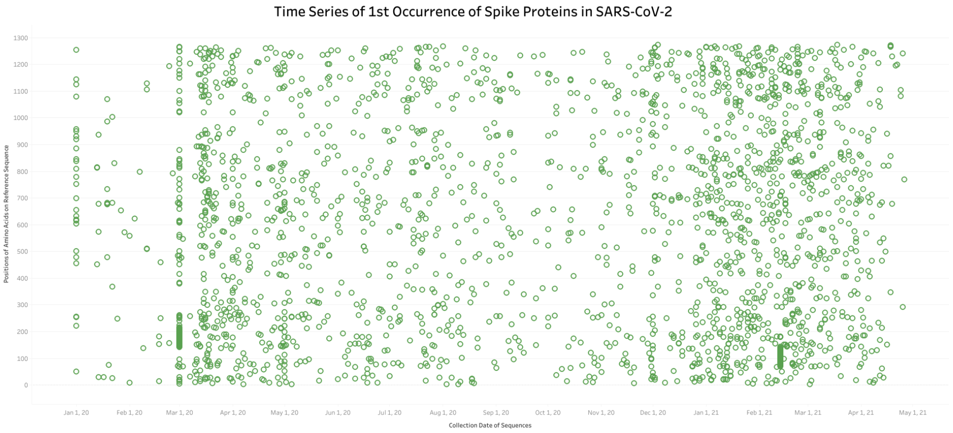


Figure 4.3: Time series of unique mutations in spike proteins of SARS-CoV-2.

To compare the mutation patterns of spike proteins of SARS-CoV-2 with two previous outbreaks - SARS-CoV and MERS-CoV, we have performed the same time series analysis using datasets 8 and 9, respectively. The number of unique mutations in the spike protein of SARS-CoV is 174, and MERS-CoV is 171. It's significantly lower than what we observed in the spike protein of SARS-CoV-2. From Figure 4.4, we see that most of the unique mutations occurred at the beginning of the SARS-CoV outbreak and then went down. We see a similar pattern for MERS-CoV in Figure 4.5 where very few unique mutations were observed.

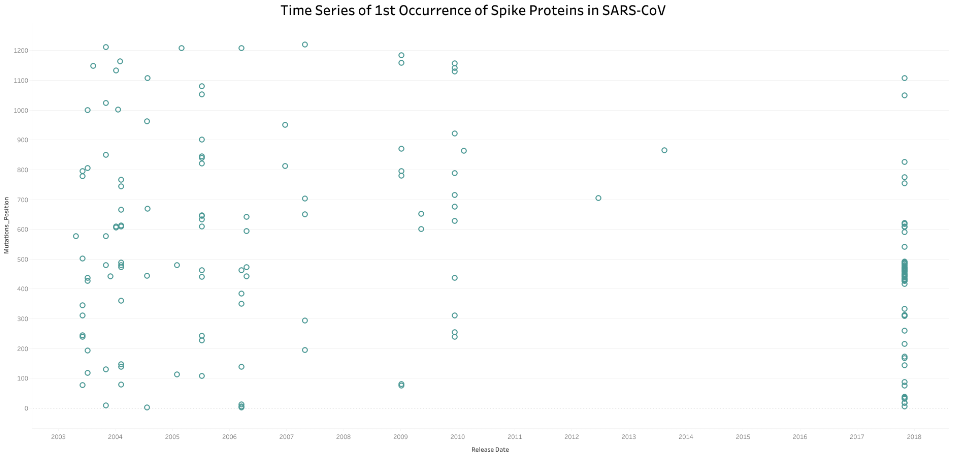


Figure 4.4: Time series of unique mutations in spike proteins of SARS-CoV.

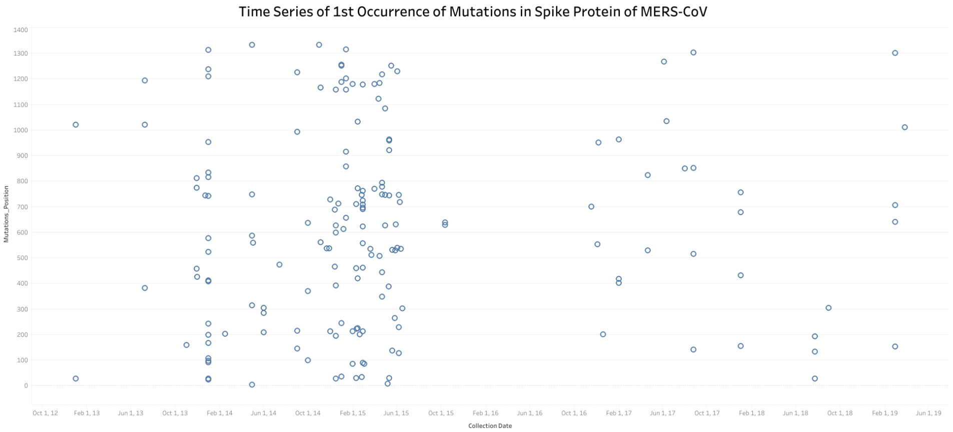


Figure 4.5: Time series of unique mutations in spike proteins of MERS-CoV.

Membrane Protein

The MSA of unique membrane protein sequences from dataset 10 is performed using MAFFT and visualized using Jalview here. The time series of unique membrane protein mutations is visualized using Tableau workbook. We have 387 unique mutations in membrane proteins. Until June 2021, the membrane protein of SARS-CoV-2 has been mutated 704 times and around 92% of mutations occurred in sequences collected from the USA. Besides this, several mutations were observed in Australia, Egypt, India, and Japan. From Figure 4.6, we see that until January 2021, the number of mutations in the membrane was low. After that, we observed many mutations in three transmembrane domains (20-40, 51-71, and 80 -100 residues) which are responsible for efficient interaction with the spike proteins. From Figure 4.7, we get a closer picture of the mutation trend of membrane protein in 2021. The number of mutations increased rapidly in March with 128 mutations, and it reached a peak in April with 153 mutations in total. After that, it shows a downward trend in May and 92 mutations observed in June. The mutation trend in membrane protein follows the COVID-19 case worldwide, as we observed a peak in COVID cases in April 2021.

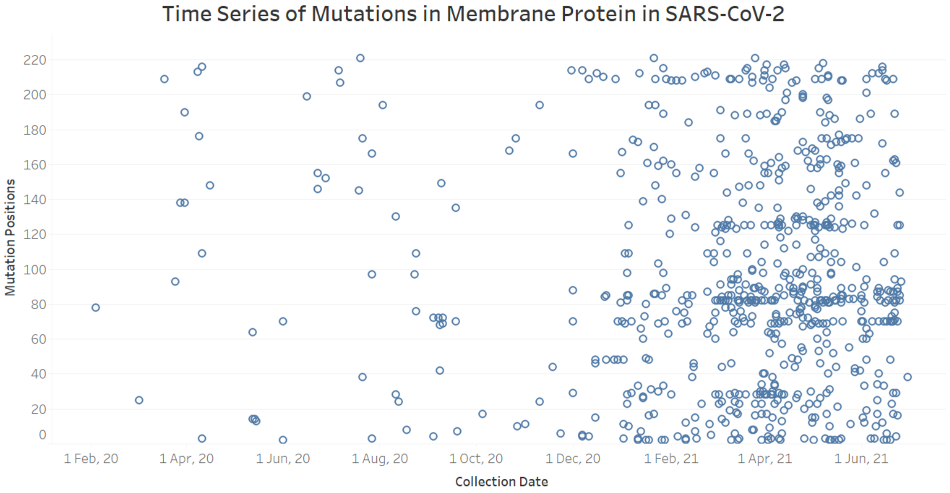


Figure 4.6: Time series of unique mutations of membrane protein of SARS-CoV-2

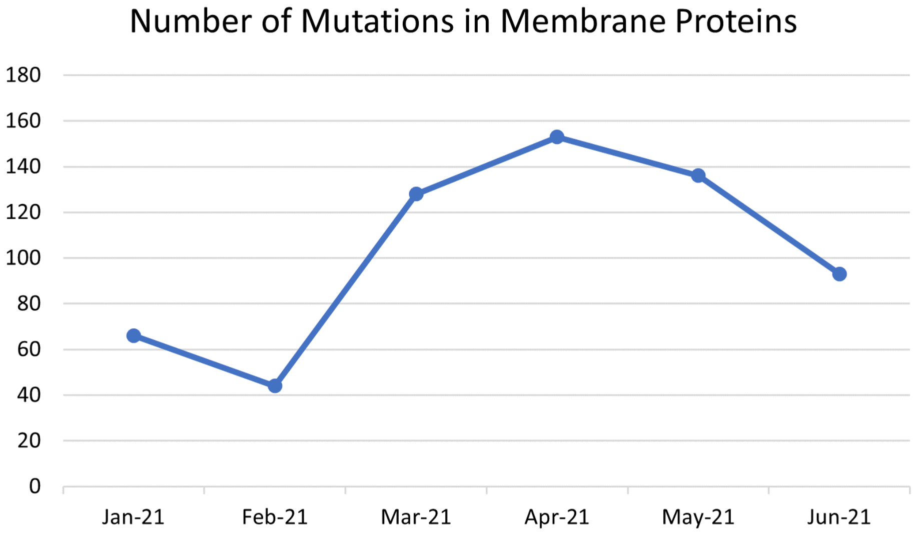


Figure 4.7: The trend of mutations in membrane protein in 2021.

Envelope Protein

The MSA of envelope protein sequences from dataset 12 is performed using MAFFT considering YP_009724393 as a reference sequence and visualized here using Jalview. In our analysis, 163 unique mutations in envelope protein are observed. Until June 2021, the envelope protein of SARS-CoV-2 has been mutated 207 times. The number of mutations in the envelope protein is comparatively lower than the membrane and spike proteins. From Figure 4.8, we observe a low number of mutations in the beginning till January 2021. Like membrane protein, the highest number of mutations is observed in April, with 53 mutations. 87% of total mutations in envelope protein are observed in the USA. In the case of SARS-CoV-2, non-synonymous mutations were observed in the envelope protein. The highest number of mutations in the envelope protein of SARS-CoV-2 occur in the C-terminus domain. Almost all mutations change from hydrophobic to hydrophobic except QKI36831 (D72Y) and QKI36855 (S68C), where the change in the R group is hydrophilic to hydrophobic, which would possibly create changes in protein functions and interactions. Although, these mutations happened only once, and both of them occurred in Guangzhou, China, in late February. Also, these two mutations are not present in other SARS-related coronaviruses, as we see in Figure 3.4.

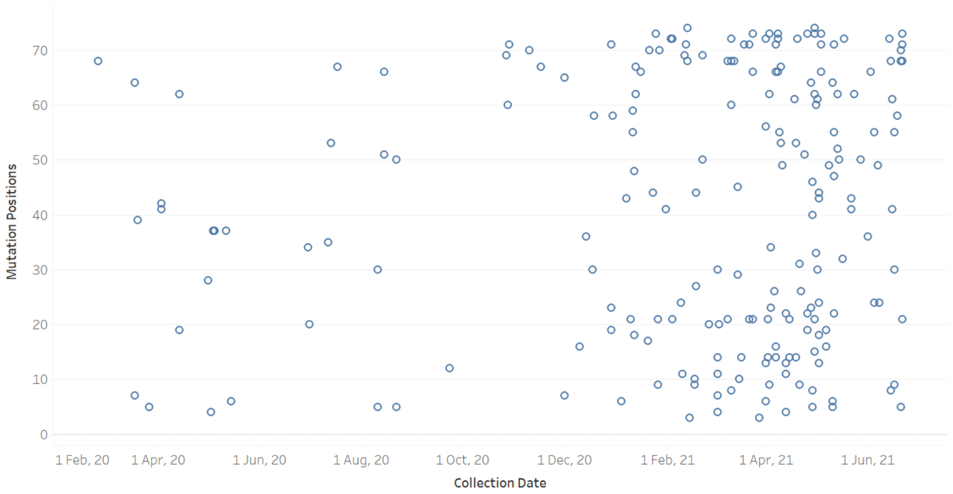


Figure 4.8: Time series of unique mutations of envelope protein of SARS-CoV-2.

4.3.2 Implication Analysis of Mutations of Spike Proteins

The implication analysis shows that 98% of the time if D614G mutation occurs, all 29 other mutations must occur in spike protein. For 14 mutations - S13I, W152C, S477N, T732A, T487K, N501Y, V1176F, P26S, T1027I, Q677P, T20N, S494P, K417T, T716I the probability is 100%. D614G mutation increases infection to human cells by assembling other functional spike proteins into the virion [190]. In our analysis, we got 13 mutation pairs where both $P(A \mid B)$ and $P(A \mid B)$ are greater than 0.7 shown in Table 4.2. Here, mutation pairs T20N, R190S, K417T, H655Y, and T1027I are characteristic mutations of P.1 lineage [33]. Other mutations - S13I, T95I, W152C, D253G, L452R, and T732A are present in the structurally important N-terminal and receptor-binding domains. Our numerical analysis is consistent with the virological data as these mutations present in different variants of concerns. The implication analysis on all 435 mutation pairs is reported in supplementary Table 1.

Mutation A	**Mutation B**	$P(A \mid B)$	$P(B \mid A)$
S13I	W152C	0.98	0.97
T20N	R190S	0.98	0.95
R190S	K417T	0.96	0.95
T20N	K417T	0.98	0.94
T95I	D253G	0.93	0.86
L452R	S13I	0.97	0.85
L452R	W152C	0.98	0.84
T1027I	T20N	0.86	0.81
T1027I	K417T	0.88	0.81
T1027I	R190S	0.86	0.79
H655Y	K417T	0.98	0.70
H655Y	T20N	0.95	0.70
T732A	T478K	0.95	0.70

Table 4.2: Mutation implications in spike proteins.

4.3.3 Recurred Co-Mutations Analysis of Spike Protein

The Pearson C.C. is plotted for three mutations pairs in Figure 4.9 where we notice that the frequency of mutations is identical to each other in each month.

In our analysis, most of the co-mutation and recurrent mutations are in N terminal domain (S1 domain) (1-302 residues), which recognizes carbohydrates to attach the SARS-CoV-2 virus to the host cell [46]. When the result is sorted by the occurrence of both mutations in the sequences at the same time, the top 10 mutations pairs in N terminal domain are L5F, T95I, D253G, S13I, W152C, T20N, R190S, P26S, and D138Y, where both coefficient values are

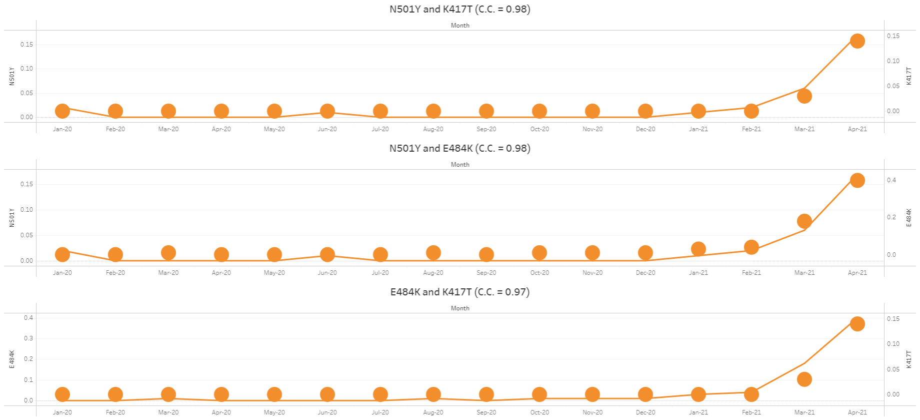


Figure 4.9: Recurrence of spike proteins mutations of SARS-CoV-2.

greater than 0.7 showed in Table 4.3. Here, C_A, C_B and C_AB denotes the occurrence of mutation A, occurrence of mutation B and occurrence of mutation A and B together respectively.

Mutation A	Mutation B	C_A	C_B	C_AB	Pearson	Phi
L5F	T95I	864	639	555	0.98	0.72
T95I	D253G	639	593	550	1.00	0.88
L5F	D253G	864	593	545	0.97	0.73
S13I	W152C	534	527	516	1.00	0.97
P26S	T20N	203	133	130	0.98	0.79
T20N	R190S	133	129	127	1.00	0.97
P26S	R190S	203	129	127	0.98	0.78
D138Y	T20N	183	133	126	0.90	0.80
D138Y	R190S	183	129	124	0.90	0.80

Table 4.3: Recurred mutations of spike proteins in NTD.

Multiple sequence alignment (MSA) of the S1A domain (N terminal domain) is also performed using the MAFFT sequence alignment program [61], visualized using Jalview in BLOSUM62 Score color code [166]. The MSA, ordered by its collection date, confirms the suspected re-occurrence and co-occurrence of the mutations - S13I and W152C; L5F, T95I, and D253G; and L18F, T20N, P26S, D138Y, and R190S. In Figure 4.10, we see these mutation pairs appeared at the same time and then disappeared for some time, then reappeared again.

L5F, T95I, and D253G mutations in the spike protein structure produce

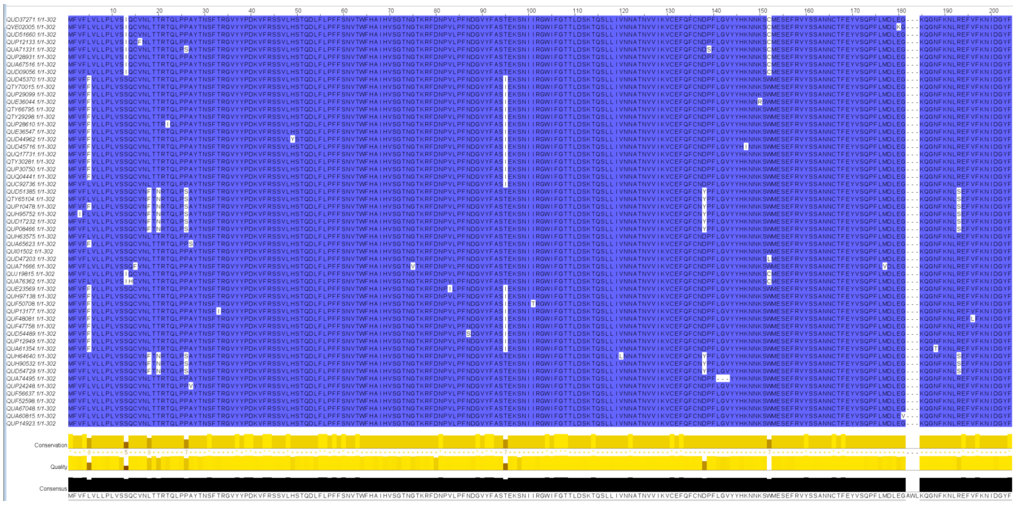


Figure 4.10: Co-mutation and recurrent mutations in spike proteins of SARS-CoV-2.

B.1.526 variants which are most prevalent in New York [196]. Recurred mutation pair S13I and W152C is found in the B.1.429 variant [105].T20N, R190S, P26S, and D138Y are four mutations in the N terminal domain, identified in the P.1 variant [161]. N1 (residues 14–26), N3 (residues 141–156), and N5 (residues 246-260) loops make up the N terminal domain supersite [91]. The T20N, P26S (N1 fold), W152C (N3 fold), and D253G (N5 fold) mutations reside in the N terminal domain supersite, which is a target for neutralizing antibodies [18]. Another recurred mutation pair in the receptor-binding domain (S1B domain) is N501Y and K417T with Pearson coefficient 0.98 and Phi coefficient 0.74, which is also present in the P.1 variant. For recurrent mutation pairs, there could be a migration factor. For example, mutation may be started in one region (e.g., USA) and later found in another region (e.g., Europe). We have checked the location of all recurrent mutation pairs present in Table 2. All sequences containing suspected recurrent mutation pairs are located in the USA except one sequence in the L5F-T95I pair in Bangladesh, and they appeared periodically. Therefore, it gives strong evidence that there is no migration factor in recurrent mutations in spike proteins.

We observed 16 highly conserved sequences in the S1A domain (N terminal domain) in spike protein where its mutations have both Pearson and Phi coefficients as 1. All sequences contain 25 mutations from G142Y to M177E and are collected this year (January 16 - April 20) from different states of the USA. Fourteen of them are from A.2.525 Lineage, which is most common in the USA (61%), and 4 of them are from B.1 lineage (45% from the USA). These sequences have deletions in residues 141-143 (LGV) in the N3 loop of the N terminal domain supersite. The N terminal domain and receptor-binding domain are essential regions in the spike protein. They act as attachment receptors and are targeted by neutralizing antibodies in vaccine design. Our numerical analysis of recurred mutation is consistent with the biological data from the literature, providing strong evidence that other mutation pairs are also functional.

4.3.4 Recurred Co-Mutations Analysis of Membrane and Envelope Protein

The co-mutation and recurrent mutation analysis will help to understand the relations of mutations in membrane and envelope proteins. We checked the individual frequency of each mutation in E and M proteins. The top 5 highest occurred mutations in membrane proteins are I82T, H125Y, F28L, V70F, and A85S are presented in Table 4.4. We observed I82T mutation in 6.4% of the total sequences. This mutation was reported within 44.0% of B.1 lineages. Around 99.7% of B.1.525 lineages carry I82T mutation. We observed 14 sequences with V70L mutations, which are present in B.1.1.7 variants and interact with D178H mutation in spike protein [140]. In envelope protein, the top 5 highest occurred mutations are V62F, L21F, L73F, L21V, and S55F in Table 4.5 and all of them are in the C terminal domain. We observed P71L mutation in envelope protein which is a defying mutation for the B.1.351 variant.

We checked how many months each mutation occurs to identify the recurred mutations. For 17 months (Mar 2020 - Jun 2021), the frequency of each mutation for each month is calculated by dividing the number of times every amino acid mutated for each month by the total number of mutations in that month. By using the frequency matrix, the pairwise correlation coefficient is calculated for all mutation pairs. The Phi coefficient is calculated for the

Mutatios	Count
I82T	34
H125Y	17
F28L	15
V70L	14
A81S	12
V70F	11
A85S	10
D209Y	9
A2S	8

Table 4.4: Highest occurred mutations in membrane protein of SARS-CoV-2.

Mutatios	Count
V62F	5
L21F	5
L73F	5
L21V	4
S55F	4
S68F	4
T30I	3
P71S	3
T9I	3

Table 4.5: Highest occurred mutations in envelope protein of SARS-CoV-2.

same mutation pairs to measure the degree of association between mutation pairs. We used the similar hypothesis of co-mutation and recurrent mutation analysis of spike protein introduced in the Section 5.3.1 to identify recurrent mutations in E and M protein. There are no mutation pairs with Phi and Pearson's coefficients greater than 0.70 for both proteins. Thurs there is no recurrent co-mutation pairs in envelope and membrane protein of SARS-CoV-2.

4.3.5 Discussions

We observed that the number of unique mutations in spike protein of SARS-CoV-2 is higher than the two previous outbreak SARS-CoV and MERS-CoV. We observed most of the mutations in SARS-CoV and MERS-CoV in the beginning of pandemic and then they disappear. That means following health measures we controlled those outbreak and when there was no patients the virus couldn't survived. But for SARS-CoV-2, the spike protein is continuously being mutated which makes it difficult to control the virus and vice versa. We observe that the number of mutations in membrane protein and envelope protein is low compared to spike protein. As spike protein plays the vital role to attach the virus with the host cell receptor and influences human-to-human transmission, it makes sense that the mutations number will be higher in this protein compared to other proteins of SARS-CoV-2. From implication analysis, our findings suggests that that D614G mutation in spike protein influences the occurrence of other frequent mutations. These co-mutated pairs are present in multiple variants. From our analysis, we see that the recurrent co-mutation pairs in spike proteins are present in functional N-terminal and receptor binding domain which are target for neutralizing antibodies. Also, these circular mutations are defying mutations for different variants of SARS-CoV-2. The coefficient values are very low for mutation pairs in membrane and envelope proteins. This indicates that these proteins are genetically drifted but not genetically shifted like spike protein. The spike protein is responsible for viral entry to human cells; this is why vaccine design and drug development target spike protein. Our analysis shows that the SARS-CoV-2 does not need special mutation pairs of envelope and membrane protein for infecting human cells.

4.4 Conclusions

In this chapter, we perform time series analysis, implication analysis, and recurrent co-mutation analysis on three structural proteins of SARS-CoV-2, such as spike, membrane, and envelope proteins. Our analysis suggests that the spike protein of SARS-CoV-2 has been continuously being mutated from the beginning of the pandemic, which is the most critical protein for binding with the host cell receptor. We identify several co-mutation and recurrent co-mutation pairs in spike protein which are present in functional N-terminal domain and receptor-binding domains. We also observed 16 high conserved sequence in the S1A domain where mutations are co-mutated. Vaccine and drug design by targeting these mutation pairs of spike protein will be effective.

Chapter 5

Mutation Analysis of Different Genotypes

The objective of this chapter is to explore different variants of SARS-CoV-2. The mutation occurs in SARS-CoV-2 during replication of genome information [119]. All RNA viruses change with time and make it diverse, which produces new variants [76]. Some variants emerge and disappear, and some variants persist in the environment. These changes may affect the virus's properties, such as how easily it spreads, the associated disease severity, or the performance of vaccines, therapeutic medicines, diagnostic tools, or other public health and social measures. At the time of writing, there are eight variants of SARS-CoV-2 that emerged around the world [65]. WHO declares four variants as variants of concern (VOC) - Alpha, Beta, Gamma, Delta and four variants as variants of interest (VOI) - Eta, Lota, Kappa, Lambda. A variant or genotype is identified as a VOC when there is evidence of transmission, hospitalizations, and deaths rate increase as well as causes a significant reduction in the effectiveness of treatments or vaccines. On the other hand, VOI causes specific genetic mutations which change the receptor binding and are suspected to reduce the efficacy of treatments or predicted increase in transmissibility or disease severity [17]. For designing proper vaccines and drugs for COVID-19, it is vital to understand the mutation change. WHO and international experts monitor the virus change to identify significant amino acid substitution that may produce a new variant. We performed the time series analysis of mutations of variants and find out responsible mutations for producing each variant.

5.1 Methods and Materials

Our analysis is on those viruses which survived. We do not analyze unseen viruses, for which mutations were probably uniform random. We analyzed sequences of six variants from GISAID [141] to identify additional mutations present in all variants, present in some variants, and unique for each variant. For B.1.1.7 (Alpha), B.1.351 (Beta), GR_501Y.V3 (Gamma), B.1.617 (Delta), B.1.525 (Eta), and GH_452R.V1 (Epsilon), we collected one sequence per day till April 2021. Using Coronavirus Genotyping Tool [23], the nucleotide and amino acid mutations are analyzed. Then a tally database is generated using Python script to see if a particular mutation exists for a variant or not. If the mutation exists, it is denoted as 1, otherwise 0. From there, we calculated unique mutation no for each day for each variant to compare the pattern of mutations of variants. The nucleotide mutations in all six variants are C241T, C3037T, C14408T, and A23403G. The only mutation present in the five variants is G11083T. In this analysis, there are 8 NT mutations which present in at least four variants. We continue this analysis to find mutations that present in at least two and three variants. Finally, unique mutations for each variant are generated by comparing all six variants' data. The unique mutations for each variant can be found here.

5.2 Time series of mutations of variants

For the time series of mutation analysis, we have performed three experiments for each variant. The 1st experiment shows how the number of total mutation change with time for each variant. The 2nd experiment shows the number of new mutations per day. Here we compare the number of mutations of each date with the previous date and plot how many new mutations were observed each day. The 3rd experiment shows the number of unique mutations changes with time for each variant. As some mutations are shared between multiple variants, we plotted the number of unique mutations per day to understand the mutation change pattern better.

5.2.1 Alpha Variant

The B.1.1.7 (Alpha) variant is one of four variants of concern announced on December 14, 2020. This variant has initially arisen in the UK and is most prevalent in Europe [108]. Alpha variant increases the transmission rate at least 56% [26] and increase the hospitalizations. This variant also increases the death rate by 35% [5]. In our analysis, we observed defying mutations N501Y, A570D, D614G, P681H, T716I, S982A, D1118H, and deletions in 69, 70, and 144 position in spike protein, D3L, R203K, G204R, S235F in nucleocapsid protein, R52I, Y73C in ORF8 protein, T1001I, A1708D, I2230T, deletion in 3675-3677 position in ORF1a protein [50]. In addition, we observed F138H, M153I, A263P, T286I, A706V, A899S, G1219V in spike protein, T208I in membrane protein, A152V in nucleocapsid protein, many mutations in ORF1a protein including L730F, S693Y, K1817E, T1567A, P748L, Q1009R, L3711F. On average, 38 mutations were observed per day for B.1.1.7 variants. The highest number of mutations per day was noticed in mid-February with 52 mutations which contain the highest 22 new mutations compared to previous sequences. The unique mutation graph shows ups and downs in the number of mutations with time where observed some pick in late December, mid-February, and early March.

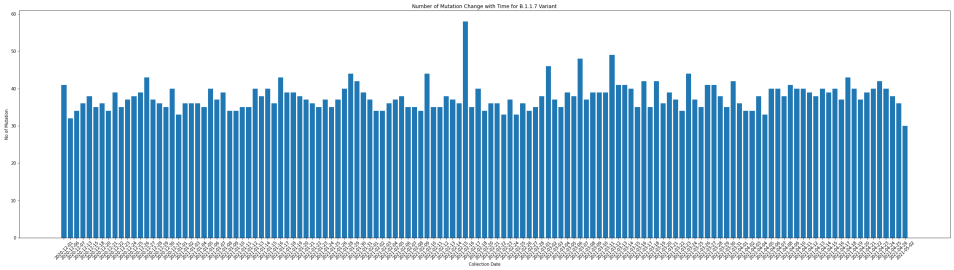


Figure 5.1: Time series of number of mutation change for B.1.1.7 variant

5.2.2 Beta Variant

B.1.351 (Beta) variant was first was originated and initially expanded on South Africa on December 18, 2020 [52]. Previously it was known as the South Africa variant. This variant has also high transmission rate like alpha

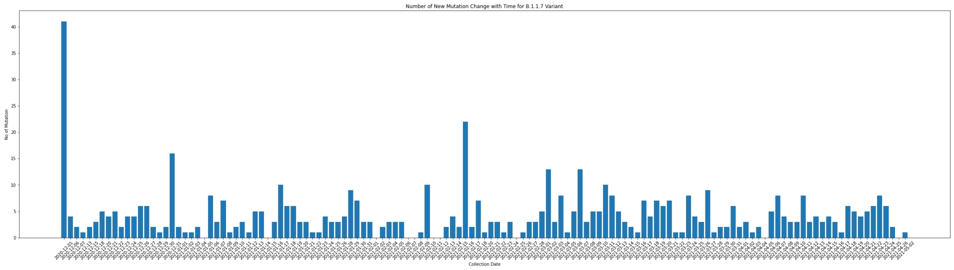


Figure 5.2: Time series of number of new mutation change for B.1.1.7 variant

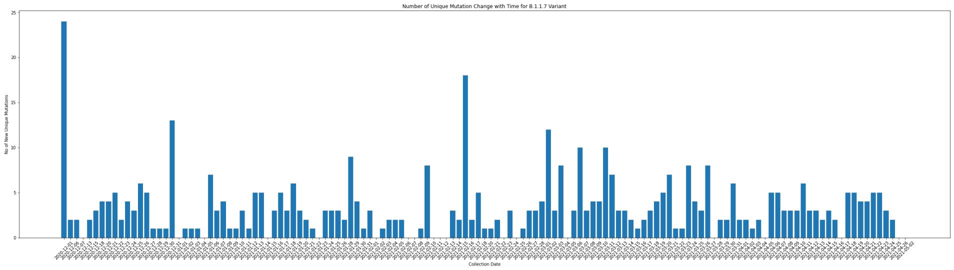


Figure 5.3: Time series of number of unique mutation change for B.1.1.7 variant

variant [162]. It contains multiple mutations in the spike proteins, including N501Y, D614G, E484K shared mutations. Other non-synonymous mutations found in the analysis are D80A, D215G, A701V, K417N mutation in spike protein, P71L mutation in envelope protein, T205I mutation in nucleocapsid protein, Q57H in ORF3a protein, P314L mutation in ORF1ab protein, T265I, K1655N, K3352R mutation in ORF1a protein. Besides, a deletion at 241-243 position in spike protein is reported. Besides, we also observed several other mutations including T19I, N148T, A222V, V308L, L571F, A522S, H655Y in spike protein, P13S, R203K, G204R, Q409L in nucleocapsid protein, L631F, G3704S, T5912L, deletion in 3675-3677 position in ORF1a protein, K16R, P4715l, S171L in ORF3a protein, and I121L in ORF8 protein. On average, this variant has been mutated 29 times per day, with the highest 48 mutations on March 20, 2020. In terms of unique mutations, B.1.351 shows an overall low trend except for a spike in March.

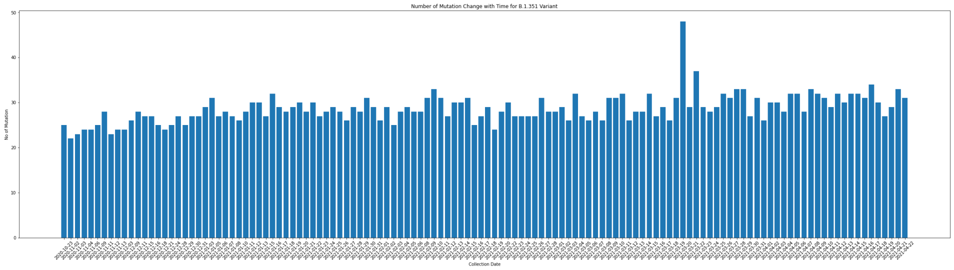


Figure 5.4: Time series of number of mutation change for B.1.351 variant

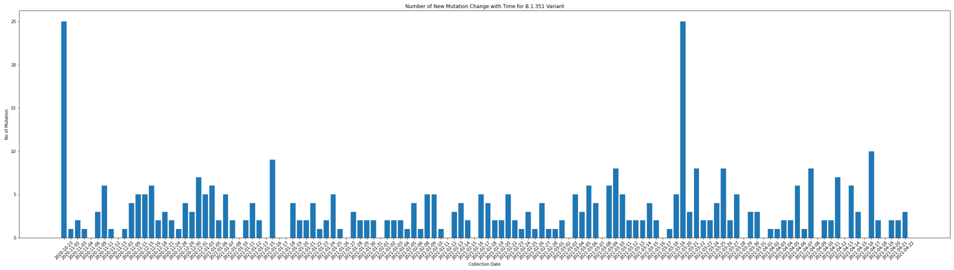


Figure 5.5: Time series of number of new mutation change for B.1.351 variant

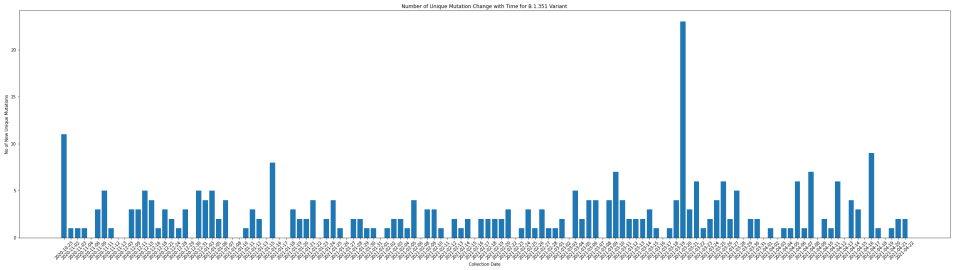


Figure 5.6: Time series of number of unique mutation change for B.1.351 variant

5.2.3 Gamma Variant

GR_501Y.V3 is known as P.1 lineage, which is renamed as Gamma variant by WHO. This variant of concern originated in Brazil and spread to 10 other countries, including the UK. This variant reduced resistance to the combination of bamlanivimab and etesevimab monoclonal antibody treatment [127]. This It contains multiple shared mutations in spike proteins, including N501Y, E484K, and D614G. It also contains multiple non-synonymous mutations in spike protein, including L18F, T20N, P26S, D138Y, R190S, K417T, W152L, H655Y, T1027I, and V1176F. Also, S253P mutation in ORF3a protein, S1188L, K1795Q in ORF1a protein, P80R, R203K, G204R in N protein, and E92K mutation in ORF8 protein are observed in our analysis. We also observed L5F, D80G, W152L, L242F, S256L, L452R, L938F in spike protein, R32H, P279Q, P383L in nucleocapsid protein, L123F, T265I, S528L, I4205V, T4443I, P4715L, D5584Y in ORF1a protein, V32L, G66S, L95F mutation in ORF8 protein. On average, the Gamma variant shows 45 mutations per day. A very low mutation rate is observed in terms of new mutations and unique mutations in this variant.

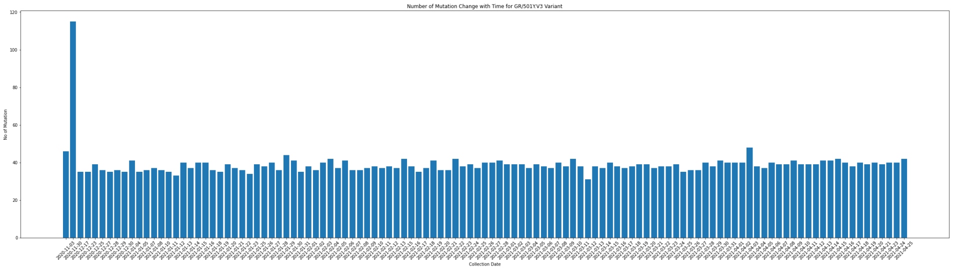


Figure 5.7: Time series of number of mutation change for GR_501.V3 variant

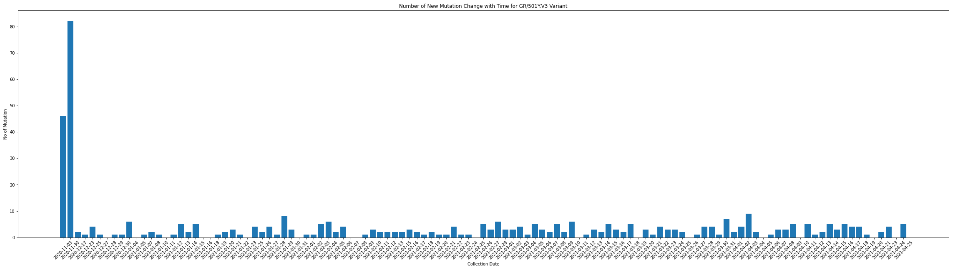


Figure 5.8: Time series of number of new mutation change for GR_501.V3 variant

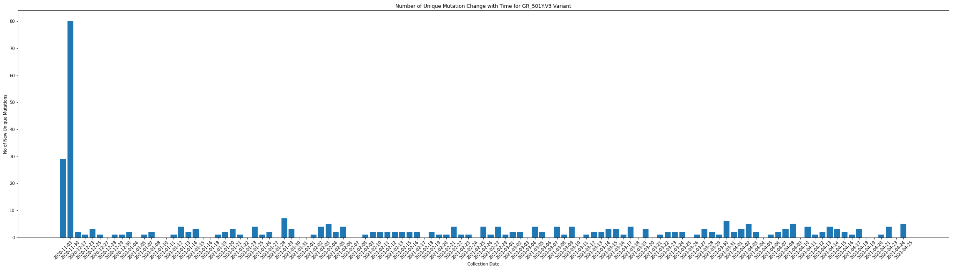


Figure 5.9: Time series of number of unique mutation change for GR_501.V3 variant

5.2.4 Delta Variant

B.1.617.2 and Delta variant is another variant of concern announced by WHO. Delta variant causes more infections and spread the virus more quickly than previous variants of COVID-19 [17]. This variant was first detected in India in late December 2020 and spread in many countries, including the UK. This Delta variant is different than the Delta coronavirus genera. Delta variant contains L452R, D614G, and P681R mutations in spike protein which influences antibody binding. In our analysis, several mutations were observed in the N-terminal domain, receptor-binding domain (RBD), and furin cleavage site of the spike protein, including T95I, E154K, A263E, E484Q, R158G, T478K, D950N, N1187H, which could affect a variety of antibodies. It also contains I82S mutation in membrane protein, D62G, R203M, D377Y mutation in nucleocapsid protein, S26L in ORF3a protein, V82A, and T120I mutation ORF7a protein. Besides these defying mutations, we also observed T95I, E156G , V382L, T478K , E583Q, D950N, V1060L, N1187H, K1245N in spike protein, D3Y, P13L, T141L, G335A, R385K in nucleocapsid protein, L116F in ORF7a protein, and V163L, P240L in ORF3a protein. On average, 33 mutations were observed per day in the Delta variant. The new and unique mutation graph shows an overall low trend with a spike in late March 2021.

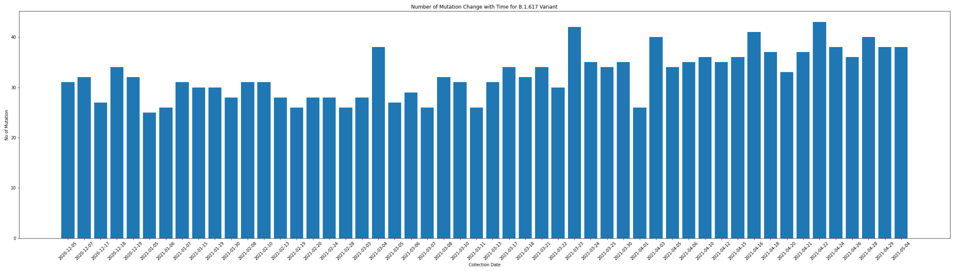


Figure 5.10: Time series of number of mutation change for B.1.617 variant

5.2.5 Eta Variant

B.1.525 and Eta variant is a variant of interest recognized by WHO. This variant was initially detected in late 2020 and spread in North America, Europe, Asia, Africa, and Australia. Eta variant potentially reduce the neutralization

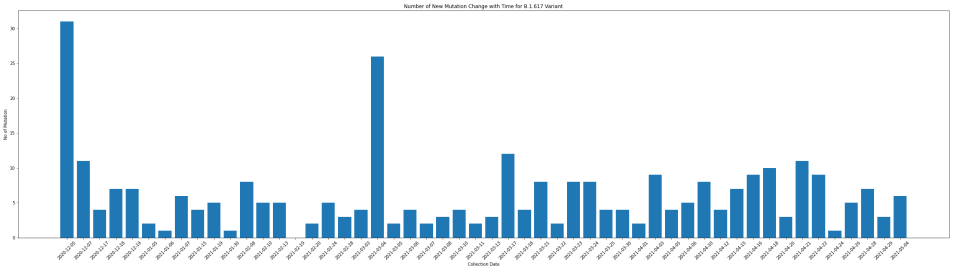


Figure 5.11: Time series of number of new mutation change for B.1.617 variant

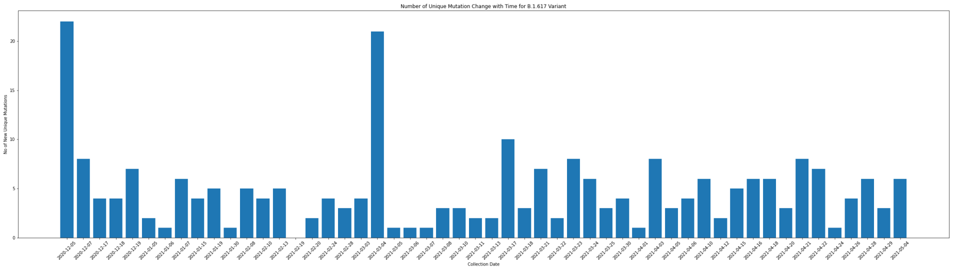


Figure 5.12: Time series of number of unique mutation change for B.1.617 variant

by some Emergency Use Authorization (EUA) monoclonal antibody treatments [17]. This variant contains multiple mutations observed in the variant of concerns like E484K, D614 mutation, deletion in 69, 70, 144 positions in spike protein. Besides, Q52R, A67V, Q677H, F888L mutation in spike protein was observed in this variant. Additionally, it contains I82T mutation in membarne protein, L21F mutation in envelope protein, A12G, T205I mutation in nucleocapsid protein. It has the same three amino acid deletion (3675 -3677) in ORF1a protein which is observed in alpha, beta, and gamma variants. On average, the Eta variant has been mutated 37 times per day. We noticed a spike in new mutations and unique mutation time series in late March.

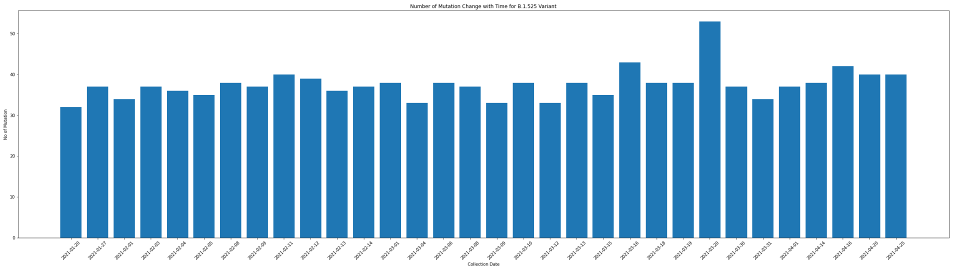


Figure 5.13: Time series of number of mutation change for B.1.525 variant

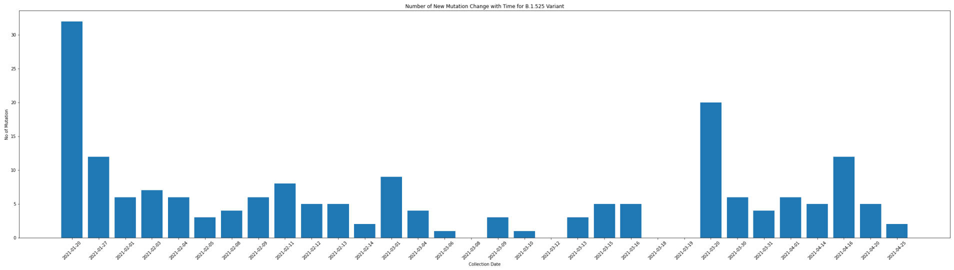


Figure 5.14: Time series of number of new mutation change for B.1.525 variant

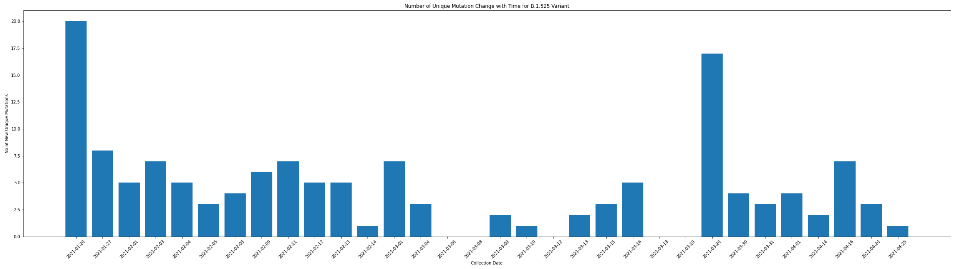


Figure 5.15: Time series of number of unique mutation change for B.1.525 variant

5.2.6 Epsilon Variant

GH_452R.V1 and Epsilon variant is part of the B.1.429/B.1.427 lineage was initially detected in California at the beginning of 2021 and spread in many countries. It contains S13I, W152C mutation in NTD, and L452R mutation in RBD in spike protein. L452R mutation helps the variant to escape the neutralizing activity of monoclonal antibodies in the RBD area. It also increases binding affinity with the ACE2 receptor. Besides, it contains D614G shared mutation in spike protein, I4205V in ORF1a protein, D1183Y in ORF1b protein, Q57H in ORF3a protein, and T205I mutation in nucleocapsid proteins. The average mutation rate of the Epsilon variant is 27. The new mutation trend is lower overall, and the unique mutation trend shows many ups and downs with time.

5.2.7 Discussions

We observed several new mutations for different variants of concerns besides the defying mutations mentioned in literature. Alpha, Beta, Gamma, and Delta variants of concerns are responsible for higher transmission and infec-

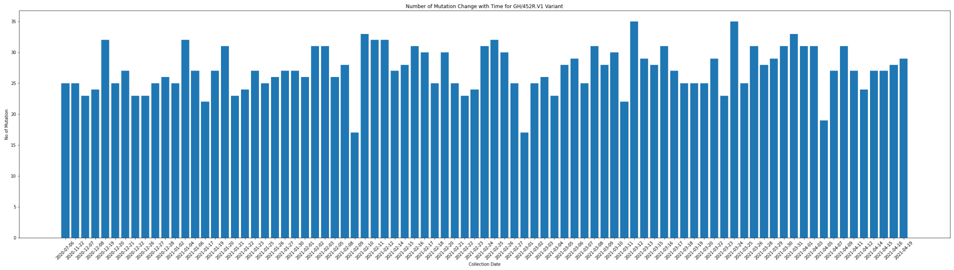


Figure 5.16: Time series of number of mutation change for GH_452R.V1 variant

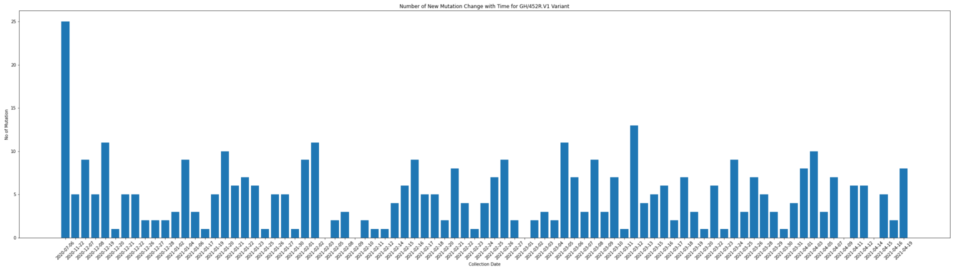


Figure 5.17: Time series of number of new mutation change for GH_452R.V1 variant

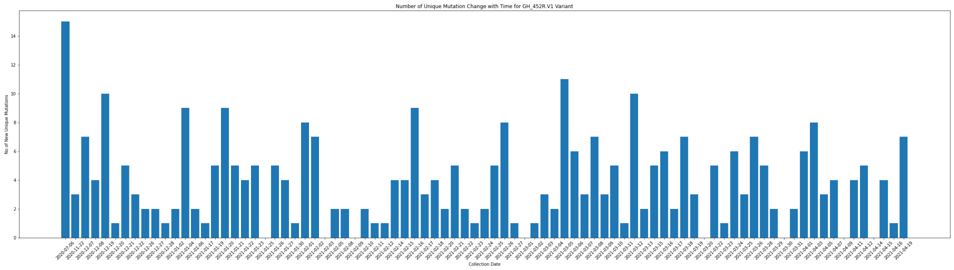


Figure 5.18: Time series of number of unique mutation change for GH_452R.V1 variant

tions rate. On average, the highest number of mutations per day was observed in Alpha and Gamma variants in our analysis. We performed three experiments for each variant to understand the mutation trend of the variants. By comparing the time series of variants, we observe that the highest number of mutations per day for the Gamma variant. Alpha and Delta variants also show an overall high mutation trend per day. Almost all variants show a low trend in terms of time series of unique mutations per day as there are a number of mutations that are shared between multiple variants.

5.3 Conclusions

In this study, we present the time series analysis of six variants of concerns and variants of interests of SARS-CoV-2. We observe that the virus is changing with time, and a set of mutations are producing new variants, which are increasing the transmission rate, death rate and also influencing the effectiveness of vaccines and drugs. Our analysis identifies several new mutations in different proteins for each variant besides the defying mutations. As the SARS-CoV-2 virus is continuously being mutated, we may observe new variants in upcoming days which might be more infectious than current variants of concerns.

Chapter 6

Mutation Prediction with Time Series Forecasting using LSTM Network

The objective of this chapter is to present methods and results of time series forecasting of mutation sites of SARS-CoV-2 using an Encoder-Decoder LSTM network. We first explain the architecture of traditional long-short term memory (LSTM) network and Encoder-Decoder LSTM network and how we are using the model for mutation site prediction. In section 3.4, we see that previous research used deep learning models and LSTM network for time series forecasting of transmission rate, death rate, confirmed cases and so on and achieved promising result. That's why, we provide time series forecasting for nucleotide mutations and spike mutations using LSTM model. As spike protein is a prime target for vaccine and drug design for SARS-CoV-2, that's why we try to predict spike mutations. As a RNA virus, SARS-CoV-2 genome sequence changes with time and it's not possible to predict all those mutations. Here, we predict those mutations which survive for a long time, not random mutations. We propose two statistical tests with bootstrapping to validate the performance of our model and each predicted mutation sites. We demonstrate that the forecasting of biologically significant mutation sites is more confident than random mutation sites.

6.1 Methods

6.1.1 LSTM Network

Time series forecasting is a forecasting process that uses the underlying relationship of historical data to develop a model. In time series analysis, the data points are ordered with time. The model is then used to provide future predictions using deep learning methods. It predicts future data based on the trends of past data. Here, the data points are recorded with time intervals, and the goal is to predict mutations of SARS-CoV-2 virus strains. As one of the recurrent neural networks (RNN), LSTM performs better time-series tasks with large data input and long-term dependencies. It improves the shortcomings, gradient vanishing, and exploding of traditional RNN to a great extent [49]. LSTM models are used in classification, prediction, making decisions based on time-series data. Like the standard RNN model, LSTM has an input, hidden, and output layer. The LSTM network model contains four interacting layers as shown in Figure 6.1.

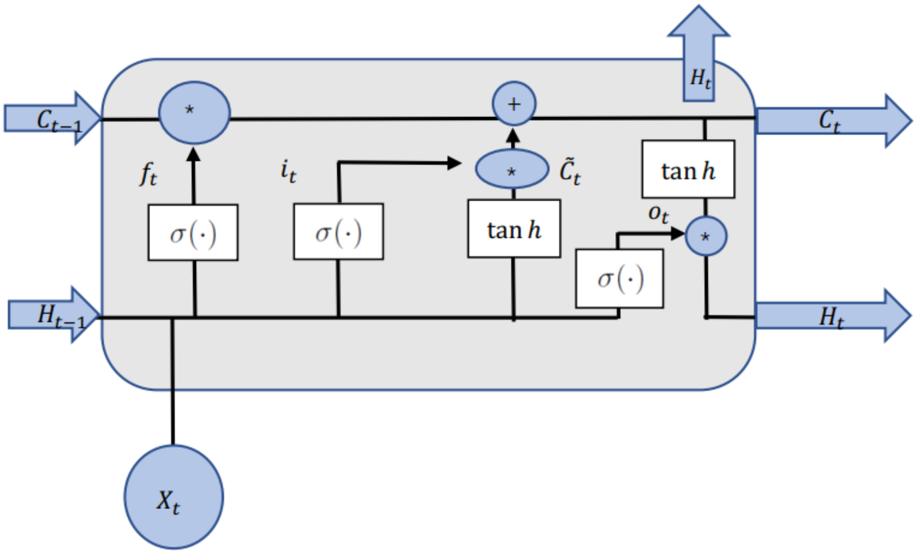


Figure 6.1: LSTM network architecture.

Forget gate layer, input gate layer, current state layer, and output layer make up a simple LSTM structure. Each layer has its own associated weight and plays a different role in time-series data processing. Unlike standard RNN, a forget layer in the LSTM model helps extract information from the data that matters the most in the recent timestep. Information that tends to contribute less to prediction will be assigned less weight and thus has less importance in producing output. A forget layer takes the input from the current time step and output from the previous timestep. The input gate layer is inherited from standard RNN and combines information streams in a time-series order. It also takes the output from the previous timestep and input of the current timestep, using the non-linear function to produce two results. Different gates in LSTM help to process information by using a sigmoid activation function where output is either 0 or 1. The gates block everything if the output is 0; otherwise, gates allow everything to pass through it. The forget gate, input gate and output gate of LSTM are calculated using equations 6.1, 6.2, 6.3.

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \tag{6.1}$$

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{6.2}$$

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{6.3}$$

$\sigma(.)$ is the sigmoid function, W_x is the relevant weight matrix. Specifically, h_{t-1} is output from the previous mutation date. x_t is input of current date's mutation information, as a one-hot vector. b_x is a biased term at gate x. The cell state, candidate cell state and the final output are calculated using the following equations 6.4, 6.5, and 6.6.

$$\tilde{C}_t = \tanh(W_C.[h_{t-1}, x_t] + b_C) \tag{6.4}$$

$$C_t = f_t * C_{t-1} + i_t * C_t \tag{6.5}$$

$$h_t = o_t * \tanh(C_t) \tag{6.6}$$

The current state layer takes the output from forget Gate and input Gate, producing the result that can be used in the output layer to generate output for the current timestep. Output from the current timestep will be passed on and then used as the input in calculations of the next timestep.

6.1.2 Encoder Decoder LSTM Networks

We use an encoder-decoder stacked sequences to sequences LSTM model to predict which mutations could survive in SARS-CoV-2. The model maps a fixed-length input vector to a fixed-length output vector [149]. The multivariate multi-step seq2seq model for time series forecasting contains two LSTMs - encoder and decoder shown in Figure 6.2. The encoder turns the input sequence into a context vector, which is a fixed-length vector. To forecast the output sequence, the decoder uses the context vector as input and the final encoder state as input decoder state. Given an input sequence $(i_1, i_2, ..., i_n)$, the conditional probability of the output sequence $(o_1, o_2, ..., o_n)$ is estimated. The design includes repeated vector layers and time-distributed dense layers to give a multivariate multi-step time series forecast. The context vector is repeated by the repeat vectors, which are then passed to the decoder as an input. We'll repeat the process n-times, where n is the number of future steps. On each time step, the time distribution will apply a fully connected dense layer and separate the output of each time step. All outputs from the preceding layer are fed to all of the neurons in a dense layer, with each neuron providing one output to the next layer. The model is optimized with Adam optimizer and the loss function is Huber loss.

To perform a time-series analysis, we used time series data from dataset 14 and dataset 15 with the sliding window techniques. Every past 90 days of data will be used as an input to predict the mutations in the next 7 days. Data were transformed into input samples based on a 97 days of the time window and trained the model to minimize the absolute errors on every future 7 days' prediction. We reshaped the input data into the form of a 3D tensor to feed into the networks. We used one encoder layer and one decode layer to train the model. For input, the layer takes sequences of the past 90 days with 123

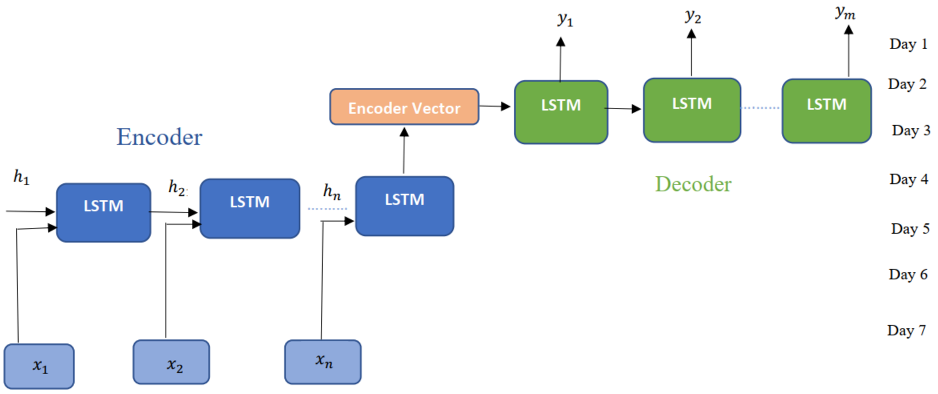


Figure 6.2: Encode-Decoder LSTM network.

different mutation sites for nucleotide mutation prediction and 103 different mutation sites for spike protein mutation prediction. For the output layer, a sigmoid function is used, which converts the calculated results of each site into a probability between 0 to 1, which represents the probability of mutation occurrence on the corresponding site. The loss function is calculated using equation 6.7.

$$y_{loss} = y_{true} - y_{predict} \tag{6.7}$$

6.1.3 Statistical Tests

It is difficult to validate the result of complex multi-layer neural network. One of the most important methods to interpret and validate the correctness of the model is to conduct various statistical tests on models with certain data sets. Due to the complexity of this neural network structure and the nature of Time-Series problems, evaluating the performance purely from predicted results is challenging and not reliable. Thus, we came up with different test methods to quantify performance for the model in general and each specific predicted site.

Bootstrapping is the primary approach we exploited in the test. Two different bootstrapping methods are used for different purposes to test the correctness and robustness of our model. The first test is conducted under the null hypothesis that the model, in general, is not performing significantly better than random prediction. In this experiment, we randomly shuffled the training dataset over its timestamp, without changing the labels (1 or 0), to break the connection between time sequence and mutation occurrence. We then used a trained model on the same test data, recording predicted values for each site and calculating the mean-squared error for each site and the whole model. This process was repeated three hundred times, and three hundred MSE was recorded to be compared with the MSE produced by the model with non-shuffled data. The calculated p-value represents the model's probability with the non-shuffled dataset outperformed by the shuffled dataset result. Here, 0 is a strong indication of better performance for non-shuffled results, while 1 is otherwise. The threshold for rejection is set between 0.0 - 0.05. The second test is under the null hypothesis that this model will not perform better if particular sites are missing and have incomplete data. In these experiments, we removed ten random mutation sites each time from the training set and trained the model with the incomplete dataset. We predicted on the same test data and recorded MSE as before, repeating the process three hundred times. The calculated p-value has the same indication as to the previous one. However, we proposed values between 0.1 - 0.7 as being robust, at which regions model with incomplete data is not significantly outperformed by the model without removing sites, meaning removing data would not affect the prediction to a great extent.

6.2 Results and Discussions

6.2.1 Prediction Performance

The model uses stochastic gradient descent to optimize the total risk by minimizing the sum of loss for every sample point. We trained the models with a batch size of 32 and an epoch of 50 after validation. For spike mutation prediction, the MSE on train data is 0.01142. We evaluate the performance of respective models with mean absolute error (MAE) and root means square error (RMSE) using equation 6.8 and 6.9.

$$MAE = \frac{1}{N}\sum_{j=1}^{N}|y_t - \hat{y_p}| \tag{6.8}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_t - \hat{y_p})^2} \tag{6.9}$$

Where y_t and y_p are the true data and predicted data, respectively. The MAE and RMSE for nucleotide mutation prediction are 0.03838 and 0.15445. For spike mutation prediction, the best MAE and RMSE are 0.01340 and 0.10422, respectively. Despite the low mean squared error of the entire model, we observed that specific sites have very high accuracy and some others have relatively low accuracy with the testing data. Based on the prediction performed on the testing set with the trained model, we calculated the prediction confidence interval for each mutation site to claim the confidence for the future forecast. We observed that areas with consistent mutation occurrence tend to have higher accuracy while sites with random mutation appearance usually have higher errors when predicted. For instance, site C241T, which has a consistent occurrence in mutation regardless of the timestep, has a reasonably low mean error when tested. This gives significant insight into the application of this experiment as to which site can be trusted in future prediction.

6.2.2 Nucleotide Mutation Prediction

For the nucleotide mutation prediction, both bootstrap methods discussed in Section 6.1.3 show P<.001 for the entire model. We thus rejected the null hypothesis for the entire model and claimed that our model is robust against random prediction. It indicates that all mutation sites are essential for predicting. We calculated the same p-value for each site and visualized in Figure 6.3 using Tableau workbook.

The p_value_m1 and p_value_m2 indicate the confidence of prediction of each mutation site in two bootstrap methods. Here, we indicate the confident mutation sites with arrow. We map the p-value of mutations with their biological significance. The LSTM model gives a low p-value for many biologically

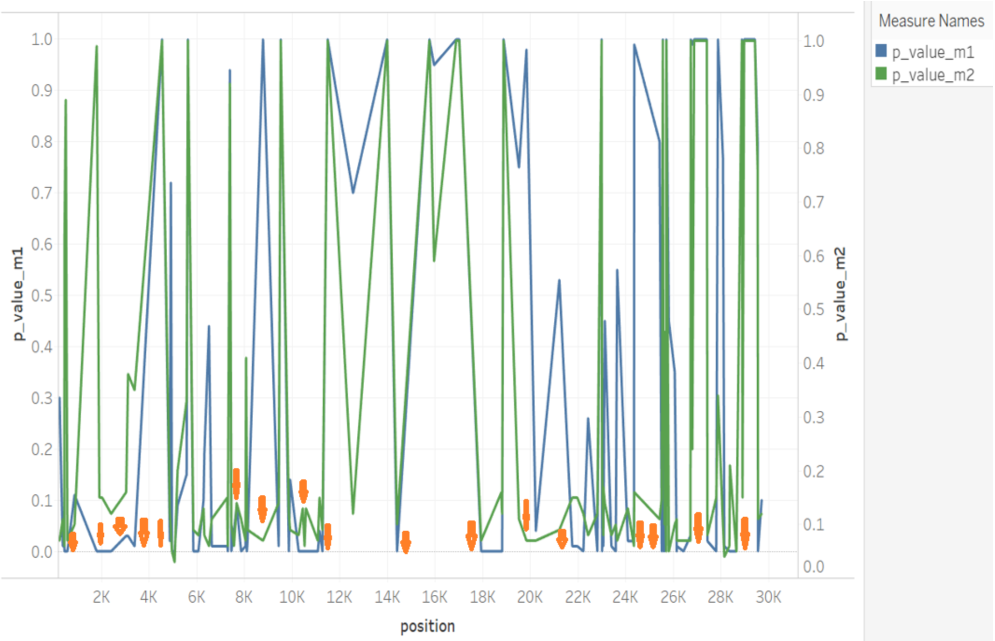


Figure 6.3: Nucleotide mutation prediction with p value.

significant mutation sites. Our model provides a p-value of 0.07 for the C241T mutation site, which has implications for viral RNA folding. It changes the hexameric loop motif of SARS-CoV-2 SL5b in C-residue. It also changes the Sl4, changing the folding of RNA [107]. G11083 has a p-value of 0, which is induced by L3606F mutation in the ORF1ab protein. The increase in mutations among the viral progeny is also aided by RNA recombination with the G11083T mutation. It may improve the carrier's fitness as a benefit allele in the future [183]. C14408T gives a p-value of 0, which shows significant association with the fatality rate [155]. C22227T gives a p-value of 0, which has a potential influence on infectivity, pathogenicity, or host adaptability alongside T445, C6286T, C26801G, and G29645T mutations due to haplotype epidemic trends [51]. A23403G is induced by D614G mutation in S protein which gives a p-value of 0.01. D614G changes S protein structure to human ACE2 binding receptor and increases infectivity [190]. It is also a shared mutation between all variants of concerns. The model provides a p-value of 1 for some mutation sites like G9526T, C7392T, G13993T, G17019T, which means these sites are not confident in the prediction. However, there is no biological functionality for these non-confident sites. Besides these mutation sites, we also observed some other mutation sites where prediction is confident but no functionality mentioned in literature yet such as G376T, G571A, G872A, C1912T, A2031G, G2407A, C3117T, A3390C, A4870G, C6628T, C7279T, T7288A and so on.

6.2.3 Spike Mutation Prediction

For the spike mutation prediction, we use the two bootstrap methods discussed in Section 6.1.3. The p-value of the entire model is P<.001 and P=0.42, respectively. Both of which land within the threshold we set for the tests, and thus we claim the robustness of the model on spike protein prediction. For the first bootstrap methods, the p-values of most sites are P<.001, indicating a low probability of being outperformed by shuffled data on those sites. For the second approach, almost all the sites have their p-value between 0.1-0.5, showing an insignificant change in the accuracy of prediction after removing random mutation sites from the data. It indicates that our model is statistically significant as it provides prediction based on inference from time-series relation among input data instead of random chances. The p-value of each mutation sites visualized in Figure 6.4 using Tableau workbook.

The p_value_b1 and p_value_b2 indicate the confidence of prediction of each mutation site in two bootstrap methods. Here, we indicate the confident mutation sites with arrow. We map back the p-value to the biological functionality of mutation sites. L5F, S13I, L18F, T20N, P26S, T95I, D138Y, W152C, R190S shows confidence in prediction with a p-value of 0.1-0.7, which we identified as recurrent mutations in spike protein. The antigenic supersite in the N-terminal domain contains these mutation locations. L5F may facilitate protein folding and assembly of virion through spike proteins mediate entry into the endoplasmic reticulum (ER). It may also enhances the hydrophobicity of the supercite [186]. Because the N terminal domain antigenic supersite was modified by a relocation of the signal peptide cleavage site and the development of a new disulfide bond in B.1.427/B.1429, the S13I and W152C mutations resulted in an entire loss of neutralization for 10 of 10 N terminal domain-specific mAbs [105]. L18F has been observed to interfere with the binding of neutralizing antibodies in the South African strain. Further, it has a replicative advan-

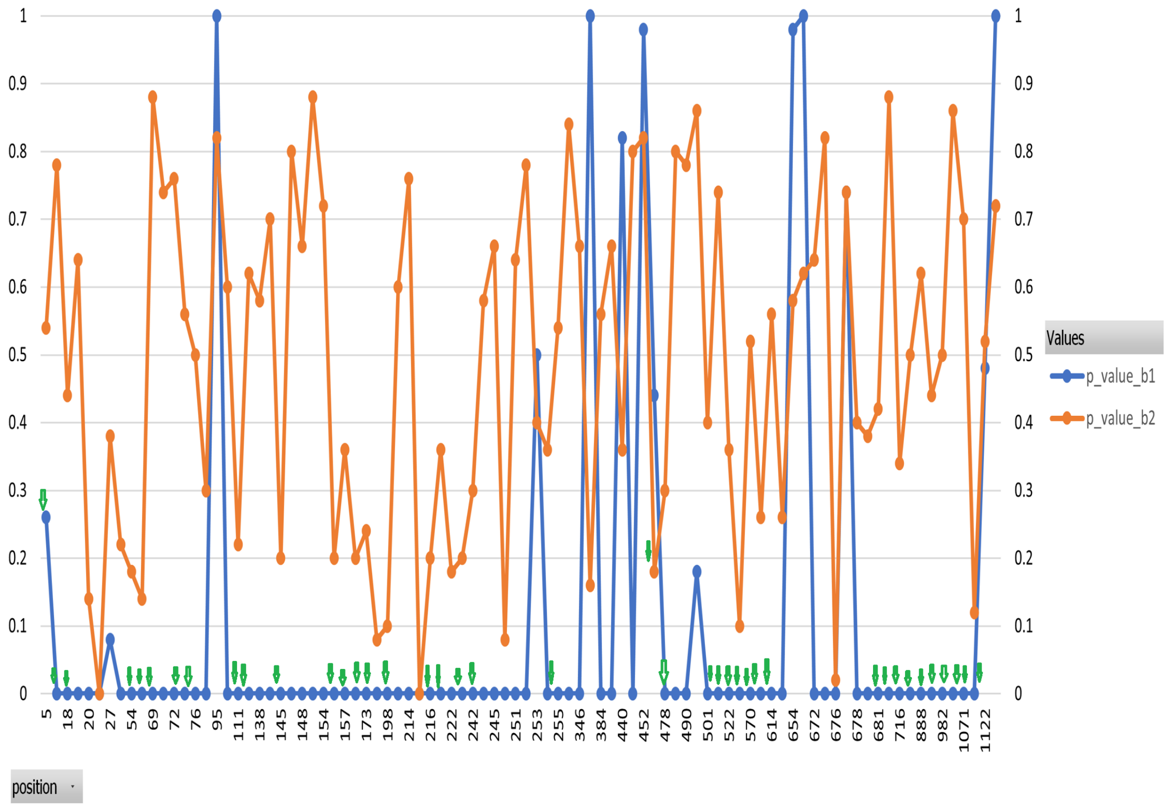


Figure 6.4: Spike mutation prediction with p value.

tage within England, South African, and Brazil strains [41]. The mutations T20N, P26S, L18F, D138Y, and R190S caused a dramatic reorganization of the surface potential in Brazil B.1.1.248 variant [40]. L18F, T20N, L452R, and N501Y are located in the N terminal domain and receptor-binding domain regions, and they are responsible for immune escape and higher transmissibility [32].T95I is a B.1.526 variant synonymous mutation that breaks all side-chain/side-chain H-bonds and/or side-chain/main-chain H-bonds created by a buried Threonine residue[152]. V367F shows a confident p-value of 0.16. Due to enhanced structural stabilization of the receptor-binding domain beta-sheet scaffold, V367F has a greater binding affinity for human ACE2 [117]. K417N and E484K lower COVID-19 vaccination effectiveness and impart resistance to several therapeutic monoclonal antibodies, resulting in a confident prediction [133]. S477N strengthens the binding of the spike protein with the hACE2 receptor, where its prediction is confident [143]. T478K affects the spike binding domain with human receptor ACE2, increasing the electrostatic potential on the interface, which shows a p-value of 0.3 [28]. In some strains, A570V caused a reduction in total spike protein stability [126]. D614G has a p-value of 0.56 and is known to regulate the balance of close to open trimers while also increasing infectivity. N679K and P681H exchange neutral residues for two positively charged residues adjacent to the cleavage sites, potentially benefiting enzyme-substrate coupling. Alongside P681H, P681R provides an additional basic residue proximal to the cleavage site, which may influence S1/S2 cleavability by furin. It also facilitates the furin-mediated spike cleavage and enhances viral infectivity. The efficacy of viral fusion and cell-cell viral spread, particularly when it occurs in the background of other S changes [164]. Our prediction is confident for both these sites with a p-value close to 0.4. V1176F mutation shows a p-value of 0.72 and is involved in the viral fusion machinery. It has the ability to eliminate a key neutralization epitope, and it contributes to antibody neutralization escape [75]. For some mutation sites like V367F, H655Y the p-value is 1, and these sites are not biologically functional. Besides these biological significant mutations, we also observed a number of mutations where p value fall into the threshold value but no functionalities mentioned in literature yet. Such as G75V, T76I, A76V, D80Y, S98F, D111N, V126A, Y145N, E156V, F157S, R158S, Q173L, R190V, D198F, L216F, S221L, A222V, L241F, L242F and so on.

6.2.4 Discussions

We perform the time series forecasting of nucleotide and spike mutations of SARS-CoV-2 using an encoder-decoder stacked sequences to sequences LSTM model. For 90 days of previous mutation information, we predict next 7 days mutation information using the sliding window technique. To evaluate the performance of the model, we measure MAE and RMSE for both prediction models and it gives low error for both nucleotide and spike mutation prediction which indicates good performance of the model. In the statistical tests with bootstrapping, both nucleotide and spike mutation prediction model give a low p-value which indicates that our model is predicting based on inference from time-series relation among input data instead of random prediction. Also, removing random mutation sites haven't affected the performance of the model for spike mutation prediction. For both nucleotide and spike mutation prediction, the p value of most of the mutation sites fall into threshold values which indicates that our prediction is confident in most of the sites. Comparatively, the predictions on spike mutation sites are more confident than the nucleotide mutation prediction. Our analysis suggests that that the forecasting is more confident in some biologically significant sites than others insignificant sites. We also observed confident predictions in some mutation sites which biological functionality is yet to be discovered.

6.3 Conclusions

In this study, we present the methods and results of time series forecasting of nucleotide and spike mutation prediction using an encoder-decoder LSTM network model. We predict those mutations which survived in SARS-CoV-2. We map back the prediction confidence of mutation sites from our model to the biological functionality from the literature. Overall, our model provides confident prediction for most of the biologically significant mutations sites. Our analysis suggests that there are several other spike mutation sites where prediction is confident but the functionality is yet to be discovered. To better understand the implications of these results, further studies should be performed to identify the functionality of these mutation sites. The prediction of mutation sites can be useful for drug and vaccine design of COVID-19.

Chapter 7

Conclusion and Future Work

The continuous evolution of SARS-CoV-2 has affected people worldwide and impacted public health measures and global research. With the persistence of current variants of concern and the emergence of new ones, it is imperative to analyze a large number of genomic sequence datasets. In this dissertation, we performed several numerical analysis on the mutations of spike, membrane and envelope proteins of SARS-CoV-2 to evaluate the implications of the co-occurrence and recurrence of mutations. We established the correlation of mutations in the N-terminal domain (NTD), receptor-binding domain (RBD), and other vital regions in the spike protein using the mutation pairs that have high rates of co-occurrence and re-occurrence by measuring their Pearson co-relation co-efficient and Phi co-efficient. With their biological implications, we identified functional mutations that potentially warrant further observation and investigation. This analysis provides valuable information that may help reduce the morbidity and mortality of COVID-19. Our analysis suggests that the membrane and envelope protein mutations are not influencing the virus transmission like the spike protein as there is no recurrent co-mutation pairs in the these proteins. We performed time-series analysis on the variants of concerns and variants of interests to see the trend of mutation of different variants. We identified several new mutations in those variants which haven't mentioned in literature yet. This research introduces an encoder-decoder LSTM network to forecast the nucleotide and spike mutations of COVID-19. The performance of the model is evaluated using MAE and RMSE. The result shows that our model has an overall low mean error in prediction. We proposed statistical tests with bootstrap methods to provide the confidence of the whole model as well as each mutation site. The experiments analyzed COVID-19 Virus strains from different aspects, and we can conclude the mutations are of certain patterns and correlations such that future mutations can be predicted to an extent. Comparatively, the spike mutation prediction is more confident than the nucleotide mutation prediction. The statistical tests show that our model is highly robust in prediction on most sites despite missing data, which are biologically significant in virus spread, high infection and transmission, and escaping immune response. The biological structure can be further learned according to the results of the co-mutation analysis and prediction confidence level to understand the virus structure mechanism better. Prediction on important sites can be used to study virus evolution and will be useful for drug and vaccine design for COVID-19.

References

[1] E. Abdollahi, D. Champredon, J. M. Langley, A. P. Galvani, and S. M. Moghadas, "Temporal estimates of case-fatality rate for covid-19 outbreaks in canada and the united states," *Cmaj*, vol. 192, pp. 666–670, June 22, 2020. DOI: https://doi.org/10.1503/cmaj.200711.

[2] J. Adler and I. Parmryd, "Quantifying colocalization by correlation: The pearson correlation coefficient is superior to the mander's overlap coefficient," *Cytometry Part A*, vol. 77, pp. 733–742, 30 March, 2010. DOI: https://doi.org/10.1002/cyto.a.20896.

[3] S. F. Ahmed, A. A. Quadeer, and M. R. McKay, "Preliminary identification of potential vaccine targets for the covid-19 coronavirus (sars-cov-2) based on sars-cov immunological studies," *Viruses*, vol. 12, pp. 1–15, 9 February, 2020. DOI: https://doi.org/10.3390/v12030254.

[4] N. Alballa and I. Al-Turaiki, "Machine learning approaches in covid-19 diagnosis, mortality, and severity risk prediction: A review," *Informatics in Medicine Unlocked*, vol. 24, p. 100 564, April 3, 2021. DOI: https://doi.org/10.1016/j.imu.2021.100564.

[5] D. M. Altmann, R. J. Boyton, and R. Beale, "Immunity to sars-cov-2 variants of concern," *Science*, vol. 371, pp. 1103–1104, March 12, 2021. DOI: https://doi.org/10.1126/science.abg7404.

[6] A. L. Arndt, B. J. Larson, and B. G. Hogue, "A conserved domain in the coronavirus membrane protein tail is important for virus assembly," *Journal of Virology*, vol. 84, pp. 11 418–11 428, August 18, 2010. DOI: https://doi.org/10.1128/JVI.01131-10.

[7] K. ArunKumar, D. V. Kalaga, C. M. S. Kumar, G. Chilkoor, M. Kawaji, and T. M. Brenza, "Forecasting the dynamics of cumulative covid-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-regressive integrated moving average (arima) and seasonal auto-regressive integrated moving average (sarima)," *Applied Soft Computing*, vol. 103, pp. 1–26, February 8, 2021. DOI: https://doi.org/10.1016/j.asoc.2021.107161.

[8] I. Astuti and Y. Ysrafil, “Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response,” *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, pp. 407–412, April 18, 2020. DOI: https://doi.org/10.1016/j.dsx.2020.04.020.

[9] D. Baud, X. Qi, K. Nielsen-Saines, D. Musso, L. Pomar, and G. Favre, “Real estimates of mortality following COVID-19 infection,” *The Lancet Infectious Diseases*, vol. 20, p. 773, July 01, 2020. DOI: https://doi.org/10.1016/S1473-3099(20)30195-X.

[10] S. Belouzard, V. C. Chu, and G. R. Whittaker, “Activation of the sars coronavirus spike protein via sequential proteolytic cleavage at two distinct sites,” *Proceedings of the National Academy of Sciences*, vol. 106, pp. 5871–5876, April 7, 2009. DOI: https://doi.org/10.1073/pnas.0809524106.

[11] M. Bianchi, D. Benvenuto, M. Giovanetti, S. Angeletti, M. Ciccozzi, and S. Pascarella, “Sars-CoV-2 envelope and membrane proteins: Structural differences linked to virus characteristics?” *BioMed Research International*, vol. 2020, pp. 1–6, May 30, 2020. DOI: https://doi.org/10.1155/2020/4389089.

[12] D. Brian and R. Baric, “Coronavirus genome structure and replication,” *Coronavirus Replication and Reverse Genetics*, vol. 287, pp. 1–30, October 25, 2005. DOI: https://doi.org/10.1007/3-540-26765-4_1.

[13] E. S. Brielle, D. Schneidman-Duhovny, and M. Linial, “The SARS-CoV-2 exerts a distinctive strategy for interacting with the ACE2 human receptor,” *Viruses*, vol. 12, pp. 1–10, April 30, 2020. DOI: https://doi.org/10.3390/v12050497.

[14] J. R. Brister, D. Ako-Adjei, Y. Bao, and O. Blinkova, “NCBI viral genomes resource,” *Nucleic Acids Research*, vol. 43, pp. D571–D577, January 28, 2015. DOI: https://doi.org/10.1093/nar/gku1207.

[15] B. Cao, Y. Wang, D. Wen, W. Liu, J. Wang, G. Fan, L. Ruan, B. Song, Y. Cai, M. Wei, *et al.*, “A trial of lopinavir–ritonavir in adults hospitalized with severe Covid-19,” *New England Journal of Medicine*, vol. 382, pp. 1787–1799, May 7, 2020. DOI: https://doi.org/10.1056/NEJMoa2001282.

[16] Y.-c. Cao, Q.-x. Deng, and S.-x. Dai, “Remdesivir for severe acute respiratory syndrome coronavirus 2 causing COVID-19: An evaluation of the evidence,” *Travel Medicine and Infectious Disease*, vol. 35, pp. 1–6, 2020. DOI: https://doi.org/10.1016/j.tmaid.2020.101647.

[17] CDC, *SARS-CoV-2 Variant Classifications and Definitions*, https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html, Accessed: 2021-08-21, 2021.

[18] G. Cerutti, Y. Guo, T. Zhou, J. Gorman, M. Lee, M. Rapp, E. R. Reddem, J. Yu, F. Bahna, J. Bimela, Y. Huang, P. S. Katsamba, L. Liu, M. S. Nair, R. Rawi, A. S. Olia, P. Wang, B. Zhang, G.-Y. Chuang, D. D. Ho, Z. Sheng, P. D. Kwong, and L. Shapiro, "Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite," *Cell Host & Microbe*, vol. 29, pp. 819–833, May 12, 2021. DOI: https://doi.org/10.1016/j.chom.2021.03.005.

[19] Z. Chagla, "The BNT162b2 (BioNTech/Pfizer) vaccine had 95% efficacy against COVID-19 7 days after the 2nd dose," *Annals of Internal Medicine*, vol. 174, JC15, February 2021. DOI: https://doi.org/10.7326/ACPJ202102160-015.

[20] R. Chandra, A. Jain, and D. S. Chauhan, *Deep learning via LSTM models for COVID-19 infection forecasting in India*, January 28, 2021. arXiv: 2101.11881 [cs.LG].

[21] M. Chatzou, C. Magis, J.-M. Chang, C. Kemena, G. Bussotti, I. Erb, and C. Notredame, "Multiple sequence alignment modeling: Methods and applications," *Briefings in Bioinformatics*, vol. 17, pp. 1009–1023, November 2016. DOI: https://doi.org/10.1093/bib/bbv099.

[22] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos, Solitons & Fractals*, vol. 135, pp. 1–6, June 2020. DOI: https://doi.org/10.1016/j.chaos.2020.109864.

[23] S. Cleemput, W. Dumon, V. Fonseca, W. Abdool Karim, M. Giovanetti, L. C. Alcantara, K. Deforche, and T. De Oliveira, "Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes," *Bioinformatics*, vol. 36, pp. 3552–3555, June 1, 2020. DOI: https://doi.org/10.1093/bioinformatics/btaa145.

[24] F. Corpet, "Multiple sequence alignment with hierarchical clustering," *Nucleic Acids Research*, vol. 16, pp. 10 881–10 890, November 25, 1988. DOI: https://doi.org/10.1093/nar/16.22.10881.

[25] M. Costanzo, M. A. De Giglio, and G. N. Roviello, "SARS-CoV-2: Recent Reports on Antiviral Therapies Based on Lopinavir/Ritonavir, Darunavir/Umifenovir, Hydroxychloroquine, Remdesivir, Favipiravir and other Drugs for the Treatment of the New Coronavirus," *Current Medicinal Chemistry*, vol. 27, pp. 4536–4541, August 5, 2020. DOI: https://doi.org/10.2174/0929867327666200416131117.

[26] N. G. Davies, S. Abbott, R. C. Barnard, C. I. Jarvis, A. J. Kucharski, J. D. Munday, C. A. B. Pearson, T. W. Russell, D. C. Tully, A. D. Washburne, T. Wenseleers, A. Gimma, W. Waites, K. L. M. Wong, K. van Zandvoort, J. D. Silverman, C. C.-1. W. Group, C.-1. G. U. (.-U. Consortium, K. Diaz-Ordaz, R. Keogh, R. M. Eggo, S. Funk, M. Jit,

K. E. Atkins, and W. J. Edmunds, “Estimated transmissibility and impact of SARS-CoV-2 lineage B. 1.1. 7 in England,” *Science*, vol. 372, pp. 1–11, April 09, 2021. DOI: https://doi.org/10.1126/science.abg3055.

[27] M. L. DeDiego, E. Álvarez, F. Almazán, M. T. Rejas, E. Lamirande, A. Roberts, W.-J. Shieh, S. R. Zaki, K. Subbarao, and L. Enjuanes, “A severe acute respiratory syndrome coronavirus that lacks the E gene is attenuated in vitro and in vivo,” *Journal of Virology*, vol. 81, pp. 1701–1713, February 15, 2007. DOI: https://doi.org/10.1128/JVI.01467-06.

[28] S. Di Giacomo, D. Mercatelli, A. Rakhimov, and F. M. Giorgi, “Preliminary report on severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) Spike mutation T478K,” *Journal of Medical Virology*, vol. 93, pp. 5638–5643, May 05, 2021. DOI: https://doi.org/10.1002/jmv.27062.

[29] M. D. Dicks, A. J. Spencer, N. J. Edwards, G. Wadell, K. Bojang, S. C. Gilbert, A. V. Hill, and M. G. Cottingham, “A Novel Chimpanzee Adenovirus Vector with Low Human Seroprevalence: Improved Systems for Vector Derivation and Comparative Immunogenicity,” *PlOS ONE*, vol. 7, pp. 1–12, July 13, 2012. DOI: https://doi.org/10.1371/journal.pone.0040385.

[30] R. C. Edgar, “MUSCLE: Multiple sequence alignment with high accuracy and high throughput,” *Nucleic Acids Research*, vol. 32, pp. 1792–1797, March 19, 2004. DOI: https://doi.org/10.1093/nar/gkh340.

[31] B. Efron, E. Halloran, and S. Holmes, “Bootstrap confidence levels for phylogenetic trees,” *Proceedings of the National Academy of Sciences*, vol. 93, pp. 13 429–13 429, November 12, 1996. DOI: https://doi.org/10.1073/pnas.93.23.13429.

[32] J. Fantini, N. Yahi, F. Azzaz, and H. Chahinian, “Structural dynamics of SARS-CoV-2 variants: A health monitoring strategy for anticipating Covid-19 outbreaks,” *Journal of Infection*, vol. 83, pp. 197–206, August 01, 2021. DOI: https://doi.org/10.1016/j.jinf.2021.06.001.

[33] N. R. Faria, T. A. Mellan, C. Whittaker, I. M. Claro, D. d. S. Candido, S. Mishra, M. A. E. Crispim, F. C. S. Sales, I. Hawryluk, J. T. McCrone, R. J. G. Hulswit, L. A. M. Franco, M. S. Ramundo, J. G. de Jesus, P. S. Andrade, T. M. Coletti, G. M. Ferreira, C. A. M. Silva, E. R. Manuli, R. H. M. Pereira, P. S. Peixoto, M. U. G. Kraemer, N. Gaburo, C. d. C. Camilo, H. Hoeltgebaum, W. M. Souza, E. C. Rocha, L. M. de Souza, M. C. de Pinho, L. J. T. Araujo, F. S. V. Malta, A. B. de Lima, J. d. P. Silva, D. A. G. Zauli, A. C. d. S. Ferreira, R. P. Schnekenberg, D. J. Laydon, P. G. T. Walker, H. M. Schlüter, A. L. P. dos Santos, M. S. Vidal, V. S. Del Caro, R. M. F. Filho, H. M. dos Santos, R. S. Aguiar,

J. L. Proença-Modena, B. Nelson, J. A. Hay, M. Monod, X. Miscouridou, H. Coupland, R. Sonabend, M. Vollmer, A. Gandy, C. A. Prete, V. H. Nascimento, M. A. Suchard, T. A. Bowden, S. L. K. Pond, C.-H. Wu, O. Ratmann, N. M. Ferguson, C. Dye, N. J. Loman, P. Lemey, A. Rambaut, N. A. Fraiji, M. d. P. S. S. Carvalho, O. G. Pybus, S. Flaxman, S. Bhatt, and E. C. Sabino, "Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil," *Science*, vol. 372, pp. 815–821, 2021. DOI: https://doi.org/10.1126/science.abh2644.

[34] A. R. Fehr and S. Perlman, "Coronaviruses: An overview of their replication and pathogenesis," in *Coronaviruses*, vol. 1282, Springer, February 12, 2015, pp. 1–23. DOI: https://doi.org/10.1007/978-1-4939-2438-7_1.

[35] R. A. Feldman, R. Fuhr, I. Smolenov, A. (Mick) Ribeiro, L. Panther, M. Watson, J. J. Senn, M. Smith, Almarsson, H. S. Pujar, M. E. Laska, J. Thompson, T. Zaks, and G. Ciaramella, "mRNA vaccines against H10N8 and H7N9 influenza viruses of pandemic potential are immunogenic and well tolerated in healthy adults in phase 1 randomized clinical trials," *Vaccine*, vol. 37, pp. 3326–3334, May 31, 2019. DOI: https://doi.org/10.1016/j.vaccine.2019.04.074.

[36] J. Gao, Z. Tian, and X. Yang, "Breakthrough: Chloroquine phosphate has shown apparent efficacy in treatment of COVID-19 associated pneumonia in clinical studies," *BioScience Trends*, vol. 14, pp. 72–73, March 16, 2020. DOI: https://doi.org/10.5582/bst.2020.01047.

[37] Y. Gautam, "Transfer Learning for COVID-19 cases and deaths forecast using LSTM network," *ISA Transactions*, pp. 1–16, January 4, 2021, ISSN: 0019-0578. DOI: https://doi.org/10.1016/j.isatra.2020.12.057.

[38] I. Glowacka, S. Bertram, M. A. Müller, P. Allen, E. Soilleux, S. Pfefferle, I. Steffen, T. S. Tsegaye, Y. He, K. Gnirss, *et al.*, "Evidence that TMPRSS2 activates the severe acute respiratory syndrome coronavirus spike protein for membrane fusion and reduces viral control by the humoral immune response," *Journal of Virology*, vol. 85, pp. 4122–4134, April 15, 2011. DOI: https://doi.org/10.1128/JVI.02232-10.

[39] N. Goldman, "Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses," *Systematic Zoology*, vol. 39, pp. 345–361, December 1990. DOI: https://doi.org/10.2307/2992355.

[40] C. E. Gómez, B. Perdiguero, and M. Esteban, "Emerging SARS-CoV-2 Variants and Impact in Global Vaccination Programs against SARS-CoV-2/COVID-19," *Vaccines*, vol. 9, pp. 1–13, March 11, 2021. DOI: https://doi.org/10.3390/vaccines9030243.

[41] F. Grabowski, M. Kochanczyk, and T. Lipniacki, “L18F substrain of SARS-CoV-2 VOC-202012/01 is rapidly spreading in England,” *MedRxiv*, February 9, 2021. DOI: https://doi.org/10.1101/2021.02.07.21251262.

[42] B. S. Graham, “Rapid COVID-19 vaccine development,” *Science*, vol. 368, pp. 945–946, May 29, 2020. DOI: https://doi.org/10.1126/science.abb8923.

[43] S. Griffin, “Covid-19: Lopinavir-ritonavir does not benefit hospitalised patients, UK trial finds,” *Bmj*, vol. 370, July 01, 2020. DOI: https://doi.org/10.1136/bmj.m2650.

[44] A. Grover and M. Oberoi, “A systematic review and meta-analysis to evaluate the clinical outcomes in COVID-19 patients on angiotensin converting enzyme inhibitors or angiotensin receptor blockers,” *European Heart Journal-Cardiovascular Pharmacotherapy*, vol. 7, pp. 148–157, March 15, 2020. DOI: https://doi.org/10.1093/ehjcvp/pvaa064.

[45] M. Guan, H. Y. Chen, S. Y. Foo, Y.-J. Tan, P.-Y. Goh, and S. H. Wee, “Recombinant protein-based enzyme-linked immunosorbent assay and immunochromatographic tests for detection of immunoglobulin G antibodies to severe acute respiratory syndrome (SARS) coronavirus in SARS patients,” *Clinical and Vaccine Immunology*, vol. 11, pp. 287–291, March 2004. DOI: https://doi.org/10.1128/CDLI.11.2.287-291.2004.

[46] L. Guruprasad, “Human SARS CoV-2 spike protein mutations,” *Proteins: Structure, Function, and Bioinformatics*, vol. 89, pp. 569–576, January 09, 2021. DOI: https://doi.org/10.1002/prot.26042.

[47] T. A. Hall *et al.*, “BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT,” in *Nucleic Acids Symposium Series*, [London]: Information Retrieval Ltd., c1979-c2000., vol. 41, 1999, pp. 95–98. [Online]. Available: https://ci.nii.ac.jp/naid/10030689140/en/.

[48] G.-Z. Han, “Pangolins harbor SARS-CoV-2-related coronaviruses,” *Trends in Microbiology*, vol. 28, pp. 515–517, April 10, 2020. DOI: https://doi.org/10.1016/j.tim.2020.04.001.

[49] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, pp. 1735–1780, November 15, 1997. DOI: https://doi.org/10.1162/neco.1997.9.8.1735.

[50] E. B. Hodcroft, *Covariants: Sars-cov-2 mutations and variants of interest*, 2021.

[51] E. B. Hodcroft, M. Zuber, S. Nadeau, T. G. Vaughan, K. H. D. Crawford, C. L. Althaus, M. L. Reichmuth, J. E. Bowen, A. C. Walls, D. Corti, J. D. Bloom, D. Veesler, D. Mateo, A. Hernando, I. Comas, F. González Candelas, S.-S. consortium, T. Stadler, and R. A. Neher, "Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020," *MedRxiv*, vol. 595, pp. 707–712, March 24, 2021. DOI: https://doi.org/10.1101/2020.10.25.20219063.

[52] M. Hoffmann, P. Arora, R. Groß, A. Seidel, B. F. Hörnich, A. S. Hahn, N. Krüger, L. Graichen, H. Hofmann-Winkler, A. Kempf, M. S. Winkler, S. Schulz, H.-M. Jäck, B. Jahrsdörfer, H. Schrezenmeier, M. Müller, A. Kleger, J. Münch, and S. Pöhlmann, "SARS-CoV-2 variants B. 1.351 and P. 1 escape from neutralizing antibodies," *Cell*, vol. 184, pp. 2384–2393, April 29, 2021. DOI: https://doi.org/10.1016/j.cell.2021.03.036.

[53] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche, M. A. Müller, C. Drosten, and S. Pöhlmann, "SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor," *Cell*, vol. 181, pp. 271–280, March 05, 2020. DOI: https://doi.org/10.1016/j.cell.2020.02.052.

[54] B. G. Hogue and C. E. Machamer, "Coronavirus structural proteins and virus assembly," *Nidoviruses*, pp. 179–200, December 07, 2007. DOI: https://doi.org/10.1128/9781555815790.ch12.

[55] M. K. Hourfar, W. K. Roth, E. Seifried, and M. Schmidt, "Comparison of two real-time quantitative assays for detection of severe acute respiratory syndrome coronavirus," *Journal of Clinical Microbiology*, vol. 42, pp. 2094–2100, May 2004. DOI: https://doi.org/10.1128/JCM.42.5.2094-2100.2004.

[56] Z. Hu, P. Hao, X.-J. Song, S.-M. Jiang, Y.-X. Liu, H. Wang, X. Rao, H.-D. Song, S.-Y. Wang, Y. Zuo, A.-H. Zheng, M. Luo, H.-L. Wang, F. Deng, H.-Z. Wang, Z.-H. Hu, M.-X. Ding, G.-P. Zhao, and H.-K. Deng, "Identification of two critical amino acid residues of the severe acute respiratory syndrome coronavirus spike protein for its variation in zoonotic tropism transition via a double substitution strategy," *Journal of Biological Chemistry*, vol. 280, pp. 29 588–29 595, August 09, 2005. DOI: https://doi.org/10.1074/jbc.M500662200.

[57] Y. Huang, X. Huang, S. Wang, Y. Yu, S. Ni, and Q. Qin, "Soft-shelled turtle iridovirus enters cells via cholesterol-dependent, clathrin-mediated endocytosis as well as macropinocytosis," *Archives of Virology*, vol. 163, pp. 3023–3033, July 31, 2018. DOI: https://doi.org/10.1007/s00705-018-3966-8.

[58] J. A. Jaimes, N. M. André, J. S. Chappie, J. K. Millet, and G. R. Whittaker, “Phylogenetic analysis and structural modeling of SARS-CoV-2 spike protein reveals an evolutionary distinct and proteolytically-sensitive activation loop,” *Journal of Molecular Biology*, vol. 432, pp. 3309–3325, May 1, 2020. DOI: https://doi.org/10.1016/j.jmb.2020.04.009.

[59] W. Ji, W. Wang, X. Zhao, J. Zai, and X. Li, “Cross-species transmission of the newly identified coronavirus 2019-nCoV,” *Journal of Medical Virology*, vol. 92, pp. 433–440, January 22, 2020. DOI: https://doi.org/10.1002/jmv.25682.

[60] S. Kang, W. Peng, Y. Zhu, S. Lu, M. Zhou, W. Lin, W. Wu, S. Huang, L. Jiang, X. Luo, and M. Deng, “Recent Progress in understanding 2019 Novel Coronavirus associated with Human Respiratory Disease: Detection, Mechanism and Treatment,” *International Journal of Antimicrobial Agents*, pp. 1–9, May 2020. DOI: https://doi.org/10.1016/j.ijantimicag.2020.105950.

[61] K. Katoh, J. Rozewicki, and K. D. Yamada, “MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization,” *Briefings in Bioinformatics*, vol. 20, pp. 1160–1166, September 06, 2017. DOI: https://doi.org/10.1093/bib/bbx108.

[62] M. Kawase, K. Shirato, L. van der Hoek, F. Taguchi, and S. Matsuyama, “Simultaneous treatment of human bronchial epithelial cells with serine and cysteine protease inhibitors prevents severe acute respiratory syndrome coronavirus entry,” *Journal of Virology*, vol. 86, pp. 6537–6545, May 24, 2012. DOI: https://doi.org/10.1128/JVI.00094-12.

[63] R. A. Khailany, M. Safdar, and M. Ozaslan, “Genomic characterization of a novel SARS-CoV-2,” *Gene Reports*, pp. 1–6, June 2020. DOI: https://doi.org/10.1016/j.genrep.2020.100682.

[64] J. S. Khalili, H. Zhu, N. S. A. Mak, Y. Yan, and Y. Zhu, “Novel coronavirus treatment with ribavirin: Groundwork for an evaluation concerning COVID-19,” *Journal of Medical Virology*, March 30, 2020. DOI: https://doi.org/10.1002/jmv.25798.

[65] J. Khateeb, Y. Li, and H. Zhang, “Emerging SARS-CoV-2 variants of concern and potential intervention approaches,” *Critical Care*, vol. 25, pp. 1–8, July 12, 2021. DOI: https://doi.org/10.1186/s13054-021-03662-x.

[66] D. Kim, J.-Y. Lee, J.-S. Yang, J. W. Kim, V. N. Kim, and H. Chang, “The architecture of SARS-CoV-2 transcriptome,” *Cell*, vol. 181, pp. 914–921, May 14, 2020. DOI: https://doi.org/10.1016/j.cell.2020.04.011.

[67] İ. Kırbaş, A. Sözen, A. D. Tuncer, and F. Ş. Kazancıoğlu, "Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches," *Chaos, Solitons & Fractals*, vol. 138, pp. 1–7, September 2020. DOI: https://doi.org/10.1016/j.chaos.2020.110015.

[68] J. Klumperman, J. K. Locker, A. Meijer, M. C. Horzinek, H. J. Geuze, and P. Rottier, "Coronavirus M proteins accumulate in the Golgi complex beyond the site of virion budding," *Journal of Virology*, vol. 68, pp. 6523–6534, October 1994. DOI: https://doi.org/10.1128/JVI.68.10.6523-6534.1994.

[69] M. D. Knoll and C. Wonodi, "Oxford–AstraZeneca COVID-19 vaccine efficacy," *The Lancet*, vol. 397, pp. 72–74, January 09, 2021. DOI: https://doi.org/10.1016/S0140-6736(20)32623-4.

[70] S. Kumar, G. Stecher, M. Li, C. Knyaz, and K. Tamura, "MEGA X: molecular evolutionary genetics analysis across computing platforms," *Molecular Biology and Evolution*, vol. 35, pp. 1547–1549, June 01, 2018. DOI: https://doi.org/10.1093/molbev/msy096.

[71] S. Kumar, "Drug and Vaccine Design against Novel Coronavirus (2019-nCoV) Spike Protein through Computational Approach," Preprints 2020. DOI: https://doi.org/10.20944/preprints202002.0071.v1.

[72] K. Kupferschmidt and J. Cohen, "Race to find COVID-19 treatments accelerates," *Science*, vol. 367, pp. 1412–1413, March 27, 2020. DOI: https://doi.org/10.1126/science.367.6485.1412.

[73] T. T.-Y. Lam, N. Jia, Y.-W. Zhang, M. H.-H. Shum, J.-F. Jiang, H.-C. Zhu, Y.-G. Tong, Y.-X. Shi, X.-B. Ni, Y.-S. Liao, W.-J. Li, B.-G. Jiang, W. Wei, T.-T. Yuan, K. Zheng, X.-M. Cui, J. Li, G.-Q. Pei, X. Qiang, W. Y.-M. Cheung, L.-F. Li, F.-F. Sun, S. Qin, J.-C. Huang, G. M. Leung, E. C. Holmes, Y.-L. Hu, Y. Guan, and W.-C. Cao, "Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins," *Nature*, vol. 583, pp. 1–4, March 26, 2020. DOI: https://doi.org/10.1038/s41586-020-2169-0.

[74] J. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang, L. Zhang, *et al.*, "Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor," *Nature*, vol. 581, pp. 215–220, March 30, 2020. DOI: https://doi.org/10.1038/s41586-020-2180-5.

[75] E. Lasek-Nesselquist, P. Lapierre, E. Schneider, K. S. George, and J. Pata, "The localized rise of a B. 1.526 variant containing an E484K mutation in New York State," *medRxiv*, March 01, 2021. DOI: https://doi.org/10.1101/2021.02.26.21251868.

[76] A. S. Lauring and E. B. Hodcroft, "Genetic variants of SARS-CoV-2—what do they mean?" *Jama*, vol. 325, pp. 529–531, January 06, 2021. DOI: https://doi.org/10.1001/jama.2020.27124.

[77] T. T. Le, Z. Andreadakis, A. Kumar, R. G. Roman, S. Tollefsen, M. Saville, and S. Mayhew, "The COVID-19 vaccine development landscape," *Nat Rev Drug Discov*, vol. 19, pp. 305–306, April 09, 2020. DOI: https://doi.org/10.1038/d41573-020-00073-5.

[78] F. Li, "Structure, function, and evolution of coronavirus spike proteins," *Annual Review of Virology*, vol. 3, pp. 237–261, September 29, 2016. DOI: https://doi.org/10.1146/annurev-virology-110615-042301.

[79] G. Li, Y. Fan, Y. Lai, T. Han, Z. Li, P. Zhou, W. Wang, D. Hu, X. Liu, Q. Zhang, and J. Wu, "Coronavirus infections and immune responses," *Journal of Medical Virology*, vol. 92, pp. 424–432, January 25, 2020. DOI: https://doi.org/10.1002/jmv.25685.

[80] W. Li, A. Cowley, M. Uludag, T. Gur, H. McWilliam, S. Squizzato, Y. M. Park, N. Buso, and R. Lopez, "The EMBL-EBI bioinformatics web and programmatic tools framework," *Nucleic Acids Research*, vol. 43, W580–W584, July 01, 2015. DOI: https://doi.org/10.1093/nar/gkv279.

[81] W. Li, C. Zhang, J. Sui, J. H. Kuhn, M. J. Moore, S. Luo, S.-K. Wong, I.-C. Huang, K. Xu, N. Vasilieva, *et al.*, "Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2," *The EMBO Journal*, vol. 24, pp. 1634–1643, April 20, 2005. DOI: https://doi.org/10.1038/sj.emboj.7600640.

[82] X. Li, J. Zai, Q. Zhao, Q. Nie, Y. Li, B. T. Foley, and A. Chaillon, "Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2," *Journal of Medical Virology*, vol. 92, pp. 602–611, February 27, 2020. DOI: https://doi.org/10.1002/jmv.25731.

[83] C. Liang, L. Tian, Y. Liu, N. Hui, G. Qiao, H. Li, Z. Shi, Y. Tang, D. Zhang, X. Xie, and X. Zhao, "A Promising Antiviral Candidate Drug for the COVID-19 Pandemic: A Mini-Review of Remdesivir," *European Journal of Medicinal Chemistry*, vol. 201, p. 112 527, September 01, 2020. DOI: https://doi.org/10.1016/j.ejmech.2020.112527.

[84] C. Liu, Q. Zhou, Y. Li, L. V. Garner, S. P. Watkins, L. J. Carter, J. Smoot, A. C. Gregg, A. D. Daniels, S. Jervey, and D. Albaiu, "Research and Development on Therapeutic Agents and Vaccines for COVID-19 and Related Human Coronavirus Diseases," vol. 6, pp. 315–331, March 12, 2020. DOI: https://doi.org/10.1021/acscentsci.0c00272.

[85] D. X. Liu, J. Q. Liang, and T. S. Fung, “Human Coronavirus-229E,-OC43,-NL63, and-HKU1 (Coronaviridae),” *Encyclopedia of Virology*, vol. 2, pp. 428–440, 2021. DOI: https://doi.org/10.1016/B978-0-12-809633-8.21501-X.

[86] J. Liu, X. Liao, S. Qian, J. Yuan, F. Wang, Y. Liu, Z. Wang, F.-S. Wang, L. Liu, and Z. Zhang, “Community transmission of severe acute respiratory syndrome coronavirus 2, Shenzhen, China, 2020,” vol. 26, pp. 1320–1223, June 2020. DOI: https://doi.org/10.3201/eid2606.200239.

[87] P. Liu, J.-Z. Jiang, X.-F. Wan, Y. Hua, L. Li, J. Zhou, X. Wang, F. Hou, J. Chen, J. Zou, and J. Chen, “Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)?” *PLOS Pathogens*, vol. 16, pp. 1–13, May 14, 2020. DOI: https://doi.org/10.1371/journal.ppat.1008421.

[88] W. J. Liu, M. Zhao, K. Liu, K. Xu, G. Wong, W. Tan, and G. F. Gao, “T-cell immunity of SARS-CoV: Implications for vaccine development against MERS-CoV,” *Antiviral Research*, vol. 137, pp. 82–92, January 2017. DOI: https://doi.org/10.1016/j.antiviral.2016.11.006.

[89] Z. Liu, X. Xiao, X. Wei, J. Li, J. Yang, H. Tan, J. Zhu, Q. Zhang, J. Wu, and L. Liu, “Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2,” *Journal of Medical Virology*, vol. 92, pp. 595–601, February 26, 2020. DOI: https://doi.org/10.1002/jmv.25726.

[90] E. H. Livingston, P. N. Malani, and C. B. Creech, “The Johnson & Johnson Vaccine for COVID-19,” *Jama*, vol. 325, pp. 1575–1575, March 1, 2021. DOI: https://doi.org/10.1001/jama.2021.2927.

[91] S.-M. Lok, “An NTD supersite of attack,” *Cell Host & Microbe*, vol. 29, pp. 744–746, May 12, 2021. DOI: https://doi.org/10.1016/j.chom.2021.04.010.

[92] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Menf, J. Wang, Y. Lin, J. Yuan, Z. Xie, J. Ma, W. J Liu, D. Wang, W. Xu, E. C Holmes, G. F Gao, G. Wu, W. Chen, W. Shi, and W. Tan, “Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding,” *The Lancet*, vol. 395, pp. 565–574, February 22, 2020. DOI: https://doi.org/10.1016/S0140-6736(20)30251-8.

[93] J. Luan, Y. Lu, X. Jin, and L. Zhang, “Spike protein recognition of mammalian ACE2 predicts the host range and an optimized ACE2 for SARS-CoV-2 infection,” *Biochemical and Biophysical Research Communications*, vol. 526, pp. 165–169, May 21, 2020. DOI: https://doi.org/10.1016/j.bbrc.2020.03.047.

[94] D.-f. Lv, Q.-m. Ying, Y.-s. Weng, C.-b. Shen, J.-g. Chu, J.-p. Kong, D.-h. Sun, X. Gao, X.-b. Weng, and X.-q. Chen, “Dynamic change process of target genes by RT-PCR testing of SARS-Cov-2 during the course of a Coronavirus Disease 2019 patient,” *Clinica Chimica Acta*, vol. 506, pp. 172–175, July 2020. DOI: https://doi.org/10.1016/j.cca.2020.03.032.

[95] L. Lv, G. Li, J. Chen, X. Liang, and Y. Li, “Comparative genomic analysis revealed specific mutation pattern between human coronavirus SARS-CoV-2 and Bat-SARSr-CoV RaTG13,” *Frontiers in Microbiology*, vol. 11, pp. 1–7, November 30, 2020. DOI: https://doi.org/10.3389/fmicb.2020.584717.

[96] S. Lyu, X. Yuan, H. Zhang, X. Hang, Y. Li, W. Shi, L. Liu, Z. Yu, and Y. Wu, “Transcriptome profiling analysis of lung tissue of Chinese soft-shell turtle infected by Trionyx sinensis Hemorrhagic Syndrome Virus,” *Fish & Shellfish Immunology*, vol. 98, pp. 653–660, March 2020. DOI: https://doi.org/10.1016/j.fsi.2019.10.061.

[97] Y. Ma, Y. Huang, T. Wang, A. P. Xiang, and W. Huang, “ACE2 shedding and FURIN abundance in target organs may influence the efficiency of SARS-CoV-2 entry,” *The Open Bioinformatics Journal*, vol. 14, pp. 1–12, March 22, 2021. DOI: https://doi.org/10.2174/1875036202114010001.

[98] I. G. Madu, S. L. Roth, S. Belouzard, and G. R. Whittaker, “Characterization of a highly conserved domain within the severe acute respiratory syndrome coronavirus spike protein S2 domain with characteristics of a viral fusion peptide,” *Journal of Virology*, vol. 83, pp. 7411–7421, August 01, 2009. DOI: https://doi.org/10.1128/JVI.00079-09.

[99] E. Mahase, “Covid-19: Moderna vaccine is nearly 95% effective, trial involving high risk and elderly people shows,” *BMJ: British Medical Journal (Online)*, vol. 371, November 17, 2020. DOI: https://doi.org/10.1136/bmj.m4471.

[100] M. Maleki, M. R. Mahmoudi, M. H. Heydari, and K.-H. Pho, “Modeling and forecasting the spread and death rate of coronavirus (COVID-19) in the world using time series models,” *Chaos, Solitons & Fractals*, vol. 140, pp. 1–12, November 2020. DOI: https://doi.org/10.1016/j.chaos.2020.110151.

[101] V. S. Mandala, M. J. McKay, A. A. Shcherbakov, A. J. Dregni, A. Kolocouris, and M. Hong, “Structure and drug binding of the SARS-CoV-2 envelope protein transmembrane domain in lipid bilayers,” *Nature Structural & Molecular Biology*, vol. 27, pp. 1202–1208, November 11, 2020. DOI: https://doi.org/10.1038/s41594-020-00536-8.

[102] I. Manopo, L. Lu, Q. He, L. L. Chee, S.-W. Chan, and J. Kwang, "Evaluation of a safe and sensitive Spike protein-based immunofluorescence assay for the detection of antibody responses to SARS-CoV," *Journal of Immunological Methods*, vol. 296, pp. 37–44, January 2005. DOI: https://doi.org/10.1016/j.jim.2004.10.012.

[103] M. Marovich, J. R. Mascola, and M. S. Cohen, "Monoclonal Antibodies for Prevention and Treatment of COVID-19," *Jama*, vol. 324, pp. 131–132, June 15, 2020. DOI: https://doi.org/10.1001/jama.2020.10245.

[104] V. Maurya, S. Kumar, M. Brahma Bhatt, and S. K. Saxena, "Therapeutic Development and Drugs for the Treatment of COVID-19," *Coronavirus Disease 2019 (COVID-19)*, pp. 109–126, April 30, 2020. DOI: https://doi.org/10.1007/978-981-15-4814-7_10.

[105] M. McCallum, J. Bassi, A. De Marco, A. Chen, A. C. Walls, J. Di Iulio, M. A. Tortorici, M.-J. Navarro, C. Silacci-Fregni, C. Saliba, K. R. Sprouse, M. Agostini, D. Pinto, K. Culap, S. Bianchi, S. Jaconi, E. Cameroni, J. E. Bowen, S. W. Tilles, M. S. Pizzuto, S. B. Guastalla, G. Bona, A. F. Pellanda, C. Garzoni, W. C. Van Voorhis, L. E. Rosen, G. Snell, A. Telenti, H. W. Virgin, L. Piccoli, D. Corti, and D. Veesler, "SARS-CoV-2 immune evasion by the B. 1.427/B. 1.429 variant of concern," *Science*, vol. 373, pp. 648–654, August 06, 2021. DOI: https://doi.org/10.1126/science.abi7994.

[106] D. L. Mckee, A. Sternberg, U. Stange, S. Laufer, and C. Naujokat, "Candidate drugs against SARS-CoV-2 and COVID-19," *Pharmacological Research*, vol. 157, pp. 1–9, July 2020. DOI: https://doi.org/10.1016/j.phrs.2020.104859.

[107] A. Mishra, A. K. Pandey, P. Gupta, P. Pradhan, S. Dhamija, J. Gomes, B. Kundu, P. Vivekanandan, and M. B. Menon, "Mutation landscape of SARS-CoV-2 reveals five mutually exclusive clusters of leading and trailing single nucleotide substitutions," *bioRxiv*, July 27, 2020. DOI: https://doi.org/10.1101/2020.05.07.082768.

[108] M. Mohammadi, M. Shayestehpour, and H. Mirzaei, "The impact of spike mutated variants of SARS-CoV2 [Alpha, Beta, Gamma, Delta, and Lambda] on the efficacy of subunit recombinant vaccines," *The Brazilian Journal of Infectious Diseases*, pp. 1–9, August 17, 2021. DOI: https://doi.org/10.1016/j.bjid.2021.101606.

[109] L. Mousavizadeh and S. Ghasemi, "Genotype and phenotype of COVID-19: Their roles in pathogenesis," *Journal of Microbiology, Immunology and Infection*, vol. 54, pp. 159–163, March 31, 2020. DOI: https://doi.org/10.1016/j.jmii.2020.03.022.

[110] N. Muralidharan, R. Sakthivel, D. Velmurugan, and M. M. Gromiha, “Computational studies of drug repurposing and synergism of lopinavir, oseltamivir and ritonavir binding with SARS-CoV-2 protease against COVID-19,” *Journal of Biomolecular Structure and Dynamics*, vol. 39, pp. 2673–2678, 2020. DOI: https://doi.org/10.1080/07391102.2020.1752802.

[111] T. F. F. Ng, C. Manire, K. Borrowman, T. Langer, L. Ehrhart, and M. Breitbart, “Discovery of a Novel Single-Stranded DNA Virus from a Sea Turtle Fibropapilloma by Using Viral Metagenomics,” *Journal of Virology*, vol. 83, pp. 2500–2509, March 2009. DOI: https://doi.org/10.1128/JVI.01946-08.

[112] J. L. Nieto-Torres, C. Verdiá-Báguena, J. M. Jimenez-Guardeño, J. A. Regla-Nava, C. Castaño-Rodriguez, R. Fernandez-Delgado, J. Torres, V. M. Aguilella, and L. Enjuanes, “Severe acute respiratory syndrome coronavirus E protein transports calcium ions and activates the NLRP3 inflammasome,” *Virology*, vol. 485, pp. 330–339, November 2015. DOI: https://doi.org/10.1016/j.virol.2015.08.010.

[113] H. Nishiura, T. Kobayashi, T. Miyama, A. Suzuki, S.-m. Jung, K. Hayashi, R. Kinoshita, Y. Yang, B. Yuan, A. R. Akhmetzhanov, and N. M. Linton, “Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19),” *International Journal of Infectious Diseases*, vol. 94, pp. 154–155, May 01, 2020. DOI: https://doi.org/10.1016/j.ijid.2020.03.020.

[114] I. K. Oboho, S. M. Tomczyk, A. M. Al-Asmari, A. A. Banjar, H. Al-Mugti, M. S. Aloraini, K. Z. Alkhaldi, E. L. Almohammadi, B. M. Alraddadi, S. I. Gerber, D. L. Swerdlow, J. T. Watson, and T. A. Madani, “2014 MERS-CoV outbreak in Jeddah—a link to health care facilities,” *New England Journal of Medicine*, vol. 372, pp. 846–854, February 26, 2015. DOI: https://doi.org/10.1056/NEJMoa1408636.

[115] M. Oostra, C. De Haan, R. De Groot, and P. Rottier, “Glycosylation of the severe acute respiratory syndrome coronavirus triple-spanning membrane proteins 3a and M,” *Journal of Virology*, vol. 80, pp. 2326–2336, March 01, 2006. DOI: https://doi.org/10.1128/JVI.80.5.2326-2336.2006.

[116] J. T. Ortega, M. L. Serrano, F. H. Pujol, and H. R. Rangel, “Role of changes in SARS-CoV-2 spike protein in the interaction with the human ACE2 receptor: An in silico analysis,” *EXCLI Journal*, vol. 19, pp. 410–417, March 18, 2020. DOI: https://doi.org/10.17179/excli2020-1167.

[117] J. Ou, Z. Zhou, R. Dai, J. Zhang, S. Zhao, X. Wu, W. Lan, Y. Ren, L. Cui, Q. Lan, L. Lu, D. Seto, J. Chodosh, J. Wu, G. Zhang, and Q. Zhang, “V367F Mutation in SARS-CoV-2 Spike RBD Emerging during

the Early Transmission Phase Enhances Viral Infectivity through Increased Human ACE2 Receptor Binding Affinity," *Journal of Virology*, vol. 95, JVI–00 617, July 26, 2020. DOI: https://doi.org/10.1128/JVI.00617-21.

[118] X. Ou, Y. Liu, X. Lei, P. Li, D. Mi, L. Ren, L. Guo, R. Guo, T. Chen, J. Hu, Z. Xiang, Z. Mu, X. Chen, J. Chen, K. Hu, Q. Jin, J. Wang, and Z. Qian, "Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV," *Nature Communications*, vol. 11, pp. 1–12, March 27, 2020. DOI: https://doi.org/10.1038/s41467-020-15562-9.

[119] M. Pachetti, B. Marini, F. Benedetti, F. Giudici, E. Mauro, P. Storici, C. Masciovecchio, S. Angeletti, M. Ciccozzi, R. C. Gallo, D. Zella, and R. Ippodrino, "Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant," *Journal of Translational Medicine*, vol. 18, pp. 1–9, April 22, 2020. DOI: https://doi.org/10.1186/s12967-020-02344-6.

[120] V. Papastefanopoulos, P. Linardatos, and S. Kotsiantis, "COVID-19: a comparison of time series methods to forecast percentage of active cases per population," *Applied Sciences*, vol. 10, pp. 1–15, June 3, 2020. DOI: https://doi.org/10.3390/app10113880.

[121] R. K. Pathan, M. Biswas, and M. U. Khandaker, "Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model," *Chaos, Solitons & Fractals*, vol. 138, pp. 1–7, September 2020. DOI: https://doi.org/10.1016/j.chaos.2020.110018.

[122] N. Petrosillo, G. Viceconte, O. Ergonul, G. Ippolito, and E. Petersen, "COVID-19, SARS and MERS: are they closely related?" *Clinical Microbiology and Infection*, vol. 26, pp. 729–734, June 2020, ISSN: 1198-743X. DOI: https://doi.org/10.1016/j.cmi.2020.03.026.

[123] T. Phan, "Genetic diversity and evolution of SARS-CoV-2," *Infection, Genetics and Evolution*, vol. 81, pp. 1–3, July 2020. DOI: https://doi.org/10.1016/j.meegid.2020.104260.

[124] D. E. L. Promislow, "A Geroscience Perspective on COVID-19 Mortality," *The Journals of Gerontology: Series A*, vol. 75, e30–e33, September 16, 2020. DOI: https://doi.org/10.1093/gerona/glaa094.

[125] R. Ralph, J. Lew, T. Zeng, M. Francis, B. Xue, M. Roux, A. Toloue, S. Rubino, N. Dawe, M. Al-Ahdal, D. Kelvin, C. Richardson, J. Kindrachuk, D. Falzarano, and A. Kelvin, "2019-nCoV (Wuhan virus), a novel Coronavirus: human-to-human transmission, travel-related cases, and vaccine readiness," *The Journal of Infection in Developing Countries*, vol. 14, pp. 3–17, January 31, 2020. DOI: https://doi.org/10.3855/jidc.12425.

[126] D. Ray, L. Le, and I. Andricioaei, “Distant residues modulate conformational opening in SARS-CoV-2 spike protein,” *bioRxiv*, 2020. DOI: https://doi.org/10.1101/2020.12.07.415596.

[127] C. Rees-Spear, L. Muir, S. A. Griffith, J. Heaney, Y. Aldon, J. L. Snitselaar, P. Thomas, C. Graham, J. Seow, N. Lee, A. Rosa, C. Roustan, C. F. Houlihan, R. W. Sanders, R. K. Gupta, P. Cherepanov, H. J. Stauss, E. Nastouli, K. J. Doores, M. J. van Gils, and L. E. McCoy, “The effect of spike mutations on SARS-CoV-2 neutralization,” *Cell Reports*, vol. 34, pp. 1–21, March 23, 2021. DOI: https://doi.org/10.1016/j.celrep.2021.108890.

[128] P. Rice, I. Longden, and A. Bleasby, “EMBOSS: the European molecular biology open software suite,” *Trends Genet*, vol. 16, pp. 276–277, June 01, 2000. DOI: https://doi.org/10.1016/s0168-9525(00)02024-2.

[129] A. Rodrıguez-Gascón, A. del Pozo-Rodrıguez, and M. Á. Solinıs, “Development of nucleic acid vaccines: use of self-amplifying RNA in lipid nanoparticles,” *International Journal of Nanomedicine*, vol. 9, pp. 1–11, April 10, 2014. DOI: https://doi.org/10.2147/IJN.S39810.

[130] M. Romano, A. Ruggiero, F. Squeglia, G. Maga, and R. Berisio, “A Structural View at SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping,” *Cells*, vol. 9, pp. 1–22, May 20, 2020. DOI: https://doi.org/10.3390/cells9051267.

[131] S. Rosa and W. Santos, “Clinical trials on drug repositioning for COVID-19 treatment,” *Revista Panamericana de Salud Pública*, vol. 44, pp. 1–7, March 20, 2020. DOI: https://doi.org/10.26633/RPSP.2020.40.

[132] T. R. Ruch and C. E. Machamer, “The coronavirus E protein: assembly and beyond,” *Viruses*, vol. 4, pp. 363–382, March 08, 2012. DOI: https://doi.org/10.3390/v4030363.

[133] D.-K. Ryu, B. Kang, S.-J. Woo, M.-H. Lee, A. S. Tijsma, H. Noh, J.-I. Kim, J.-M. Seo, C. Kim, M. Kim, E. Yang, G. Lim, S.-G. Kim, S.-K. Eo, J.-a. Choi, S.-S. Oh, P. M. Nuijten, M. Song, H.-Y. Chung, C. A. van Baalen, K.-S. Kwon, and S.-Y. Lee, “Therapeutic efficacy of CT-P59 against P. 1 variant of SARS-CoV-2,” *bioRxiv*, July 09, 2021. DOI: https://doi.org/10.1101/2021.07.08.451696.

[134] P. Saha, R. Majumder, S. Chakraborty, and A. Kumar, “Mutations in Spike protein of SARS-CoV-2 modulate receptor binding, membrane fusion and immunogenicity: an insight into viral tropism and pathogenesis of COVID-19,” May 19, 2020. DOI: https://doi.org/10.26434/chemrxiv.12320567.

[135] N. Saitou and M. Nei, "The neighbor-joining method: A new method for reconstructing phylogenetic trees.," *Molecular Biology and Evolution*, vol. 4, pp. 406–425, July 01, 1987. DOI: https://doi.org/10.1093/oxfordjournals.molbev.a040454.

[136] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, January 2015. DOI: https://doi.org/10.1016/j.neunet.2014.09.003.

[137] D. Schoeman and B. C. Fielding, "Coronavirus envelope protein: Current knowledge," *Virology Journal*, vol. 16, pp. 1–22, May 27, 2019. DOI: https://doi.org/10.1186/s12985-019-1182-0.

[138] J. Shang, G. Ye, K. Shi, Y. Wan, C. Luo, H. Aihara, Q. Geng, A. Auerbach, and F. Li, "Structural basis of receptor recognition by SARS-CoV-2," *Nature*, vol. 581, pp. 221–224, March 30, 2020. DOI: https://doi.org/10.1038/s41586-020-2179-y.

[139] K. Shen, Y. Yang, T. Wang, D. Zhao, Y. Jiang, R. Jin, Y. Zheng, B. Xu, Z. Xie, L. Lin, Y. Shang, X. Lu, S. Shu, Y. Bai, J. Deng, M. Lu, L. Ye, X. Wang, Y. Wang, and L. Gao, "Diagnosis, treatment, and prevention of 2019 novel coronavirus infection in children: experts' consensus statement," *World Journal of Pediatrics*, vol. 16, pp. 223–231, February 07, 2020. DOI: https://doi.org/10.1007/s12519-020-00343-7.

[140] L. Shen, J. D. Bard, T. J. Triche, A. R. Judkins, J. A. Biegel, and X. Gai, "Emerging variants of concern in SARS-CoV-2 membrane protein: a highly conserved target with potential pathological and therapeutic implications," *Emerging Microbes & Infections*, vol. 10, pp. 885–893, May 09, 2021. DOI: https://doi.org/10.1080/22221751.2021.1922097.

[141] Y. Shu and J. McCauley, "GISAID: Global initiative on sharing all influenza data–from vision to reality," *Euro Surveillance*, vol. 22, pp. 1–3, March 30, 2017. DOI: https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

[142] A. Shulla, T. Heald-Sargent, G. Subramanya, J. Zhao, S. Perlman, and T. Gallagher, "A transmembrane serine protease is linked to the severe acute respiratory syndrome coronavirus receptor and activates virus entry," *Journal of Virology*, vol. 85, pp. 873–882, January 15, 2011. DOI: https://doi.org/10.1128/JVI.02062-10.

[143] A. Singh, G. Steinkellner, K. Köchl, K. Gruber, and C. C. Gruber, "Serine 477 plays a crucial role in the interaction of the SARS-CoV-2 spike protein with the human receptor ACE2," *Scientific Reports*, vol. 11, pp. 1–11, February 22, 2021. DOI: https://doi.org/10.1038/s41598-021-83761-5.

[144] A. K. Singh, A. Singh, A. Shaikh, R. Singh, and A. Misra, “Chloroquine and hydroxychloroquine in the treatment of COVID-19 with or without diabetes: A systematic search and a narrative review with a special reference to India and other developing countries,” *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, pp. 241–246, May-June 2020. DOI: https://doi.org/10.1016/j.dsx.2020.03.011.

[145] Z. Song, Y. Xu, L. Bao, L. Zhang, P. Yu, Y. Qu, H. Zhu, W. Zhao, Y. Han, and C. Qin, “From SARS to MERS, thrusting coronaviruses into the spotlight,” *Viruses*, vol. 11, pp. 1–23, January 14, 2019. DOI: https://doi.org/10.3390/v11010059.

[146] G. Spinato, C. Fabbris, J. Polesel, D. Cazzador, D. Borsetto, C. Hopkins, and P. Boscolo-Rizzo, “Alterations in smell or taste in mildly symptomatic outpatients with SARS-CoV-2 infection,” *Jama*, vol. 323, pp. 2089–2090, April 22, 2020. DOI: https://doi.org/10.1001/jama.2020.6771.

[147] S. Su, G. Wong, W. Shi, J. Liu, A. C. Lai, J. Zhou, W. Liu, Y. Bi, and G. F. Gao, “Epidemiology, genetic recombination, and pathogenesis of coronaviruses,” *Trends in Microbiology*, vol. 24, pp. 490–502, June 01, 2016. DOI: https://doi.org/10.1016/j.tim.2016.03.003.

[148] R. Sumirtanurdin and M. I. Barliana, “Coronavirus disease 2019 vaccine development: An overview,” *Viral Immunology*, vol. 34, pp. 134–144, April 16, 2021. DOI: https://doi.org/10.1089/vim.2020.0119.

[149] I. Sutskever, O. Vinyals, and Q. V. Le, *Sequence to sequence learning with neural networks*, December 14, 2014. arXiv: 1409.3215v3 [cs.CL].

[150] W. Tai, L. He, X. Zhang, J. Pu, D. Voronin, S. Jiang, Y. Zhou, and L. Du, “Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine,” *Cellular & Molecular Immunology*, vol. 17, pp. 613–620, March 19, 2020. DOI: https://doi.org/10.1038/s41423-020-0400-4.

[151] W. Tai, X. Zhang, Y. He, S. Jiang, and L. Du, “Identification of SARS-CoV RBD-targeting monoclonal antibodies with cross-reactive or neutralizing activity against SARS-CoV-2,” *Antiviral Research*, vol. 179, pp. 1–6, July 2020. DOI: https://doi.org/10.1016/j.antiviral.2020.104820.

[152] C. N. Thompson, S. Hughes, S. Ngai, J. Baumgartner, J. C. Wang, E. McGibbon, K. Devinney, E. Luoma, D. Bertolino, C. Hwang, K. Kepler, C. D. Castillo, M. Hopkins, H. lee, A. K. DeVito, J. L. Rakeman, and A. D. Fine, “Rapid Emergence and Epidemiologic Characteristics of the SARS-CoV-2 B. 1.526 Variant—New York City, New York, January 1–

April 5, 2021," *Morbidity and Mortality Weekly Report*, vol. 70, pp. 712–716, May 14, 2021. DOI: https://doi.org/10.15585/mmwr.mm7019e1.

[153] Y.-T. Tseng, C.-H. Chang, S.-M. Wang, K.-J. Huang, and C.-T. Wang, "Identifying SARS-CoV membrane protein amino acid residues linked to virus-like particle assembly," *PLOS ONE*, vol. 8, May 20, 2013. DOI: https://doi.org/10.1371/journal.pone.0064013.

[154] M. Uddin, F. Mustafa, T. Rizvi, T. Loney, H. Al Suwaidi, A. Al-Marzouqi, A. Eldin, N. Alsabeeha, T. Adrian, C. Stefanini, N. Nowotny, A. Alsheikh-Ali, and A. Senok, "SARS-CoV-2/COVID-19: Viral genomics, epidemiology, vaccines, and therapeutic interventions," *Viruses*, vol. 12, pp. 1–18, May 10, 2020. DOI: https://doi.org/10.3390/v12050526.

[155] O. M. UĞUREL, O. ATA, and D. BALIK, "An updated analysis of variations in SARS-CoV-2 genome," *Turkish Journal of Biology*, vol. 44, pp. 157–167, June 21, 2020. DOI: https://doi.org/10.3906/biy-2005-111.

[156] M. Ujike and F. Taguchi, "Incorporation of spike and membrane glycoproteins into coronavirus virions," *Viruses*, vol. 7, pp. 1700–1725, April 03, 2015. DOI: https://doi.org/10.3390/v7041700.

[157] L. van Dorp, M. Acman, D. Richard, L. P. Shaw, C. E. Ford, L. Ormond, C. J. Owen, J. Pang, C. C. Tan, F. A. Boshier, A. T. Ortiz, and F. Balloux, "Emergence of genomic diversity and recurrent mutations in SARS-CoV-2," *Infection, Genetics and Evolution*, vol. 83, pp. 1–9, 2020. DOI: https://doi.org/10.1016/j.meegid.2020.104351.

[158] T. P. Velavan and C. G. Meyer, "The COVID-19 epidemic," *Tropical Medicine & International Health*, vol. 25, pp. 278–280, February 12, 2020. DOI: https://doi.org/10.1111/tmi.13383.

[159] P. Venkatagopalan, S. M. Daskalova, L. A. Lopez, K. A. Dolezal, and B. G. Hogue, "Coronavirus envelope (E) protein remains at the site of assembly," *Virology*, vol. 478, pp. 75–85, April 2015. DOI: https://doi.org/10.1016/j.virol.2015.02.005.

[160] K. Wang, P. Zuo, Y. Liu, M. Zhang, X. Zhao, S. Xie, H. Zhang, X. Chen, and C. Liu, "Clinical and laboratory predictors of in-hospital mortality in patients with coronavirus disease-2019: a cohort study in Wuhan, China," *Clinical infectious diseases*, vol. 71, pp. 2079–2088, October 15, 2020. DOI: https://doi.org/10.1093/cid/ciaa538.

[161] P. Wang, R. G. Casner, M. S. Nair, M. Wang, J. Yu, G. Cerutti, L. Liu, P. D. Kwong, Y. Huang, L. Shapiro, and D. D. Ho, "Increased resistance of SARS-CoV-2 variant P. 1 to antibody neutralization," *bioRxiv*, April 09, 2021. DOI: https://doi.org/10.1101/2021.03.01.433466.

[162] P. Wang, M. S. Nair, L. Liu, S. Iketani, Y. Luo, Y. Guo, M. Wang, J. Yu, B. Zhang, P. D. Kwong, *et al.*, "Antibody resistance of SARS-CoV-2 variants B. 1.351 and B. 1.1. 7," *Nature*, vol. 593, pp. 130–135, March 08, 2021. DOI: https://doi.org/10.1038/s41586-021-03398-2.

[163] Q. Wang, Y. Zhang, L. Wu, S. Niu, C. Song, Z. Zhang, G. Lu, C. Qiao, Y. Hu, K.-Y. Yuen, *et al.*, "Structural and functional basis of SARS-CoV-2 entry by using human ACE2," *Cell*, vol. 181, pp. 894–904, April 09, 2020. DOI: https://doi.org/10.1016/j.cell.2020.03.045.

[164] J. Warwicker, "A model for pH coupling of the SARS-CoV-2 spike protein open/closed equilibrium," *Briefings in Bioinformatics*, vol. 22, pp. 1499–1507, February 25, 2021. DOI: https://doi.org/10.1093/bib/bbab056.

[165] R. Watanabe, S. Matsuyama, K. Shirato, M. Maejima, S. Fukushi, S. Morikawa, and F. Taguchi, "Entry from the cell surface of severe acute respiratory syndrome coronavirus with cleaved S protein as revealed by pseudotype virus bearing cleaved S protein," *Journal of Virology*, vol. 82, pp. 11 985–11 991, December 1, 2008. DOI: https://doi.org/10.1128/JVI.01412-08.

[166] A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton, "Jalview version 2—a multiple sequence alignment editor and analysis workbench," *Bioinformatics*, vol. 25, pp. 1189–1191, January 16, 2009. DOI: https://doi.org/10.1093/bioinformatics/btp033.

[167] W. E. Wei, Z. Li, C. J. Chiew, S. E. Yong, M. P. Toh, and V. J. Lee, "Presymptomatic Transmission of SARS-CoV-2—Singapore, January 23–March 16, 2020," *Morbidity and Mortality Weekly Report*, vol. 69, pp. 411–415, April 10, 2020. DOI: https://doi.org/10.15585/mmwr.mm6914e1.

[168] M. C. Wong, S. J. J. Cregeen, N. J. Ajami, and J. F. Petrosino, "Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019," *bioRxiv*, February 13, 2020. DOI: https://doi.org/10.1101/2020.02.07.939207.

[169] P. C. Woo, Y. Huang, S. K. Lau, and K.-Y. Yuen, "Coronavirus genomics and bioinformatics analysis," *Viruses*, vol. 2, pp. 1804–1820, August 24, 2010. DOI: https://doi.org/10.3390/v2081803.

[170] A. G. Wrobel, D. J. Benton, P. Xu, C. Roustan, S. R. Martin, P. B. Rosenthal, J. J. Skehel, and S. J. Gamblin, "SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects," *Nature Structural & Molecular Biology*, vol. 27, pp. 763–767, July 09, 2020. DOI: https://doi.org/10.1038/s41594-020-0468-7.

[171] A. Wu, Y. Peng, B. Huang, X. Ding, X. Wang, P. Niu, J. Meng, Z. Zhu, Z. Zhang, J. Wang, *et al.*, "Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China," *Cell Host & Microbe*, vol. 27, pp. 325–328, March 11, 2020. DOI: https://doi.org/10.1016/j.chom.2020.02.001.

[172] C. Wu, Y. Liu, Y. Yang, P. Zhang, W. Zhong, Y. Wang, Q. Wang, Y. Xu, M. Li, X. Li, M. Zheng, L. Chen, and H. Li, "Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods," *Acta Pharmaceutica Sinica B*, vol. 10, pp. 766–788, May 2020. DOI: https://doi.org/10.1016/j.apsb.2020.02.008.

[173] Y.-C. Wu, C.-S. Chen, and Y.-J. Chana, "The outbreak of COVID-19: An overview," *J Chin Med Assoc*, vol. 83, pp. 217–220, March 2020. DOI: https://doi.org/10.1097/JCMA.0000000000000270.

[174] F. Wu, S. Zhao, B. Yu, Y. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. Holmes, and Y.-Z. Zhang, "A new coronavirus associated with human respiratory disease in China," *Nature*, vol. 579, pp. 265–269, February 03, 2020. DOI: https://doi.org/10.1038/s41586-020-2008-3.

[175] C. Xiong, L. Jiang, Y. Chen, and Q. Jiang, "Evolution and variation of 2019-novel coronavirus," *bioRxiv*, January 30, 2020. DOI: https://doi.org/10.1101/2020.01.30.926477.

[176] J. Xu, S. Zhao, T. Teng, A. E. Abdalla, W. Zhu, L. Xie, Y. Wang, and X. Guo, "Systematic comparison of two animal-to-human transmitted human coronaviruses: SARS-CoV-2 and SARS-CoV," *Viruses*, vol. 12, pp. 1–17, February 22, 2020. DOI: https://doi.org/10.3390/v12020244.

[177] L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jin, M. Zhang, X. Huang, Y. Xiao, H. Cao, Y. Chen, T. Ren, F. Wang, Y. Xiao, S. Huang, X. Tan, N. Huang, B. Jiao, Y. Zhang, A. Luo, L. Mombaerts, J. Jin, Z. Cao, S. Li, H. Xu, and Y. Yuan, "A machine learning-based model for survival prediction in patients with severe COVID-19 infection," *medRxiv*, March 17, 2020. DOI: https://doi.org/10.1101/2020.02.27.20028027.

[178] Y. Yan, Z. Zou, Y. Sun, X. Li, K.-F. Xu, Y. Wei, N. Jin, and C. Jiang, "Anti-malaria drug chloroquine is highly effective in treating avian influenza A H5N1 virus infection in an animal model," *Cell Research*, vol. 23, pp. 300–302, February 2013. DOI: https://doi.org/10.1038/cr.2012.165.

[179] J. Yang, "Inhibition of SARS-CoV-2 Replication by Acidizing and RNA Lyase-Modified Carbon Nanotubes Combined with Photodynamic Thermal Effect," *Journal of Exploratory Research in Pharmacology*, vol. 5, pp. 18–23, April 2020. DOI: https://doi.org/10.14218/JERP.2020.00005.

[180] X. Yang, Y. Yu, J. Xu, H. Shu, J. Xia, H. Liu, Y. Wu, L. Zhang, Z. Yu, M. Fang, T. Yu, Y. Wang, S. Pan, X. Zou, S. Yuan, and Y. Shang, "Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study," *The Lancet Respiratory Medicine*, vol. 8, pp. 475–481, May 01, 2020. DOI: https://doi.org/10.1016/S2213-2600(20)30079-5.

[181] Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai, J. Liang, X. Liu, S. Li, Y. Li, F. Ye, W. Guan, Y. Yang, F. Li, S. Luo, Y. Xie, B. Liu, Z. Wang, S. Zhang, Y. Wang, N. Zhong, and J. He, "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions," *Journal of Thoracic Disease*, vol. 12, pp. 165–174, March 2020. DOI: https://doi.org/10.21037/jtd.2020.02.64.

[182] T.-T. Yao, J.-D. Qian, W.-Y. Zhu, Y. Wang, and G.-Q. Wang, "A systematic review of lopinavir therapy for SARS coronavirus and MERS coronavirus—A possible reference for coronavirus disease-19 treatment option," *Journal of Medical Virology*, vol. 92, pp. 556–563, February 2020. DOI: https://doi.org/10.1002/jmv.25729.

[183] T.-Y. Yeh and G. P. Contreras, "Viral transmission and evolution dynamics of SARS-CoV-2 in shipboard quarantine," *Bulletin of the World Health Organization*, vol. 99, pp. 486–495, April 30, 2020. DOI: https://doi.org/10.2471/BLT.20.255752.

[184] Y. Yin and R. G. Wunderink, "MERS, SARS and other coronaviruses as causes of pneumonia," *Respirology*, vol. 23, pp. 130–137, February 2018. DOI: https://doi.org/10.1111/resp.13196.

[185] C. Y. Yong, H. K. Ong, S. K. Yeap, K. L. Ho, and W. S. Tan, "Recent Advances in the Vaccine Development Against Middle East Respiratory Syndrome-Coronavirus," *Frontiers in Microbiology*, vol. 10, pp. 1–18, August 02, 2019. DOI: https://doi.org/10.3389/fmicb.2019.01781.

[186] M. Zandi, S. Saber, S. Sanami, and A. Rasooli, "Spike Protein Mutations and the Effects on SARS-CoV-2 Pathogenesis," *Journal of Cellular & Molecular Anesthesia*, vol. 6, pp. 148–153, April 16, 2021. DOI: https://doi.org/10.22037/jcma.v6i2.33800.

[187] C.-Y. Zhang, J.-F. Wei, and S.-H. He, "Adaptive evolution of the spike gene of SARS coronavirus: changes in positively selected sites in different epidemic groups," *BMC Microbiology*, vol. 6, pp. 1–10, October 04, 2006. DOI: https://doi.org/10.1186/1471-2180-6-88.

[188] J.-j. Zhang, X. Dong, Y.-Y. Cao, Y.-d. Yuan, Y.-b. Yang, Y.-q. Yan, C. Akdis, and Y.-D. Gao, "Clinical characteristics of 140 patients infected by SARS-CoV-2 in Wuhan, China," *Allergy*, vol. 75, pp. 1730–1741, February 19, 2020. DOI: https://doi.org/10.1111/all.14238.

[189] J. Zhang, D. Finlaison, M. Frost, S. Gestier, X. Gu, J. Hall, C. Jenkins, K. Parrish, A. Read, M. Srivastava, K. Rose, and P. Kirkland, "Identification of a novel nidovirus as a potential cause of large scale mortalities in the endangered Bellinger River snapping turtle (*Myuchelys georgesi*)," *PLOS ONE*, vol. 13, pp. 1–19, October 24, 2018. DOI: https://doi.org/10.1371/journal.pone.0205209.

[190] L. Zhang, C. Jackson, H. Mou, A. Ojha, H. Peng, B. Quinlan, R. Erumbi, A. Pan, A. Vanderheiden, M. Suthar, W. Li, T. Izard, C. Rader, M. Farzan, and H. Choe, "SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity," *Nature Communications*, vol. 11, pp. 1–9, November 26, 2020. DOI: https://doi.org/10.1038/s41467-020-19808-4.

[191] A.-R. Zhang, W.-Q. Shi, K. Liu, X.-L. Li, M.-J. Liu, W.-H. Zhang, G.-P. Zhao, J.-J. Chen, X.-A. Zhang, D. Miao, W. Ma, W. Liu, Y. Yang, and L.-Q. Fang, "Epidemiology and evolution of Middle East respiratory syndrome coronavirus, 2012–2020," *Infectious Diseases of Poverty*, vol. 10, pp. 1–13, May 08, 2021. DOI: https://doi.org/10.1186/s40249-021-00853-0.

[192] T. Zhang, Q. Wu, and Z. Zhang, "Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak," *Current Biology*, vol. 30, pp. 1346–1351, April 20, 2020. DOI: https://doi.org/10.1016/j.cub.2020.03.063.

[193] Y.-Z. Zhang and E. C. Holmes, "A genomic perspective on the origin and emergence of SARS-CoV-2," *Cell*, vol. 181, pp. 223–227, April 16, 2020. DOI: https://doi.org/10.1016/j.cell.2020.03.035.

[194] N. Zhong, B. Zheng, Y. Li, L. Poon, Z. Xie, K.-H. Chan, P. Li, S. Tan, Q. Chang, J. Xie, X. Liu, J. Xu, D. Li, Y. ky, J. S. Peiris, and Y. Guan, "Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003," *The Lancet*, vol. 362, pp. 1353–1358, October 25, 2003. DOI: https://doi.org/10.1016/S0140-6736(03)14630-2.

[195] P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang, *et al.*, “A pneumonia outbreak associated with a new coronavirus of probable bat origin,” *Nature*, vol. 579, pp. 270–273, February 03, 2020. DOI: https://doi.org/10.1038/s41586-020-2012-7.

[196] W. Zhou and W. Wang, “Fast-spreading SARS-CoV-2 variants: challenges to and new design strategies of COVID-19 vaccines,” *Signal Transduction and Targeted Therapy*, vol. 6, pp. 1–6, 2021. DOI: https://doi.org/10.1038/s41392-021-00644-x.