Correlation between SARS-CoV-2 Spike Protein Sequence Variants and Structure Variance Using Phylogenetic Tree and Protein Simulation

Carnegie Mellon University

02-251 Great Ideas in Computational Biology, Spring 2021

LCM Team Members: Katrina Liu, Kefan Cao, Christina Ma

November 2, 2022

1 Abstract

The spike protein of SARS-CoV-2 plays a key role when the virus enters the host, and small mutations at spike protein may lead to very prevalent SARS-CoV-2 strands. Building a phylogenetic tree on the sequences of SARS-CoV-2 spike protein can serve as a classification model on SARS-CoV-2 spike protein variants. Based on the classification results, we used protein simulation to construct the structures of the spike protein variants and obtained the similarity and difference of protein structures. By analyzing the phylogenetic tree and comparing it to the results from protein simulation similarities, we found some moderate correlations between the spike protein sequence variance and its structure. However, there may be some flaws in our data and measurements, so we may not reach a conclusive statement.

2 Introduction

2.1 Background

SARS-CoV-2 is a single-stranded RNA virus with a large number of glycosylated S proteins covering its surface. When the S protein binds to the receptor, the membrane proteins promote virus entry into the cell by activating the S protein. Therefore, S protein is very critical to sustaining the viral life cycle and it is a very good target for drug therapies and vaccination development[6]. For example, researchers have shown that drugs targeting S protein to be experimentally effective, which includes ACE2-based peptide, 3CLpro inhibitor (3CLpro-1), and a novel vinyl sulfone protease inhibitor[12].

Albeit the relatively few mutations, spike proteins have in general, some of the most prevalent SARS-CoV-2 variants in this global pandemic are caused by mutations in spike proteins, which lead to functional and structural changes of the virus and make it more virulent. For example, the amino acid change in the D614G variant was caused by an A-to-G single-nucleotide mutation at position 23,403 in the reference genome, and structural mapping of amino acid changes and spike proteins' variations reveals that this mutation may have eliminated a side-chain hydrogen bond and thus increased the main-chain flexibility[8]. Therefore, it would be interesting to examine the variations of spike protein sequences and their structural-functional differences.

2.2 Research Question

Our project plans to address the research question: Is there a correlation between SARS-CoV-2 spike protein sequence variants and their structural-functional differences? The differences of SARS-CoV-2 spike protein sequence variants can be approximated by distance scoring methods based on the phylogenetic tree, and the difference of structural-functional differences of spike proteins can be derived from protein simulations.

2.3 Approach

Our project takes the approach of randomly selecting 200 amino acid sequences for each month from all amino acid sequence data retrieved from the GenBank database[3] via

the National Center for Biotechnology Information (NCBI) datasets project[1]. Then, we select three sequences with the highest branch lengths in major branches for each month based on the visualization of the phylogenetic tree constructed for each group of sequences we selected for each month. Finally, we will choose to use the SWISS model[18] to build homologous models based on their amino acid sequences. We conduct comparisons between the structures and model 6VYB [17] and comparisons between models built from sequences in each month. From there, we retrieved the QH, RMSD, and PI values by using the tool Visual Molecular Dynamics (VMD)[7] and plugins MultiSeq [4] and Stamp structural alignment [14].

2.4 Contributions

Researchers have spent huge efforts in discovering antiviral drugs and developing vaccines targeting specific regions of the SARS-CoV-2 spike protein for the treatment and prevention of SARS-CoV-2 infections. For example, Chi et al. [2] have discovered a neutralized therapeutic antibody targeting a region of the receptor-binding domain in the spike protein. Thus the variation in the region might have an impact on the effectiveness of the vaccines and drugs. However, few have tried tying protein sequence variants to their structural changes. We hope that establishing a correlation between the sequence variants of SARS-CoV-2 spike protein and changes in the protein structures can be useful in predicting how a newly discovered variant would affect the spike protein structure and thereby how it would affect the functions of spike protein in the process of viral infection.

3 Methodology

3.1 Parameter settings

In our project, we will introduce some parameters that are used to quantify the variables we want to compare. There are two variables we want to compare: 1. How different the amino acid sequences of the SARS-CoV-2 spike proteins are. 2. How different the S protein structures of the SARS-CoV-2 are.

To compare the amino acid sequences, a good indicator is the phylogenetic tree dis-

tances between different sequences. To quantify structural similarities, we choose to use the parameters provided by VMD: Q_H[13], Root-Mean-Square Deviation (RMSD), and Percent Identity(PI). Q_H is a structural homology measure that takes into account the, which ranges from 0 to 1, where $Q_H = 1$ indicates identical structures. RMSD indicates the difference of the structures with a higher value indicating a larger amount of difference. Percent Identity is used to demonstrate the similarity between structures.

3.2 Sequence data processing

We used the NCBI datasets project to retrieve all protein sequences related to the SARS-CoV-2 virus and filtered them based on their GenBank definitions. We only include the proteins with definitions "surface glycoprotein" and "spike glycoprotein", which both correspond to the S protein of the SARS-CoV-2 virus. Then, we categorized the sequence data based on the dates they were uploaded. Due to the size of the data, we used the python random package[16] to randomly selected 200 sequences for each natural month. Lastly, we converted the chosen sequences into FASTA[5] format and prepare them for multiple sequence alignment and tree building.

3.3 Phylogenetic tree building

Since the amount of genomic data is large for the recent months, we cannot construct a full phylogenetic tree for each month. Therefore, we randomly picked 200 sequences from each month for the phylogenetic tree construction. We used Multiple Sequence Comparison by Log-Expectation (MUSCLE)[11] for multiple sequence alignment and Simple Phylogeny[11] for phylogenetic tree construction.

We visualized the phylogenetic trees of each month with iTOL[10] and cherry-picked three sequences from each tree with the largest distances in different three branches from the root node for further analysis. Figure 1. is a sample phylogenetic tree constructed from the genomes we selected from April 2021. We have include the tree data and project related results in a GitHub repository (link: `https://github.com/katrina-liu/02251-lcm-project`).

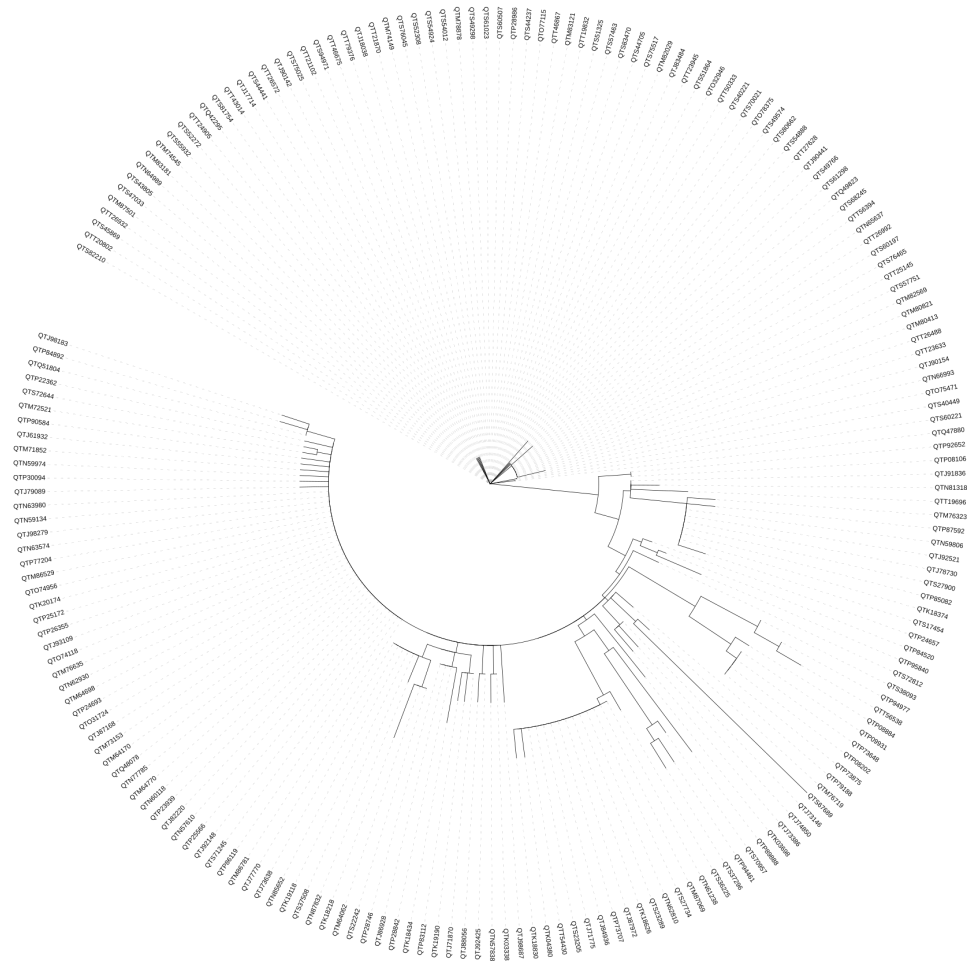

Figure 1: Sample phylogenetic tree of April 2021.

3.4 Homologous model building

We used the SWISS modeling for homogenous modeling of the picked sequences. Since the SWISS model gives more than one model, we selected the model with the highest Global Model Quality Estimation (GQME) and Qualitative Model Energy Analysis (QMEAN), which indicates good modeling of the protein structure. GQME values range from 0 to 1; close to 1 represents the reliability of the predicted 3D structure and QMEAN value < 4.0 shows the trustworthiness of structure[9]. Structural comparison is done using tool VMD With the model of structures for each sequence, we compared their distance on phylogenetic tree measured in the patristic distance on the tree with the difference in their protein structure, measured in Q_H, RMSD, and percent identity. We compared the sequences and structures with 6VYB, which is the root node for the phylogenetic tree,

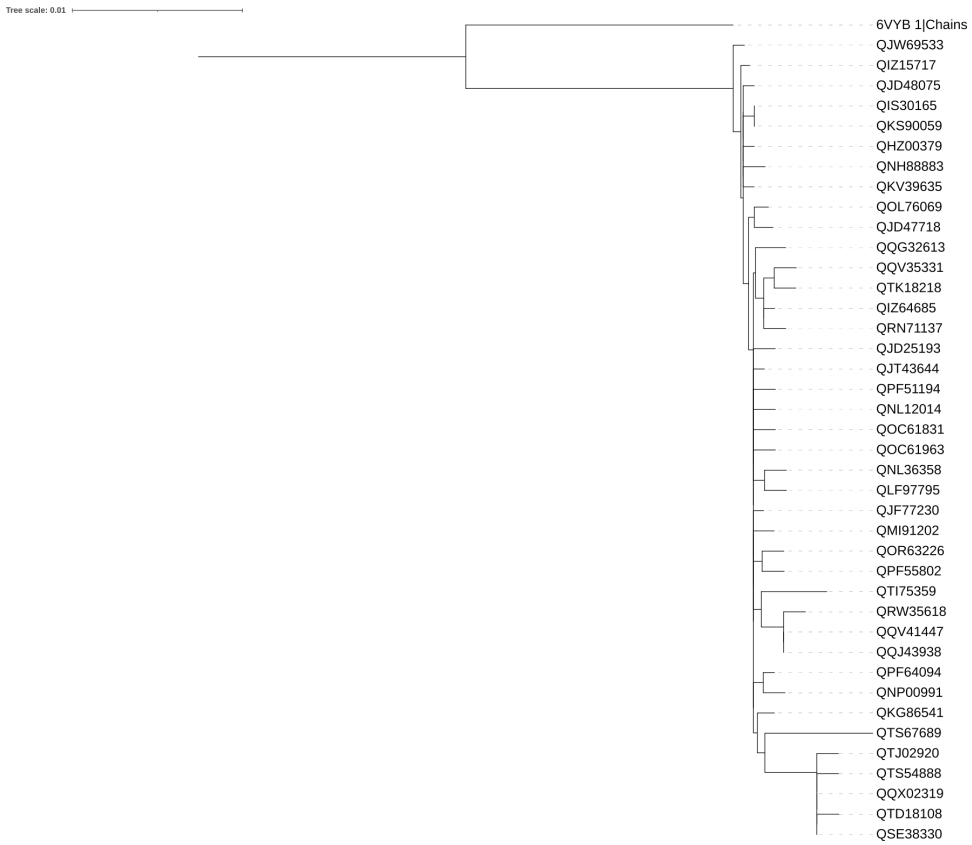

Figure 2: Phylogenetic tree of the selected sequences rerooted at 6VYB.

and we also made a pairwise comparison for each pair of genome sequences within the same month.

For comparisons with 6VYB, we obtained the distance from the phylogenetic tree by rerooting the tree at 6VYB and summing up the branch lengths from the sequence to 6VYB. Figure 2. is the tree we used to calculate this distance. For the structure analysis, we first loaded the two molecules into VMD, apply the MultiSeq plugin [4] and then performed stamp alignment[14]. The three scores: Q_H, Root-Mean-Square Deviation (RMSD), and Percent Identity(PI), can be obtained by selecting the corresponding chains we want to compare. However, since the structure of 6VYB is from electron-microscopy and our proteins are modeled by SWISS modeling, we calculated the three properties by selecting the corresponding chain with the lowest RMSD score after stamp alignment[14]. High RMSD scores (>100) indicate a problem in protein modeling, probably because SWISS used the incorrect template. Thus these proteins are excluded from the analysis.

For the pairwise comparison within each month, we calculated the distance for pairwise comparison using python DendroPy package[15], and we performed the same structure analysis.

4 Results

4.1 Experimental Setup

Given the distances and structural similarity data from both the within-month pairwise comparison of protein sequences and the comparisons between all selected sequences against the model 6VYB, we performed two linear regressions to analyze the correlation between protein sequence distance and structural similarity in these two conditions. Since the spike protein consists of three sub-units that have different alignment performance, we group the data into three chains and separately conduct the data analysis.

We also visualize the distances and structural similarity data of the selected sequences based on a chronological order to examine any abnormalities in the data we used for analysis, or if they conform with our expectation and are representative of the general trend that protein sequence variances increase throughout the time.

Table 1: Correlation Statistics

Spearman's ρ	**Distance by Q_H**			**Distance by RMSD**			**Distance by PI**		
Chain	***A***	***B***	***C***	***A***	***B***	***C***	***A***	***B***	***C***
within month	-0.594	-0.626	-0.298	0.633	0.562	0.395	-0.857	-0.731	-0.727
against 6VYB	0.270	0.367	-0.254	-0.043	0.068	0.079	0.255	0.016	-0.273

4.2 Results

According to the Spearman correlation coefficient, ρ in Table 1, the correlations between distances and structural similarities in the within-month condition is moderate in general, and the negative correlations between distance and percent identity even seem to be

strong. The directions of the correlations in the within-month condition indeed capture some trends that we predicted (Figure 3), as the less similar the protein sequences (increasing distance), the less similar their protein structures (increasing RMSD, decreasing RH and PI).

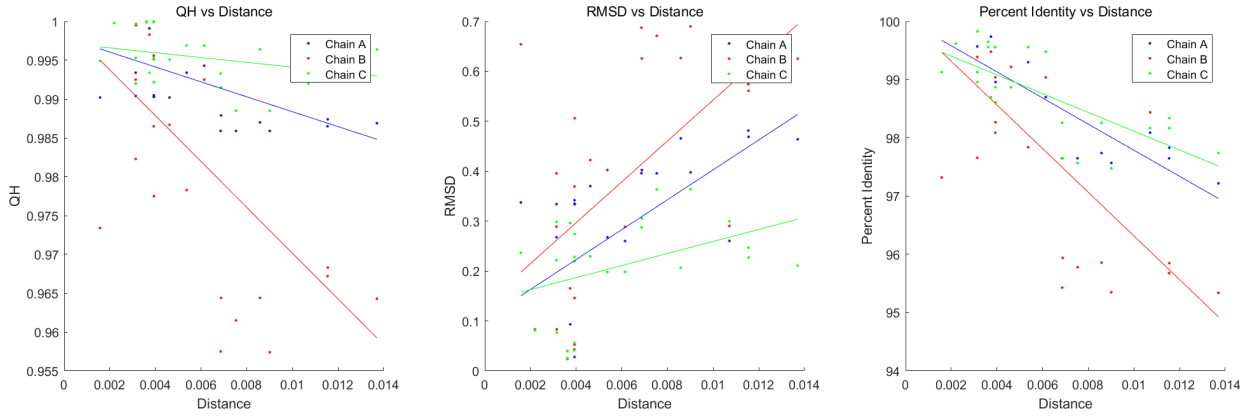

Figure 3: Distance by Structural similarity (within month)

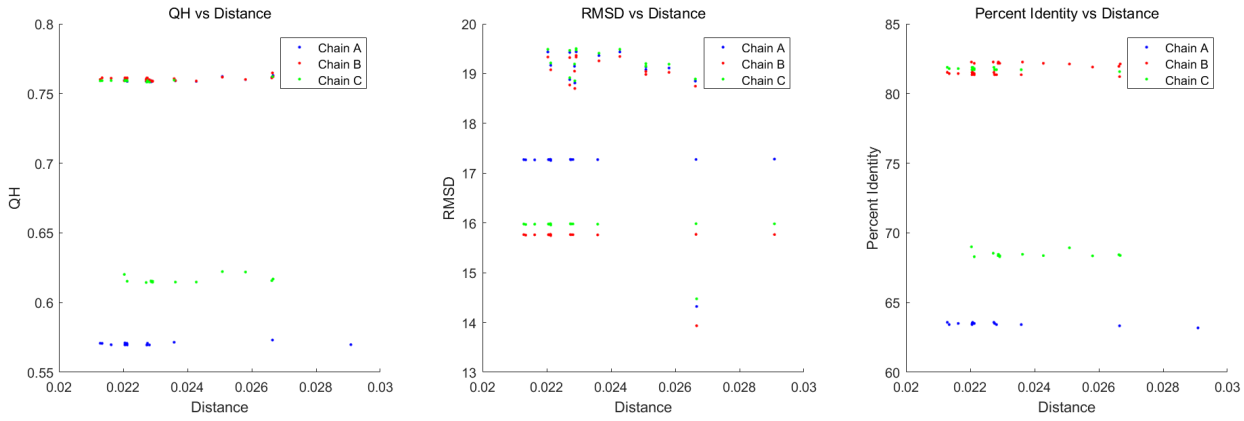

Figure 4: Distance by Structural similarity (against 6VYB)

However, the correlations in the against-6VYB condition are much weaker (the absolute values of the correlation coefficient ρ are mostly smaller than 0.3 as in Table 1). The directions of the correlations also seem rather random, so we cannot reach any confident conclusion regarding our hypothesis.

Similarly, in the scatterplots of distance vs. structural similarities when comparing against the 6VYB model, we cannot find any patterns of correlation (Figure 4). But we

do observe some clusters, which means it is possible that some variations originated from mutations of key position had a similar impact on protein structures, or it may reflect some potential systematic errors, as further elaborated in section 5.2.

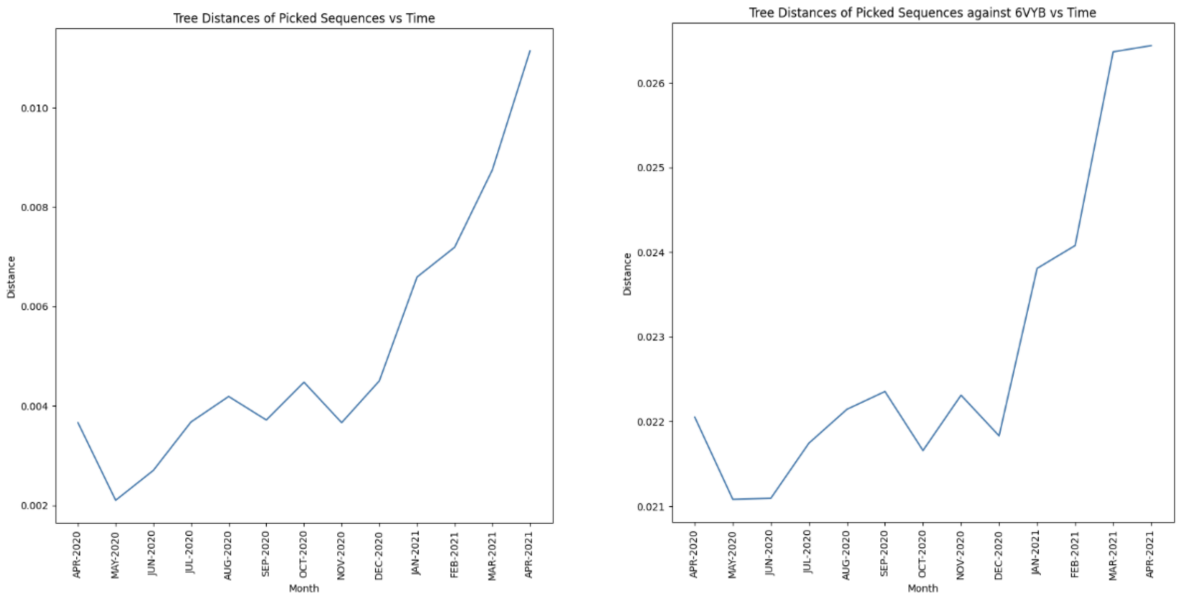

Figure 5: Distance by Time within month (left) & against 6VYB (right)

In the chronological analysis, we can observe that the distances increase with time in both comparing within the month and against 6VYB conditions (Figure 5), which may indicate that genetic diversity of the virus within months has increased, and the accumulation of mutation resulted in a larger divergence against the original sequence 6VYB throughout the time.

The visualizations of protein structural similarity in the chronological order reflect very similar patterns as the correlations in the two conditions (Figure 6, Figure 7), so it also captures the oddity of the correlations we have found in the against 6VYB condition (Figure 7). This oscillation pattern between the protein structural similarity and time can be more clearly observed in Figure 8, and more interpretations can be found in section 5.1.

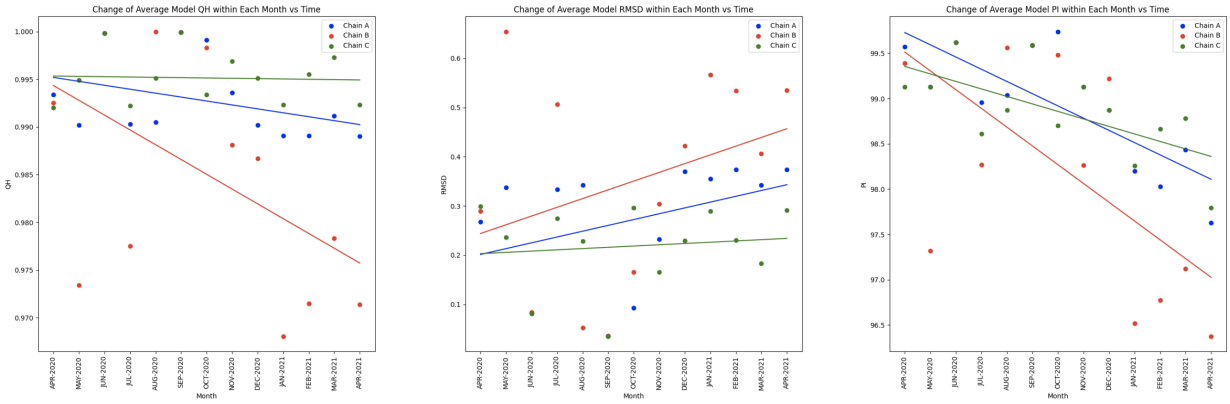

Figure 6: Structural similarity (QH, RMSD, PI) by Time (within month)

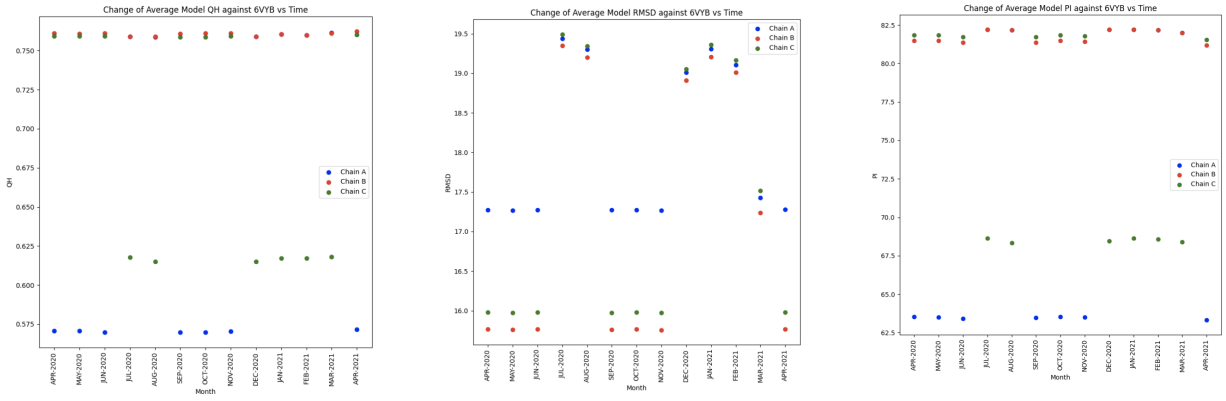

Figure 7: Structural similarity (QH, RMSD, PI) by Time (against 6VYB)

4.3 Experimental Evaluation

There may be some errors in our protocol and data used for analysis. For example, a major flaw is that the number of data points may be too small so that they are insufficient to support our conclusions. Given a huge amount of amino acid sequence data, we only randomly select 200 sequences per month to build the phylogenetic trees (out of about ten thousand total sequences). Out of the phylogenetic trees for each month, we cherry-picked about 3 sequences for further structural comparisons, so both the random and manual selection processes may have left out important data. Also, we fixed the model 6VYB as the reference protein sequence and we compare all selected sequences against this model, which may have a bias in itself. Therefore, the results we found based on our

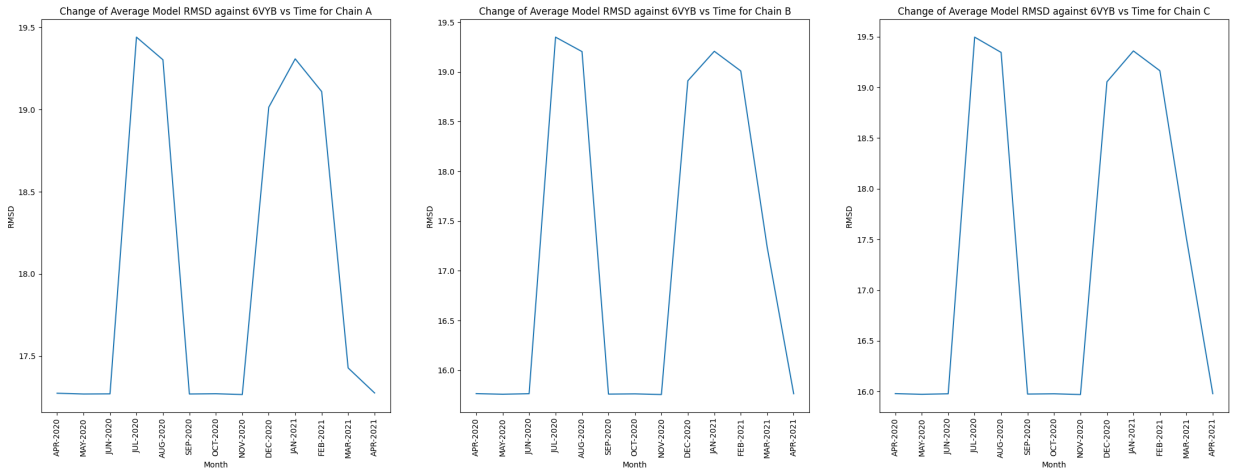

Figure 8: Average Model RMSD by Time (against 6VYB)

data may not be representative of the general trend.

Additionally, the use of the SWISS model has caused serious problems in Stamp structural alignment in VMD, as the Swiss model may have compared the sequences against a different template, and thus even very similar protein structures could have returned large RMSD scores which have to be eliminated from the analysis, and further contribute to the problem of insufficient data in this project.

5 Surprises

5.1 Oscillation of distances and structural similarity chronologically

We also analyzed the collected data based on the order of months they were selected from. We assume that the sequences and models will have a larger variance if they are discovered more recently. However, the data we collected does not exhibit a trend of increasing average tree distances over time or decreasing average structural similarities over time. Our data display an oscillation. Figure 8 demonstrates the monthly average RMSD value over time. The data increases from APR-2020 to JUL-2020. Then, there is a sudden drop for the distances and structural similarities from month SEP-2020 to NOV-

2020, which is followed by another cycle of oscillation. We do not have an explanation for this behavior other than a possible flaw in experimental design. Possible experimental design flaws that might cause this behavior include:

5.2 Clustering of values of distances and structural similarity against model 6VYB

As we observed in Figure 4 and Figure 7, instead of demonstrating a negative correlation between distances and structural similarities, the relationship between distances and structural similarities displays clustering of values of structural similarities. These clusters are formed by three major different settings of values of the structural similarity of three chains. From Figure 4, the clusters span the range of distances. That is, the clustering behavior does not involve tree distances. We proposed a plausible explanation for this other than possible errors during experiment conduction: Each cluster of data represents a type of mutations. This might be able to explain why the structures have about the same similarity scores despite their differences in distances. That is, their structures are similar to model 6VYB in the same way. However, we do not have the time to prove this assumption.

6 Conclusions

In conclusion, our project attempts to find the correlation between the phylogenetic tree distances of amino acid sequences of the SARS-CoV-2 virus and structural similarities of their protein structures. We adopted two different approaches: one is to compare selected sequences with model 6VYB; the other approach is the pairwise comparison of selected sequences within each month.

For comparison against 6VYB, The data does not display a significant positive correlation between structural similarity and tree distance, but there is a clustering of similarity regardless of tree distances. For pairwise comparison within each month, The structural similarity vs. tree distance for each month exhibits negative correlations for Q_H and PI and a positive correlation for RMSD. Although the correlations are insignificant, this is

still consistent with our hypothesis that the structural differences. of proteins is correlated with their sequence variations.

7 Future Work

Various improvements and extensions can be done regarding this project. Possible improvements to minimizes bias produced by an imperfect sampling method include sampling more data points from each month, selecting sequences using better criteria, generating the protein structure of 6VYB like other amino acids sequences using the SWISS model.

Possible extensions include building a more computational and automated way to collect data and perform analysis as many of our work is done manually, comparing against closed state spike protein model, for example, model 6VXX[17], and shortening the time interval for selecting sequences to 10 days or 5 days to get more data. Lastly, the assumption of clustering of data points implies different types of mutations should be verified.

References

[1] National Center for Biotechnology Information Bethesda (MD): National Library of Medicine (US). National center for biotechnology information (NCBI). `https://www.ncbi.nlm.nih.gov/`. Accessed: 2021-05-07.

[2] Xiangyang Chi, Renhong Yan, Jun Zhang, Guanying Zhang, Yuanyuan Zhang, Meng Hao, Zhe Zhang, Pengfei Fan, Yunzhu Dong, Yilong Yang, Zhengshan Chen, Yingying Guo, Jinlong Zhang, Yaning Li, Xiaohong Song, Yi Chen, Lu Xia, Ling Fu, Lihua Hou, Junjie Xu, Changming Yu, Jianmin Li, Qiang Zhou, and Wei Chen. A neutralizing human antibody binds to the n-terminal domain of the spike protein of sars-cov-2. *Science*, 369(6504):650–655, 2020.

[3] Lipman DJ Ostell J Sayers EW. Clark K, Karsch-Mizrachi I. Genbank. *Nucleic acids research*, 44(D1):D67-D72, 2016.

[4] Dan Wright Elijah Roberts, John Eargle and Zaida Luthey-Schulten. MultiSeq: Unifying sequence and structure data for evolutionary analysis. *BMC Bioinformatics*, 22, 7:382, 2006.

[5] Tom Goldstein, Christoph Studer, and Richard Baraniuk. Fasta: A generalized implementation of forward-backward splitting, January 2015. http://arxiv.org/abs/1501.04979.

[6] Yuan Huang, Chan Yang, Xin-feng Xu, Wei Xu, and Shuwen Liu. Structural and functional properties of sars-cov-2 spike protein: potential antivirus drug development for covid-19. *Acta Pharmacologica Sinica*, 41:1–9, 08 2020.

[7] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[8] Bette Korber, Will M. Fischer, Sandrasegaram Gnanakaran, Hyejin Yoon, James Theiler, Werner Abfalterer, Nick Hengartner, Elena E. Giorgi, Tanmoy Bhattacharya, Brian Foley, Kathryn M. Hastie, Matthew D. Parker, David G. Partridge, Cariad M. Evans, Timothy M. Freeman, Thushan I. de Silva, Adrienne Angyal, Rebecca L.

Brown, Laura Carrilero, Luke R. Green, Danielle C. Groves, Katie J. Johnson, Alexander J. Keeley, Benjamin B. Lindsey, Paul J. Parsons, Mohammad Raza, Sarah Rowland-Jones, Nikki Smith, Rachel M. Tucker, Dennis Wang, Matthew D. Wyles, Charlene McDanal, Lautaro G. Perez, Haili Tang, Alex Moon-Walker, Sean P. Whelan, Celia C. LaBranche, Erica O. Saphire, and David C. Montefiori. Tracking changes in sars-cov-2 spike: Evidence that d614g increases infectivity of the covid-19 virus. *Cell*, 182(4):812–827.e19, 2020.

[9] Rohit Kumar, Sainitin Donakonda, Stephan A. Müller, Kai Bötzel, Günter U. Höglinger, and Thomas Koeglsperger. Fgf2 affects parkinson's disease-associated molecular networks through exosomal rab8b/rab31. *Frontiers in Genetics*, 11:1153, 2020.

[10] Ivica Letunic and Peer Bork. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, 04 2021. gkab301.

[11] Fábio Madeira, Young Mi Park, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, Adrian R N Tivey, Simon C Potter, Robert D Finn, and Rodrigo Lopez. The embl-ebi search and sequence analysis tools apis in 2019. *Nucleic acids research*, 47(W1):W636—W641, July 2019.

[12] Lalonde T. Xu S. Liu W. R. Morse, J. S. Learning from the past: Possible urgent prevention and treatment options for severe acute respiratory infections caused by 2019-ncov. *Chembiochem : a European journal of chemical biology*, 21(5), 730–738, 2020.

[13] Patrick O'Donoghue and Zaida Luthey-Schulten. On the evolution of structure in aminoacyl-trna synthetases. *Microbiology and molecular biology reviews : MMBR*, 67,4 (2003): 550-73, 2003.

[14] G. J. Barton R. B. Russell. Multiple protein sequence alignment from tertiary structure comparison. *PROTEINS: Struct. Funct. Genet.*, 14, 309-323, 1992.

[15] J. Sukumaran and Mark T. Holder. Dendropy: A python library for phylogenetic computing. *Bioinformatics*, 26: 1569-1571, 2010.

[16] Guido Van Rossum. *The Python Library Reference, release 3.8.2*. Python Software Foundation, 2020.

[17] Alexandra C. Walls, Young-Jun Park, M. Alejandra Tortorici, Abigail Wall, Andrew T. McGuire, and David Veesler. Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell*, 181(2):281–292.e6, 2020.

[18] Bertoni M. Bienert S. Studer G. Tauriello G. Gumienny R. Heer F.T. de Beer T.A.P. Rempfer C. Bordoli L. Lepore R. Schwede T. Waterhouse, A. Swiss-model: homology modelling of protein structures and complexes. *Nucleic acids research*, 46(D1):W296-W303, 2018.

8 Appendix: Individual Contributions

8.1 Contributions of Katrina Liu

1. Write python scripts to Process raw sequence data, which includes filtering, categorizing, and random selecting spike protein sequence data for monthly tree building. Write scripts to query picked sequences. Write scripts to get sequence distances from phylogenetic tree data. Write scripts to parse and plot data for chronological analysis.
2. Use MUSCLE and Simple Phylogeny to get raw phylogenetic tree files.
3. Use online SWISS model to obtain PDB files of the homologous models built from picked sequences.
4. Use VMD to obtain structural similarity data.

8.2 Contributions of Kefan Cao

1. Use Simple Phylogeny to get raw phylogenetic tree files.
2. Use online SWISS model to obtain PDB files for a third of selected sequences.
3. Use VMD to obtain structural similarity data for half of selected sequences.
4. Write scripts to plot the structural similarity against distance on phylogenetic tree and perform data analysis.

8.3 Contributions of Christina Ma

1. Write scripts to download amino acid sequences from datasets tool
2. Use iTOL to construct phylogenetic trees and cherry pick sequences per month
3. Use online SWISS model to obtain PDB files from the picked sequences.
4. Code structural similarity data from VMD. Write scripts to get sequence distances from phylogenetic tree data.