**Coronavirus worldwide variants study comparisons in distribution comparison and structural differences of SARS-CoV-2 spike(S) protein RBDs**

*Ravi Potlapalli1, Thanusha Sai Ande1, Kimberly Yanchen Wang1, Jiawei Chen2,*

*Abhimanyu Malik1, Gary Xue1, Lichao Liu3, Ying Xie1, Chloe Yixin Xie1\**

1College of Computing and Software Engineering, Kennesaw State University, Marietta, GA 30060

2 College of Letter and Science, University of California, Berkeley, CA 94720

3 Stanford Cardiovascular Institute, Stanford University School of Medicine, Palo Alto, CA 94304

\*Correspondence to Chloe Yixin Xie (E-mail: [yxie11@kennesaw.edu](mailto:yxie11@kennesaw.edu))

Article Outline:

1. Introduction: COVID-19 background, variants, recent research, current methods, etc.
2. Current variants map distribution, data visualizations, etc.
3. Different variants records, like the recovery time comparisons, symptoms comparisons, etc.
4. Structural comparisons, surface comparisons, etc.
5. Potential calculations, force calculations, etc.
6. Binding comparison, with clinical data comprehensive analysis.
7. Conclusion: show the importance of our work.

Introduction Outline:

Outline: 1) why covid-19 related topics are important; 2) overview of covid-19 variants; 3) explain why we are interested in studying COVID-19 variants; 4) data science applications in covid-19 study, and the advantage of using computational tools to answer biological questions; 5) our approaches in studying covid-19 variants

KEYWORDS

COVID-19, SARS-CoV-2, Coronavirus, Protein structure, Data Analysis, Data Visualization, Distribution

ABSTRACT:

The COVID-19 pandemic has presented a significant challenge to the global healthcare system, requiring innovative and data-driven solutions to combat its spread. In this article, we explore the application of data science tools such as Python, SQL, and Power BI in analyzing and visualizing COVID-19 data. We first discuss the challenges in working with COVID-19 data, including data quality and completeness issues, and the need for timely and accurate reporting. We then outline the various data science techniques used in the analysis, including data cleaning, feature engineering, and machine learning algorithms. We demonstrate the power of these tools in visualizing and analyzing COVID-19 data, including trends in infection rates, mortality rates, and vaccination rates across different regions and demographics. Furthermore, we showcase how data science can be leveraged to model the spread of COVID-19 and predict its impact on healthcare systems. Finally, we discuss the importance of data transparency and collaboration in the fight against COVID-19, emphasizing the need for open data sharing and standardization. Our article highlights the potential of data science to provide insights into the COVID-19 pandemic and inform policy decisions, paving the way for more effective and data-driven responses to future healthcare challenges.

1. INTRODUCTION

In the year 2019 covid-19 was discovered which is an infection with the respiratory virus SARS-CoV-2 in the Chinese city of Wuhan. The Latin word "corona," which means "crown~~,~~" is the source of the term "coronavirus"[1, 2].It quickly expanded in the subsequent months to South Korea, Singapore, Thailand, Japan, and Iran[3]. This was followed by widespread viral spread throughout the world, including in the United Arab Emirates, Italy, the United States, Spain and the United Kingdom. SARS-CoV-2 is a novel coronavirus that infects both humans and other vertebrate animals, causing damage to the respiratory, intestinal, liver, and central nervous systems[4-6]. It belongs to the CoV's newest evolutionary branch. The virus spreads mostly through respiratory droplets and microscopic particles expelled during sneezing, coughing, or talking by an infected person[7, 8]. In crowded or inadequately ventilated interior environments, the virus can easily spread. Mild to severe illness is possible, though not everyone who contracts the virus exhibits symptoms. Serious sickness is more likely to affect those in their sixties and older and people of any age who have underlying medical issues[9, 10]. Since changes to the genetic code (from genetic mutations) take place at some stage in genome replication, SARS-CoV-2 is constantly evolving[11]. A lineage is a collection of genetically distinct virus variants that have a common ancestor. A variety of the SARS-CoV-2 viruses has one or more mutations (A single alteration to a virus's genome is referred to as a mutation.) that set it apart from other variants. Recombinant means combining genetic material from two separate variants to form a new variant. Numerous SARS-CoV-2 variants have been reported in the global laboratory as expected. A lineage or set of lineages may be categorized as a variant of concern (VOC), a variant of interest (VOI), Variant being monitored (VBM) due to shared traits, features that may call for public health intervention. The SARS-CoV-2 Interagency Group (SIG) variant classification scheme defines four classes of SARS-CoV-2 variants: Variant Being Monitored (VBM): Iota (B.1.526),Kappa (B.1.617.1),1.617.3,Beta (B.1.351 and descendent lineages), Delta (B.1.617.2 and AY lineages), Alpha (B.1.1.7 and Q lineages), Epsilon (B.1.427 and B.1.429), Eta (B.1.525),Gamma (P.1 and descendent lineages), Mu (B.1.621, B.1.621.1),Zeta (P.2), Variant of Interest (VOI):Till now there are not any variant of this group, Variant of Concern (VOC):Omicron (BA.1, BA.1.1, BA.2, BA.3, BA.4 ,BA.5, B.1.1.529, BQ.1.1 and XBB.1 lineages)[12-14]Variant of High Consequence (VOHC): Till now at hand are not any variant of this group. The quality of life and lifespan have been constrained by infections, leading to a rise in morbidity and needless deaths. The development of public hygiene, vaccinations, and antibiotics has decreased the likelihood of contracting an infectious disease, but recent occurrences have caused panic due to the deluge of information on prospective epidemics. The standardization of molecular biology in diagnostic testing has altered the research of respiratory viral diseases, but the level of expertise of individuals using molecular techniques does not significantly affect them. Assays that can identify much larger panels of infectious organisms are now being developed[15].

In 73rd world health Assembly held on 19 May 2020; the WHO's Director-General spoke about the coronavirus disease pandemic (COVID-19). He is  extremely disturbed by the COVID-19 pandemic's morbidity and mortality, adverse effects on social, mental, and physical health, adverse effects on economies, and aggravation of global disparities as a result expressing sympathy to all the COVID-19 victim families and solidarity with all the pandemic-affected nations, highlighting that governments have the primary responsibility for adopting and executing COVID-19 pandemic responses that are unique to their country situation and for mobilizing the resources necessary. He also advised to provide access to personal safety gear, other essential supplies, and instruction, including by offering psychosocial support; to take action to ensure their safety at work, facilitating their access to employment and ensuring their adequate compensation and to consider implementing task-sharing and task-shifting to maximize the usage of resources[16]. The COVID-19 is a pandemic with 3 732 046 confirmed cases and 261 517 fatalities as of May 6, 2020. It is a member of the Nidovirales order and spreads through contact and droplet transfer. During July to November 2022, the terrible acute respiratory syndrome coronavirus 2(SARS-CoV-2) omicron BA.5 subvariant dominated, and it significantly outperformed earlier variations in terms of neutralization escape[17, 18]. Two further omicron variations the BA.2 lineage recombinant XBB.1, the BA.5 sub lineages BF.7 and BQ.1.1, the BA.2 sub lineage BA.2.75.2, the BA.4 sub lineage BA.4.6 have just been discovered.

Our work in this paper.

1. METHODS

**2.1. Data preparation: source, cleaning, selections**

**GISAID database, how did we get access to download, whats the standard to select, how did we clean data, etc.**

Data preparation is a crucial step in the data analysis process. Source cleaning involves identifying and addressing errors, inconsistencies, or missing data in the source data. Selection involves identifying the relevant data for the analysis and discarding irrelevant or redundant data. Both source cleaning and selection are essential for accurate and meaningful analysis, as they ensure that the data used is accurate, consistent, and relevant. By investing time and effort into data preparation, analysts can ensure that their results are reliable, meaningful, and actionable[19].In this project we used data from the GISAID database. There are a few easy steps you must take in order to obtain COVID-19 datasets from the GISAID website. First, visit the GISAID website and click on the "EpiCoV™" tab in the top menu bar. Next, you will need to agree to the terms and conditions by clicking on the "Acknowledgement of Data Access Responsibilities" button. Then, select the "Global database" option to access the dataset. To search for specific COVID-19 datasets, enter "SARS-CoV-2" in the search box and select the desired dataset. Once you have selected your dataset, click on the "Download" button. You will be prompted to fill out a "Registration Form" with the necessary information, and then click on "Submit Registration". It is important to note that access to GISAID's data is restricted to registered users, and approval for access may take some time. Once your registration is approved, you will receive an email with a link to download the dataset.

**2.2. Data analysis: initial conclusion**

**Excel, SQL, etc.**

Data analysis is an essential process for businesses and organizations to make informed decisions. Excel and SQL are two popular tools used for data analysis. Excel is a spreadsheet application that provides various functions and tools for data manipulation, calculation, and visualization. SQL is a programming language used to manage and manipulate relational databases. Excel is useful for small to medium-sized datasets or for quick and simple data analysis. Excel provides an easy-to-use interface with features such as charts, pivot tables, and filters that allow users to visualize and analyze data. However, Excel can be limited in terms of handling large datasets or complex queries[20]. SQL, on the other hand, is designed for managing and analyzing large datasets. SQL provides more advanced capabilities for complex data queries and can handle datasets with millions of records. SQL is highly efficient in handling relational databases and can be used to retrieve, manipulate, and transform data. However, SQL requires a steep learning curve compared to Excel and may require programming skills. Excel is useful for small to medium-sized datasets or for quick and simple data analysis, while SQL is designed for managing and analyzing large datasets with more advanced capabilities. Choosing the right tool ultimately depends on the size and complexity of the data and the specific needs of the analysis[17, 18].

**2.3. Data visualization: line graph, circle pie graph, map distribution, etc.**

**PowerBI, python, etc.**

Power BI is a powerful data visualization tool that can help you design and visualize your data in a user-friendly and interactive way. To begin, you would need to connect to your data source by selecting the type of source you want to use such as Excel, CSV, SQL Server, or Oracle. Once your data is loaded, you can create a report by selecting the type of visualization you want to use such as bar chart, line chart, pie chart, or map. You can add multiple visualizations to a single report and drag and drop fields from your data source onto the visualizations to add data to the report. Power BI also provides a lot of customization options for your visualizations such as changing the color, font, title, and other properties to make them more appealing and informative. Once your report is complete, you can create a dashboard by pinning the visualizations to a single page, providing an overview of your data, and allowing you to interact with the visualizations. Finally, you can publish your report to the Power BI service and share it with your colleagues or clients or embed the report in other applications or websites[21]**.**

Python is a powerful programming language that is commonly used for data analysis and visualization. There are several libraries in Python that make it easy to design and visualize data. In this paragraph, we will discuss how to design and visualize data using Python. One of the most popular libraries for data visualization in Python is Matplotlib. Matplotlib is a 2D plotting library that can be used to create a wide range of visualizations, including line plots, scatter plots, bar charts, and histograms. Matplotlib provides a wide range of customization options, such as adjusting axis labels, adding legends, and changing colors and fonts. Another popular library for data visualization in Python is Seaborn. Seaborn is built on top of Matplotlib and provides a higher-level interface for creating statistical visualizations. Seaborn provides a range of built-in functions for creating visualizations such as scatterplots, line plots, heatmaps, and bar charts. Pandas is another popular library for data analysis and visualization in Python. Pandas provide data structures and functions for manipulating numerical tables and time series. Pandas can be used to load and clean data from various sources and perform data analysis and visualization.Overall, Python provides a range of powerful tools for data analysis and visualization. By using libraries such as Matplotlib, Seaborn, and Pandas, you can easily design and visualize data in Python[22].

**2.4. Spike(s) Protein structure preparation**

**Protein data bank, VOC, VOI, etc.**

The spike protein is a key component of the SARS-CoV-2 virus and is the target of most COVID-19 vaccines. Structural preparation of the spike protein involves several steps to ensure that the protein structure is ready for further analysis, such as molecular dynamics simulations. The first step in structural preparation is to obtain a high-resolution structure of the spike protein using X-ray crystallography or cryogenic electron microscopy (cryo-EM)[23-25]. The obtained structure is usually in a complex form, which includes the spike protein and other proteins, such as the receptor-binding domain or the ACE2 receptor[26]. The next step is to isolate the spike protein from the complex structure using molecular visualization software, such as PyMOL or Chimera. The isolated spike protein is then subjected to energy minimization, which involves optimizing the positions of the atoms to minimize the energy of the system. This is done to remove any steric clashes or other structural abnormalities. In short, structural preparation of the spike protein involves several steps, including obtaining a high-resolution structure, isolating the spike protein, energy minimization, solvation, and equilibration[27]. These steps ensure that the spike protein structure is ready for further analysis and provide a basis for the design and development of effective COVID-19 vaccines and therapeutics using machine learning Techniques by using libraries such as PyTorch, TorchDrug.

1. RESULTS

**3.1. Map distribution**

Map

Description automatically generated

**Figure:3.1.1: Geographical Design for Omicron 2023-January**

**Figure:3.1.2: Geographical Design for Omicron 2023-** **February**

**Figure:3.1.3: Geographical Design for Omicron 2023-March**

**3.2. Dominant SARS-CoV-2 variant in selected countries**

A picture containing transport, wheel

Description automatically generated

**Figure:3.2.1: Dominant variant with respect to year in United Kingdom**

**Chart, sunburst chart

Description automatically generated**

**Figure:3.2.2: Dominant variant with respect to year in USA**

**Diagram

Description automatically generated**

**Figure:3.2.3: Dominant variant with respect to year in India**

**3.3. Comparisons among selected countries regarding one certain variant**

Chart, bar chart, waterfall chart

Description automatically generated

**Figure:3.3.1: Percentage of variant distribution with respect to years.**

**3.4. s protein structural comparison**

**Map

Description automatically generated with medium confidence**

**Figure:3.4.1: Structural comparison of Spike-protein mutation on Variants of Concern**

1. CONCLUSION

Our work indicates what...

Our future work will be on …

 The COVID-19 pandemic has posed a significant challenge to the global healthcare system. Despite the increasing number of vaccinations, we cannot predict the future cases and evolution of Covid. In this article, the ongoing COVID-19 pandemic has concluded Dominant variant among top Variants regarding year and map distribution of Omicron during January, February, and march of 2023. It also highlighted the need for further research on the mechanisms of SARS-CoV-2 Spike protein binding with ACE2 to develop effective drug treatments. One of the best approaches involves using machine learning techniques, such as binding behavior and pattern matching of various protein structures, with Biotite, PyTorch, and Torch Drug. These libraries in python can aid in predicting protein-protein interactions using graph convolutional networks and feed-forward neural networking. The mutual attention networks can extract relevant geometric features related to query protein pairs. Such research will not only help design effective treatments for COVID-19 but also other coronavirus-caused diseases. Finally, machine learning techniques can be employed to predict protein-protein interactions for further research in drug design, and future work in this area has suggesting applications for public health.

ACKNOWLEDGMENT

This research is funded by Grant SC1GM132043-01 from National Institutes of Health (NIH); Grant 5U54MD007592 from the National Institutes on Minority Health and Health Disparities (NIMHD), a component of the NIH; The project was partially supported by National Science Foundation IIS-1924092.

References:

1. Pagani, I., et al., *Origin and evolution of SARS-CoV-2.* Eur Phys J Plus, 2023. **138**(2): p. 157.

2. Singh, D. and S.V. Yi, *On the origin and evolution of SARS-CoV-2.* Exp Mol Med, 2021. **53**(4): p. 537-547.

3. Wu, F., et al., *A new coronavirus associated with human respiratory disease in China.* Nature, 2020. **579**(7798): p. 265-269.

4. Bui, L.T., et al., *Chronic lung diseases are associated with gene expression programs favoring SARS-CoV-2 entry and severity.* Nat Commun, 2021. **12**(1): p. 4314.

5. Sweed, D., et al., *SARS-CoV-2-associated gastrointestinal and liver diseases: what is known and what is needed to explore.* Egypt Liver J, 2021. **11**(1): p. 64.

6. Wieske, L., et al., *Risk factors associated with short-term adverse events after SARS-CoV-2 vaccination in patients with immune-mediated inflammatory diseases.* BMC Med, 2022. **20**(1): p. 100.

7. Ning, S., et al., *SARS-CoV-2: Origin, Evolution, and Targeting Inhibition.* Front Cell Infect Microbiol, 2021. **11**: p. 676451.

8. Voskarides, K., *SARS-CoV-2: tracing the origin, tracking the evolution.* BMC Med Genomics, 2022. **15**(1): p. 62.

9. Safiabadi Tali, S.H., et al., *Tools and Techniques for Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)/COVID-19 Detection.* Clin Microbiol Rev, 2021. **34**(3).

10. Weiss, S.R. and S. Navas-Martin, *Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus.* Microbiol Mol Biol Rev, 2005. **69**(4): p. 635-64.

11. Xu, J., et al., *Systematic Comparison of Two Animal-to-Human Transmitted Human Coronaviruses: SARS-CoV-2 and SARS-CoV.* Viruses, 2020. **12**(2).

12. Hadj Hassine, I., *Covid-19 vaccines and variants of concern: A review.* Rev Med Virol, 2022. **32**(4): p. e2313.

13. Rotondo, J.C., et al., *SARS-CoV-2 Infection: New Molecular, Phylogenetic, and Pathogenetic Insights. Efficacy of Current Vaccines and the Potential Risk of Variants.* Viruses, 2021. **13**(9).

14. Fiolet, T., et al., *Comparing COVID-19 vaccines for their characteristics, efficacy and effectiveness against SARS-CoV-2 and variants of concern: a narrative review.* Clin Microbiol Infect, 2022. **28**(2): p. 202-221.

15. Filardo, S., et al., *Oxidative Stress and Inflammation in SARS-CoV-2- and Chlamydia pneumoniae-Associated Cardiovascular Diseases.* Biomedicines, 2021. **9**(7).

16. Raoult, D., et al., *Coronavirus infections: Epidemiological, clinical and immunological features and hypotheses.* Cell Stress, 2020. **4**(4): p. 66-75.

17. Shao, W., et al., *Effectiveness of COVID-19 vaccines against SARS-CoV-2 variants of concern in real-world: a literature review and meta-analysis.* Emerg Microbes Infect, 2022. **11**(1): p. 2383-2392.

18. Zeng, B., et al., *Effectiveness of COVID-19 vaccines against SARS-CoV-2 variants of concern: a systematic review and meta-analysis.* BMC Med, 2022. **20**(1): p. 200.

19. Van den Broeck, J., et al., *Data cleaning: detecting, diagnosing, and editing data abnormalities.* PLoS Med, 2005. **2**(10): p. e267.

20. Trepalin, S.V. and A.V. Yarkov, *CheD: chemical database compilation tool, Internet server, and client for SQL servers.* J Chem Inf Comput Sci, 2001. **41**(1): p. 100-7.

21. van Duijn, S., et al., *Public-private partnership to rapidly strengthen and scale COVID-19 response in Western Kenya.* Front Public Health, 2022. **10**: p. 837215.

22. Ziegler, C.G.K., et al., *Impaired local intrinsic immunity to SARS-CoV-2 infection in severe COVID-19.* Cell, 2021. **184**(18): p. 4713-4733 e22.

23. Sheng, Y., et al., *Host Cell Glycocalyx Remodeling Reveals SARS-CoV-2 Spike Protein Glycomic Binding Sites.* Front Mol Biosci, 2022. **9**: p. 799703.

24. Xie, Y., et al., *The pH Effects on SARS-CoV and SARS-CoV-2 Spike Proteins in the Process of Binding to hACE2.* Pathogens, 2022. **11**(2).

25. Xie, Y. and M. Butler, *Quantitative profiling of N-glycosylation of SARS-CoV-2 spike protein variants.* Glycobiology, 2023. **33**(3): p. 188-202.

26. Singh, V., et al., *Impact of Video Modules-Based Training on Knowledge, Attitude, and Practices of Cleaning and Disinfection Among Housekeeping Staff at a Tertiary Care Center During the COVID-19 Pandemic.* Cureus, 2021. **13**(10): p. e19125.

27. Jia, Z. and W. Gong, *Will Mutations in the Spike Protein of SARS-CoV-2 Lead to the Failure of COVID-19 Vaccines?* J Korean Med Sci, 2021. **36**(18): p. e124.