Journal of Data Science 19 (2), 314–333
April 2021
DOI: 10.6339/21-JDS1006
COVID-19 Special Issue

Statistical Challenges in the Analysis of Sequence and Structure Data for the COVID-19 Spike Protein

Shiyu He[1] and Samuel W.K. Wong[1,*]

[1]*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada*

Abstract

As the major target of many vaccines and neutralizing antibodies against SARS-CoV-2, the spike (S) protein is observed to mutate over time. In this paper, we present statistical approaches to tackle some challenges associated with the analysis of S-protein data. We build a Bayesian hierarchical model to study the temporal and spatial evolution of S-protein sequences, after grouping the sequences into representative clusters. We then apply sampling methods to investigate possible changes to the S-protein's 3-D structure as a result of commonly observed mutations. While the increasing spread of D614G variants has been noted in other research, our results also show that the co-occurring mutations of D614G together with S477N or A222V may spread even more rapidly, as quantified by our model estimates.

Keywords *Bayesian hierarchical models; compositional data analysis; conformational sampling; mutant clusters; SARS-CoV-2*

1 Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a strain of novel coronavirus that caused the COVID-19 outbreak in Wuhan, China in December 2019, has quickly spread across the world and has been characterized as a global pandemic (Zhou et al., 2020a; Wu et al., 2020). As of December 19, 2020, there have been over 74 million probable or confirmed cases of COVID-19, and the illness has been associated with 1.66 million deaths around the world (WHO, 2020a). The development of vaccines and antibody-based therapeutic agents has been initiated since the beginning of the pandemic and several have moved into phase III trials (Krammer, 2020; WHO, 2020b). Results concerning the long-term immunogenicity and efficacy of these vaccine candidates are a subject of continued research. Meanwhile, the virus has been found to mutate in human-to-human transmissions over time, and these changes can potentially alter the efficacy of these interventions. Therefore, it is also of vital importance to identify and study mutations with possible fitness advantages and increased infectiousness.

SARS-CoV-2 is a single-stranded RNA virus, and RNA viruses are known to have high mutation rates and genetic diversity compared to DNA viruses (Duffy, 2018; Lauring and Andino, 2010). Their ability to evolve underlies why they can adapt to novel hosts and develop resistance to either vaccine or infection-induced immunity. Often, the most consequential mutations in terms of viral functions and resistance to neutralizing antibodies are those that alter the surface proteins of the virus. For instance, the mutation A82V in the Ebola virus glycoprotein was confirmed to have enhanced infectivity and increased the severity of the EVD epidemic (Diehl et al., 2016). Further, co-occurring mutations of A143V and R148K in the influenza H7N9

*Corresponding author. Email: samuel.wong@uwaterloo.ca.

Received December 31, 2020; Accepted January 18, 2021

surface protein led to a 10-fold reduction in its sensitivity to neutralizing antibodies (Ning et al., 2019). As a result, mutations in the SARS-CoV-2 genome are being continuously monitored over time, and a major public repository for sequenced genomes is GISAID (https://gisaid.org). In addition to collecting viral genome data, GISAID also provides tools for visualizing the spread of various mutations, organized into phylogenetic clusters (also known as *clades*) over space and time.

Four structural proteins – spike (S), envelope (E), membrane (M), and nucleocapsid (N) – are the building blocks for the SARS-CoV-2 virus particle (Phan, 2020). Out of these four proteins, the S-protein plays the most critical role in attachment and entry into host cells, through its binding with the human ACE2 receptor (Wan et al., 2020). For this reason, the S-protein is the major target of many vaccines and neutralizing antibodies against SARS-CoV-2 (Amanat and Krammer, 2020). In the event of infection, these antibodies can disrupt the spike protein's ability to bind with the ACE2 receptor, thereby blocking its entry into host cells. The analyses in this paper focus on mutations in the amino acid sequence of the S-protein due to its particular importance.

Among all currently known S-protein sequence variants resulting from mutations in the underlying genome, D614G has been the most extensively studied due to its relatively early emergence and subsequent prevalence. The notation "D614G" means that the amino acid D (aspartic acid) in position 614 of the original (or *reference*) sequence has mutated to G (glycine), where the letters are used to denote the 20 different amino acid types. A rapid increase in D614G was observed in many regions after its initial appearance, which suggested fitness advantages and the hypothesis that variants with D614G are likely more infectious (Korber et al., 2020). This was later corroborated by experimental evidence that D614G, either by itself or in conjunction with other mutations, is significantly more infectious than the reference S-protein sequence (Li et al., 2020). Overall, the continued evolution of the virus has resulted in thousands of distinct S-protein sequence variants recorded in GISAID, although many of these only differ by a few mutated sequence positions. While clinical or laboratory experiments can test the infectivity of specific mutations, it is challenging to analyze large numbers of sequence variants.

Computational researchers have thus used clustering as a means to gain interpretable insight into the effect of S-protein mutations across different geographical regions. Temporal changes in the prevalence of S-protein mutations have also been studied in related research. For instance, Chen et al. (2020b) clustered mutations occurring in the receptor binding domain (RBD) of the S-protein and studied binding affinity changes for each cluster. Based on common amino acid mutations, Toyoshima et al. (2020) classified 28 countries into three clusters and studied correlations between fatality rate and S-protein D614G variants. In addition, the hypotheses of monotonic trends for D614G and various other mutations have been tested using isotonic regression by Korber et al. (2020) and in the COVID-19 pipelines of the Los Alamos National Laboratory (https://cov.lanl.gov). However, to the best of our knowledge, few authors have built comprehensive statistical models for the evolution of S-protein mutant clusters (i.e., groups of closely related sequence variants) over space and time. Such models can have an important practical value in providing forecasts and early warnings for countries where S-protein sequence variants with potential fitness advantages or higher infectiousness are actively being transmitted. To that end, this paper presents one such Bayesian hierarchical model for multinomial time series that can help tackle this problem.

The 3-D structure of a protein corresponding to its amino acid sequence is a crucial part of the puzzle for understanding how the protein functions; thus, of particular interest here

is the 3-D structure of the SARS-CoV-2 S-protein and its mutated variants. Often, sequence mutations associated with changes in viral infectivity can be attributed to changes in protein structure (Schaefer and Rost, 2012). The first 3-D structure of the SARS-CoV-2 S-protein was released in mid-February 2020 (Wrapp et al., 2020), and since then many other S-protein structures have been added to the publicly available Protein Data Bank (PDB) (Bernstein et al., 1977). However, laboratory experiments to determine protein structure are laborious and costly, and ultimately some prove to be intractable. For this reason, the structural impacts of common S-protein mutations are not yet well-documented in the PDB, and computational methods are needed to predict their impact. Different tools for 3-D protein structures have been used for this purpose thus far, including protein-protein binding affinity prediction (Chen et al., 2020b), comparative modeling with known PDB structures (Sedova et al., 2020), and Monte Carlo sampling of protein segments (Wong, 2020). In this paper we follow the illustrative analysis in Wong (2020), applying similar statistical sampling approaches to assess the potential local structural changes to the S-protein for common mutations in the current mutant clusters considered.

Overall then, our goal in this paper to illustrate statistical ideas for tackling the aforementioned challenges associated with the analysis of S-protein data, both their sequence and structure aspects. Specifically based on presently available data, we study temporal and spatial changes in the mutations of S-protein sequences and their structural impact, with the aim to better understand the ongoing evolution of the disease. Our contribution can be summarized in three parts. First, we develop a Bayesian hierarchical model to study the evolution of mutant clusters. Second, we apply sampling methods to analyze the local structural changes of the most frequently occurring protein sequence mutations in these clusters. Third, we discuss our findings and relate them to other recent work reported in the literature.

2 Data Description and Exploratory Analysis

2.1 Sequence Dataset

The S-protein sequence dataset for SARS-CoV-2 was obtained from GISAID on Oct 14th, 2020, with the number of sequences totaling 98,699 after incomplete sequences were removed. The full S-protein, based on the first discovered reference sequence, is 1273 amino acids long. Out of all complete sequences, 3,205 of them are unique, indicating that viral evolution has resulted in substantial genetic diversity. Our analysis of complete sequences shows that D614G is the most frequent mutation, appearing in 86.5% of recorded sequences, followed by S477N (6.3%), A222V (3.6%), L18F (1.9%), and L5F (0.99%), R21I (0.98%), and D936Y (0.74%). Many of these mutations are also mentioned in the recent literature where mutation analysis was considered (Korber et al., 2020; Chen et al., 2020a; Hodcroft et al., 2020). The sequences from GISAID are indexed by country and date of deposition, which allows us to conveniently group them for subsequent analysis. Sequences are separately deposited by local laboratories, therefore sequence counts vary widely by country and may not be well-correlated with actual case counts. Due to this concern, we instead focus on the relative prevalence, i.e., proportions of counts, throughout this paper.

To analyze the large numbers of sequence variants, we first implemented hierarchical clustering to group the unique sequences according to their similarities. The distance matrix for hierarchical clusters was based on the number of pairwise mismatched letters and the Ward-D linkage criterion, which creates groups such that variance is minimized within clusters (Ward,

Table 1: Relative frequencies of the most common mutations present in each cluster. The top 9 mutations in descending order and their corresponding relative frequencies are shown in the columns for each of clusters I–V. For example, the D614G mutation is present in 89% of the sequences belonging to cluster I.

Rank	Cluster I		Cluster II		Cluster III		Cluster IV		Cluster V	
	Mutation	Freq	Mutation	Freq	Mutation	Freq	Mutation	Freq	Mutation	Freq
1	D614G	89%	P863H	3%	L5F	98%	D614G	100%	D614G	99%
2	S477N	3%	A262T	2%	D614G	87%	S477N	100%	A222V	98%
3	L5F	2%	Y453F	2%	H655Y	15%	T632N	8%	L18F	54%
4	D936Y	2%	T572I	1%	A222V	13%	L822F	4%	A262S	13%
5	A222V	2%	V615I	1%	V3G	10%	S939F	4%	P272L	9%
6	R21I	2%	K77M	1%	D574Y	10%	W258L	4%	D1163Y	9%
7	L54F	1%	A845S	1%	S459Y	6%	E1144Q	4%	L5F	7%
8	P1263L	1%	L8V	1%	M1229I	4%	G566S	4%	G1167V	6%
9	Q677H	1%	H655Y	1%	T859I	4%	P330A	2%	L176F	4%

1963). For our illustrative analysis, we chose to use a total of five clusters, which aims to achieve a balance between separability of the different clusters and interpretability of the results. Table 1 shows the most common mutations in each cluster and their frequencies, where it can be seen that these clusters all have identifiable patterns. For example, D614G is dominant in cluster I and present in 89% of sequences within that cluster, while cluster II has the lowest frequency of mutations, indicating that it is composed of sequences with a high level of similarity to the original reference sequence. In clusters III, IV, and V, D614G frequently occurred together with L5F, S477N, A222V respectively, and these paired mutations are observed in almost all sequences within those clusters, evidencing a high probability for the co-occurrence of some common mutations. Evidence for these co-occurrences can also be seen in Table 2, which displays the top three unique sequences in each cluster (as ranked by frequency within that cluster) and their specific mutation positions.

We selected 9 countries from the GISAID database to study based on the larger numbers of sequences deposited, which are United States (US), Canada (CA), United Kingdom (UK), Netherlands (NL), France (FR), Spain (SP), China (CN), India (IN), and Australia (AU). The pie charts in Figure 1 show the composition of clusters for each country for sequences accumulated since the outbreak. Countries with distinctly different compositions are China and Australia, where China has the majority of its sequences from cluster II, and Australia has cluster IV as its major cluster. Cluster I is the largest cluster for the rest of the countries, followed by cluster II, while cluster V has a noticeable presence in Europe, especially the UK.

The composition of clusters also changes over time, and as examples we show the temporal trend of the daily cluster counts (upper panels) and proportions (lower panels) for the United States (US) and the United Kingdom (UK) in Figure 2. The graphs show a major difference between the US and UK composition trends over time: cluster I in the UK peaked around May to July and cluster V saw a surge since late August or early September, while cluster I remains dominant in the US. Similarly, most European countries have cluster V surging during this period, which suggests the most common mutations in cluster V, D614G and A222V, might be related to the rise of infections across Europe during the late summer and early autumn of 2020

Table 2: Top three unique sequences in each cluster, ranked by frequency. For each unique sequence, its specific mutation positions are shown in the right column. For example, 58,271 sequences belonging to cluster I had exactly the one mutation D614G, while 777 sequences in cluster I had exactly the two mutations R21I and D614G.

Cluster	Frequency	Mutation Positions
I	58,271	D614G
	777	R21I, D614G
	602	D614G, D936Y
II	11,617	Reference Sequence
	117	A829T
	43	L8V
III	486	L5F, D614G
	121	L5F
	43	L5F, A222V, D574Y, D614G, H655Y
IV	5,472	S477N, D614G
	85	S477N, D614G, T632N
	51	S477N, D614G, A930V
V	1,388	A222V, D614G
	1,369	L18F, A222V, D614G
	150	A222V, A262S, P272L, D614G

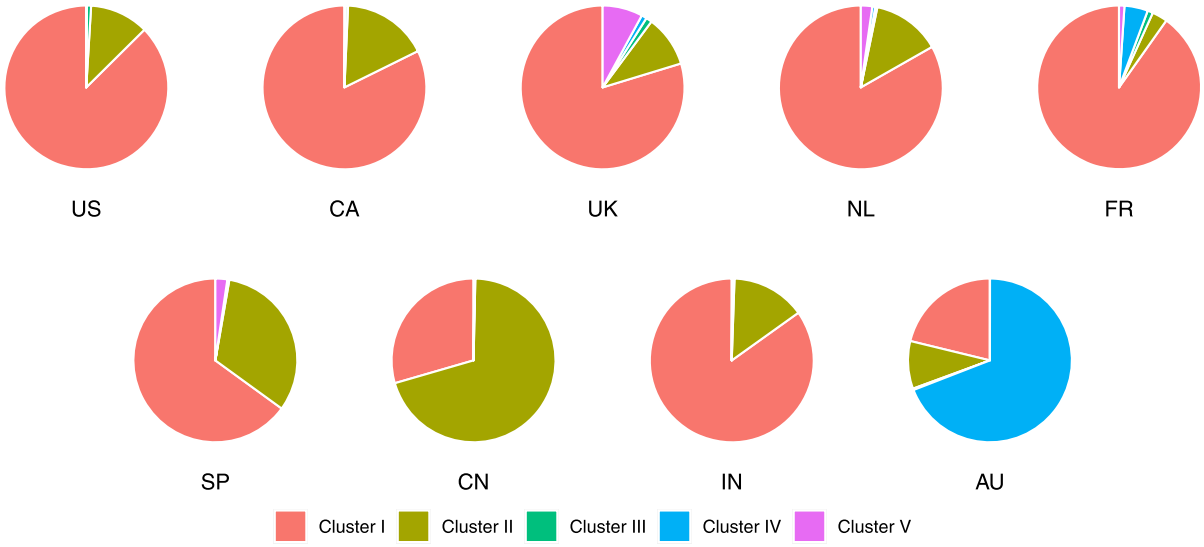

Figure 1: Pie charts of cluster proportions for 9 countries, based on all complete sequences accumulated in GISAID. For each country, the proportions of cluster I to cluster V are represented by the 5 different colors as indicated.

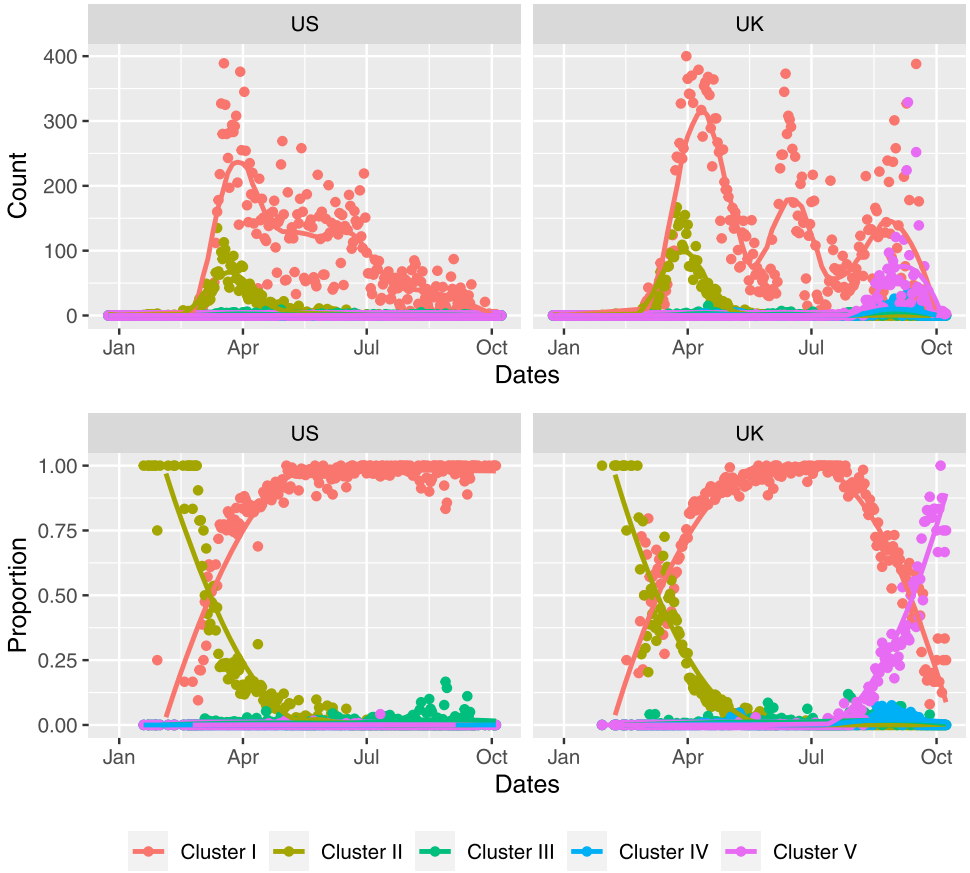

Figure 2: Temporal trend of cluster counts (upper panels) and cluster proportions (lower panels) in the US and UK. The points indicate the observed proportions and counts on each day, while the solid lines show the corresponding smoothed LOESS curves.

(Dong et al., 2020). In addition to this, the majority of countries have cluster II proportions decreasing over time in direct correspondence to the emergence of cluster I infections over time. It can be seen that the GISAID sequence counts in the upper panels of Figure 2 vary widely over time and do not necessarily correspond well with the actual case counts in the US and UK during this period.

2.2 Structure Dataset

The 3-D structure data for the S-protein were obtained from the Protein Data Bank (PDB) (Bernstein et al., 1977). The first laboratory-determined of a standalone 3-D structure of the SARS-CoV-2 S-protein was contributed in mid-February 2020 by a team of scientists at UT Austin using cryo-EM techniques (Wrapp et al., 2020), with PDB accession code 6VSB. Since then, many other groups around the world have contributed to the effort of studying different aspects of the S-protein using laboratory techniques. As of Oct 14th, 2020, there were 108 3-D structures publicly available in the PDB associated with the SARS-CoV-2 S-protein: 40 of these considered the S-protein in isolation, under different conformational states and sequence variants; 11 of these studied the structure of S-protein when bound together with ACE2; the remaining

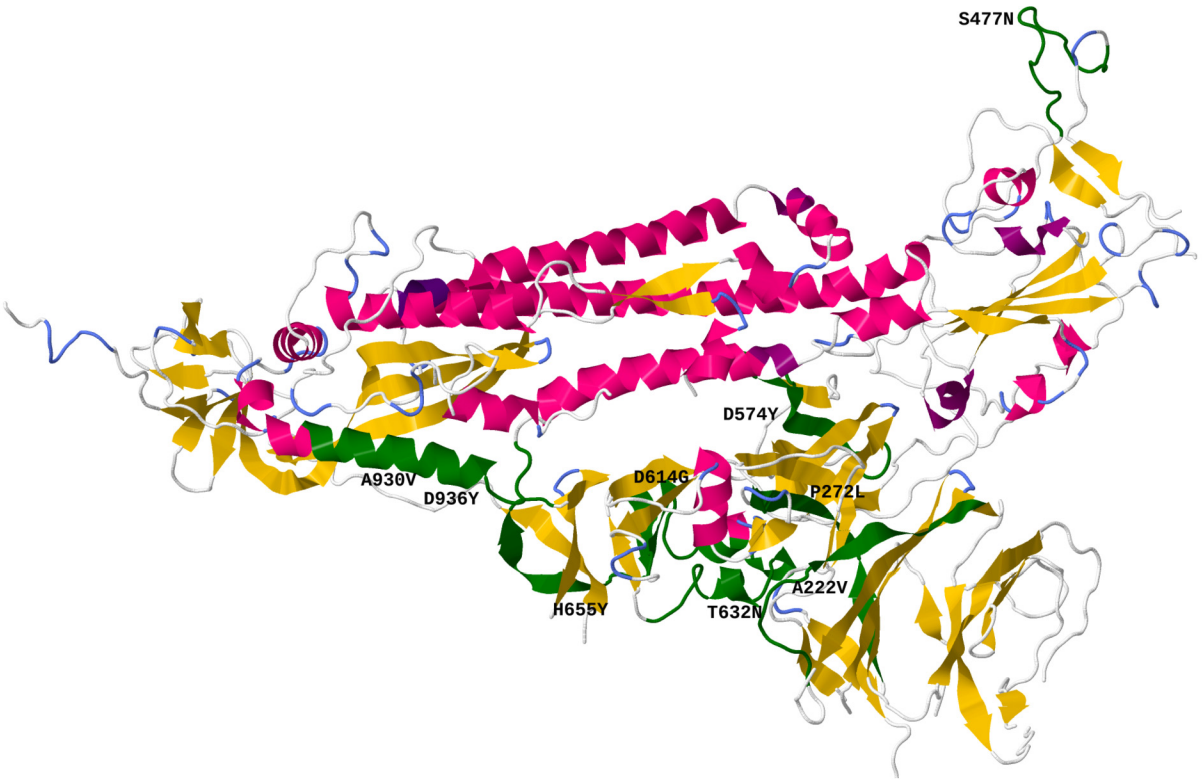

Figure 3: 3-D structure of the reference sequence S-protein (PDB accession code 6XM0). Protein segments containing common mutations are highlighted in green and labeled with the corresponding mutation in Table 2; segments with mutations that have incomplete 3-D structure data are not shown.

57 studied the structure of S-protein when interacting with different potential antibodies. Of the 68 structures containing the S-protein bound together with ACE2 or an antibody, 32 of these focused on a specific region of the S-protein, known as the receptor binding domain (RBD), primarily to decipher the binding behaviour of the S-protein to these molecules. Otherwise, for the majority of the structures (76 out of 108), experimenters attempted to determine the structure for the full S-protein.

Together, the PDB reflects the current state of knowledge for the S-protein structure, including the attempts made to assess possible neutralizing antibodies in the development of therapeutic interventions. Overall, 3-D structure determination has not kept pace with genome sequencing: among all the amino acid mutations listed in Table 1, only D614G has been studied in the laboratory. Thus, we cannot leverage the PDB to ascertain the potential changes in the 3-D structure of the S-protein as a result of those mutations. Even in the best consensus 3-D structure for the reference sequence to date (Zhou et al., 2020b), parts of the S-protein have not been successfully determined by the laboratory experiments, resulting in missing data. This consensus structure (PDB accession code 6XM0) is visualized in Figure 3, where the mutations identified in Table 2 are labelled; mutation locations where there is no 3-D structural information available are omitted from the figure.

3 Methods and Models

3.1 Bayesian Hierarchical Model for Mutant Clusters of Sequences

In this section we present a Bayesian hierarchical model for the observed sequence counts, by day and country, in each of the five clusters identified via our exploratory analysis. In motivating the model, we recall that the number of deposited GISAID sequences varies widely over time and by country. Thus, it is sensible to focus inference on the fraction of sequences belonging to each cluster (i.e., cluster proportions) in each country, as in compositional data analysis (Aitchison, 1982). Also, while LOESS identified some temporal trends for these cluster proportions in each country, it cannot be used for prediction or for comparing the overall growth rates of different clusters. Thus, our model contains the following key features. First, it provides estimates of the cluster proportions for each country on any given day. Second, it assumes a growth rate parameter for each cluster that is common across all countries, through which differences in prevalence among the S-protein mutant clusters can be quantified. Third, the temporal evolution of the cluster proportions is allowed to be dependent across countries via correlated errors.

Denote the observed sequence counts by the vectors $\boldsymbol{y}_{it} = (y_{it1}, \ldots, y_{it5})$, where y_{itc} is the number of sequences observed in GISAID for country $i = 1, \ldots, K$ on day $t \geqslant 1$ belonging to cluster $c = 1, \ldots 5$ (corresponding to clusters I–V in Section 2.1). We then let

$$\boldsymbol{y}_{it} \sim \mathit{Multinomial}\left(n_{it}, \frac{1}{14}\sum_{j=t-13}^{t} \boldsymbol{p}_{ij}\right) \text{ for } t \geqslant 14$$

where $n_{it} = \sum_{c=1}^{5} y_{itc}$ is the total number of GISAID sequences recorded for country i on day t, and $\boldsymbol{p}_{it} = (p_{it1}, \ldots, p_{it5})$ is the vector of probabilities representing the true underlying cluster proportions for country i on day t such that $\sum_{c=1}^{5} p_{itc} = 1$. Thus the model assumes that the observed sequences represent a random sample from the population of infected individuals in a country, with a reporting delay uniformly at random over the commonly assumed 14 day incubation period for the virus (Lauer et al., 2020). The sampling fraction (i.e., number of reported GISAID sequences out of the number of infected individuals in the country) may vary over time, and this setup permits inference on $\boldsymbol{p}_{it}$ in a manner that accommodates that sampling variability. Prior to Jan 20th, 2020 (i.e., when $t < 14$ in the model), we simply take the multinomial probability to be the average of the days modeled thus far, namely $\frac{1}{t}\sum_{j=1}^{t} \boldsymbol{p}_{ij}$.

We apply a log-ratio transformation on the cluster proportions, as commonly used in compositional data analysis (Aitchison, 1999). Treating cluster I as the baseline, we define

$$\tilde{\boldsymbol{p}}_{it} = \left[\log\left(\frac{p_{it2}}{p_{it1}}\right), \log\left(\frac{p_{it3}}{p_{it1}}\right), \log\left(\frac{p_{it4}}{p_{it1}}\right), \log\left(\frac{p_{it5}}{p_{it1}}\right)\right]$$

and model these transformed $\tilde{\boldsymbol{p}}_{it}$ values according to

$$\tilde{\boldsymbol{p}}_{it} = \tilde{\boldsymbol{p}}_{i,t-1} + \boldsymbol{\alpha} + \boldsymbol{\epsilon}_{it} \text{ for } t > 1 \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_2, \alpha_3, \alpha_4, \alpha_5)$ is a vector of growth rate parameters for clusters II to V, and $\boldsymbol{\epsilon}_{it} = (\epsilon_{it2}, \epsilon_{it3}, \epsilon_{it4}, \epsilon_{it5})$ represents a random noise vector with mean zero that governs daily fluctuations. Note that when $\boldsymbol{\epsilon}_{it} = \mathbf{0}$, equation (1) implies

$$p_{itc} = \frac{p_{i,t-1,c}\exp(\alpha_c)}{\sum_{c=1}^{5} p_{i,t-1,c}\exp(\alpha_c)} \propto p_{i,t-1,c}\exp(\alpha_c)$$

for $c = 1, \ldots, 5$ with α_1 defined to be 0, so that $\exp(\boldsymbol{\alpha})$ can be interpreted as the multiplicative daily growth rates of clusters II to V relative to cluster I, with the proportions normalized to sum to 1, similar to a model recently used to study the prevalence of different flu strains (Huddleston et al., 2020). The $\boldsymbol{\alpha}$ are assumed to be the same for all countries, to reflect the intrinsic fitness of each mutant cluster. Define $\boldsymbol{\epsilon}_{1:K,t,c} = (\epsilon_{1tc}, \ldots, \epsilon_{Ktc})$, namely the random noise vector on day t across all K countries for cluster c, where $c \in \{2, 3, 4, 5\}$. Then we let $\boldsymbol{\epsilon}_{1:K,t,c} \sim N_K(\mathbf{0}, \Sigma)$ independently for each day and cluster, where N_K denotes a K-variate Normal distribution and Σ is a covariance matrix. This formulation allows for spatial dependence in the sense that the noise term can be correlated among countries; intuitively, if a certain cluster experiences faster than expected growth in one country for a period of time, nearby or geographically linked countries may experience similar changes. In practice, we could set Σ to be of block diagonal form, for example, with each block as countries within the same continent, assuming correlations between different continents are likely negligible. Overall, our model setup follows the general structure of Gaussian dynamic models for multinomial time series described in Cargnoni et al. (1997).

We complete the model specification with the choice of priors. During the early outbreak period from Dec 24th, 2019 to Jan 6th, 2020, a total of 40 sequences worldwide were deposited in GISAID, with two that we would now classify to be in cluster I and 38 in cluster II. Thus we set a Dirichlet prior for the cluster proportions in the model for the starting date of the model (Jan 7th, 2020), $\boldsymbol{p}_{i1} \sim Dirichlet(3, 39, 1, 1, 1)$ independently for all $i = 1, \ldots K$ countries, obtained by adding the cluster I and II observations as pseudocounts to a uniform Dirichlet distribution. Weakly informative Cauchy priors with scale 0.5 are assigned to each unique diagonal (variance) element of Σ. A weakly informative LKJ correlation distribution with shape parameter 2 is assigned as the prior for the correlation matrices corresponding to the blocks of Σ. Finally, uniform priors are assigned for $\boldsymbol{\alpha}$.

To obtain the samples for the posterior distribution of the parameters, Markov chain Monte Carlo sampling for the model was carried out via Stan (Carpenter et al., 2017). Four parallel chains were run, with 5000 iterations each and the first half discarded as burn-in.

3.2 Sampling Methods for Local Protein Structure Analysis

Certain mutant clusters may have a higher prevalence than others, as identified via the estimates for $\boldsymbol{\alpha}$. A natural follow-up question is to ask whether these differences might be related to changes in the 3-D structure of the S-protein as a result of the common mutations shown in Table 2. To compare two 3-D protein structures, it is standard practice to compute the root-mean-square deviation (RMSD) between its corresponding backbone atoms; the four backbone atoms (N, C_α, C, O) are common to all amino acids, so this RMSD calculation can be applied even when amino acid mutations are present and provides a simple metric for assessing structural changes. However, as described in Section 2.2, currently the PDB lacks laboratory-determined structures for all but the D614G mutation, and thus modeling approaches are needed.

A protein structure is represented by the arrangement of its atoms in 3-D space, which is known as a *conformation*. Letting x denote a conformation and H a given scalar energy (or potential) function, a statistics-based approach to the problem is to draw samples from the Boltzmann distribution

$$\pi(x) \propto \exp\{-H(x)/T\}, \tag{2}$$

where T is the effective temperature. According to the energy landscape theory (Onuchic et al.,

1997), a protein structure tends to be most stable around the lowest energy conformation. While nature's 'true' energy function is not known, various energy approximations H have been developed to mimic this property for use in computational protein structure prediction (Zhang et al., 2007b). Thus in this context, the goal is to draw samples from equation (2) corresponding to the amino acid sequence before and after mutation, to assess possible structure differences among the low-energy conformations sampled.

We focus here on local structural impacts, that is, possible changes to the protein structure in the segment of amino acids near the mutation position. To do so, we treat the PDB structure in Zhou et al. (2020b) (with accession code 6XM0, and visualized in Figure 3) as the reference consensus 3-D structure for the S-protein corresponding to the reference sequence. Then we may sample conformations for specific segments of amino acids while holding the rest of the structure fixed at the coordinates in this reference 3-D structure. Segment lengths of up to approximately 15 amino acids have been recognized to be a rough upper bound where current sampling methods can perform adequately (Webb and Sali, 2017). For an individual mutation occurring at position j, we thus sample conformations for the length 15 segment of amino acids from positions $j-7$ to $j+7$. These length 15 segments for each individual mutation listed in Table 2 are highlighted in green in Figure 3.

Since proteins are composed of a linear sequence of amino acids, a sequential sampling approach can effectively exploit that property, by incrementally adding one amino acid at a time to construct approximate samples from equation (2). The idea of devising sequential sampling algorithms as a way to stochastically search for realistic low-energy conformations was originally proposed in Zhang et al. (2007a) and tested on lattice representations of proteins. Subsequently, extensions of the method applicable to real protein structures have been developed, including distance-guided chain growth (DiSGro, Tang et al., 2014) and sequential Monte Carlo (SMC, Wong et al., 2018). The implementations of these two algorithms also use slightly different approximations for the energy function H, and together can provide a more complete picture of the energy landscape. Thus we apply these algorithms to sample conformations for the segments shown in Figure 3, on both the reference sequence and the mutated sequence. Then following Wong (2020), we may compute the probability distribution of RMSDs between pairs of sampled conformations, as a way to assess potential differences in the low-energy conformational space as a result of the mutation. Specifically for the D614G mutation, we can use its known structure in the PDB to validate the results of these sampling methods.

4 Results

We present our results from fitting the proposed Bayesian hierarchical model in Section 4.1 and the results of sampling 3-D protein conformations in Section 4.2.

4.1 Estimates of Growth for the Mutant Clusters of Sequences

The posterior distribution of the parameters in the Bayesian hierarchical model (Table 3) shows that the daily growth rate parameters of clusters II to V relative to cluster I, estimated via their posterior means, are -0.05 $(-0.07, -0.03)$, 0.00 $(-0.01, 0.02)$, 0.02 $(-0.01, 0.04)$, 0.03 $(0.00, 0.05)$, with 95% credible intervals in brackets. These estimates show that during the study period, the sequences from clusters IV and V tend to have higher growth relative to clusters I, II, and III, of which mutant cluster II clearly has the weakest growth, as seen via its posterior interval that does not overlap the others. Referring to the mutation positions in Table 2, this indicates that

Table 3: Posterior means and 95% credible intervals (represented by the 2.5% and 97.5% percentiles) for the parameters in the Bayesian hierarchical model. $\boldsymbol{\alpha}$ represents the daily growth rate parameters for clusters II to V relative to cluster I. σ_C is the random noise standard deviation for countries within continent C, and $\boldsymbol{p}_{\cdot,1}$ represents the initial proportions on Jan 7th, 2020 for all countries. Σ_C for each continent together forms the diagonal block matrix Σ that represents the covariance matrix for the random noise vector.

Parameter	Mean	2.5%	97.5%
α_2	−0.0495	−0.0658	−0.0344
α_3	0.0022	−0.0142	0.0178
α_4	0.0172	−0.0073	0.0398
α_5	0.0267	0.0047	0.0478
σ_{NA}	0.2432	0.1664	0.3820
σ_{EU}	0.3802	0.2915	0.4771
σ_{AS}	0.3578	0.1885	0.6661
σ_{AU}	0.2768	0.1989	0.3926
$p_{\cdot,1,1}$	0.0999	0.0524	0.1660
$p_{\cdot,1,2}$	0.8993	0.8327	0.9473
$p_{\cdot,1,3}$	0.0008	0.0001	0.0030
$p_{\cdot,1,4}$	0.0000	0.0000	0.0001
$p_{\cdot,1,5}$	0.0000	0.0000	0.0001
$\Sigma_{\text{NA1,1}}$	0.0618	0.0277	0.1459
$\Sigma_{\text{NA1,2}}$	−0.0059	−0.0475	0.0344
$\Sigma_{\text{NA2,1}}$	−0.0059	−0.0475	0.0344
$\Sigma_{\text{NA2,2}}$	0.0618	0.0277	0.1459
$\Sigma_{\text{EU1,1}}$	0.1468	0.0850	0.2276
$\Sigma_{\text{EU1,2}}$	0.0486	−0.0353	0.1429
$\Sigma_{\text{EU1,3}}$	0.0431	−0.0309	0.1313
$\Sigma_{\text{EU1,4}}$	−0.0133	−0.1055	0.0841
$\Sigma_{\text{EU2,1}}$	0.0486	−0.0353	0.1429
$\Sigma_{\text{EU2,2}}$	0.1468	0.0850	0.2276
$\Sigma_{\text{EU2,3}}$	0.0178	−0.0610	0.1048
$\Sigma_{\text{EU2,4}}$	0.0320	−0.0926	0.1294
$\Sigma_{\text{EU3,1}}$	0.0431	−0.0309	0.1313
$\Sigma_{\text{EU3,2}}$	0.0178	−0.0610	0.1048
$\Sigma_{\text{EU3,3}}$	0.1468	0.0850	0.2276
$\Sigma_{\text{EU3,4}}$	0.0153	−0.0802	0.1114
$\Sigma_{\text{EU4,1}}$	−0.0133	−0.1055	0.0841
$\Sigma_{\text{EU4,2}}$	0.0320	−0.0926	0.1294
$\Sigma_{\text{EU4,3}}$	0.0153	−0.0802	0.1114
$\Sigma_{\text{EU4,4}}$	0.1468	0.0850	0.2276
$\Sigma_{\text{AS1,1}}$	0.1422	0.0355	0.4436
$\Sigma_{\text{AS1,2}}$	0.0051	−0.1227	0.1803
$\Sigma_{\text{AS2,1}}$	0.0051	−0.1227	0.1803
$\Sigma_{\text{AS2,2}}$	0.1422	0.0355	0.4436

sequences with amino acid D in position 614 (as in the reference sequence) will tend to decrease in prevalence over time in the presence of the other mutant clusters. In addition, cluster I, which is mostly characterized by the lone D614G mutation, may have a growth disadvantage if clusters IV or V are also spreading in the country. The clusters with the strongest growth (IV and V) are primarily composed of variants with the co-occurrence of D614G with S477N or A222V.

The covariance matrix $\boldsymbol{\Sigma}$ for the random noise vector is set up in block diagonal form: each block represents countries within the same continent and parameterized as $\boldsymbol{\Sigma}_C = diag(\sigma_C) \times \boldsymbol{\Omega}_C \times diag(\sigma_C)$, where σ_C is the standard deviation of the daily noise term for countries within continent C, and $\boldsymbol{\Omega}_C$ is the corresponding correlation matrix. Specifically, we defined four continents: North America (NA) as (US, CA); Europe (EU) as (UK, NL, FR, SP); Asia (AS) as (CN, IN); and Australia (AU) as its own continent. The posterior means for σ_{EU} and σ_{AS} are relatively larger than σ_{NA} and σ_{AU}, indicating that overall cluster growth trends are somewhat more predictable in North America and Australia than Europe or Asia. On the other hand, spatial dependence in the noise terms in general is low for countries studied, with no estimated correlations exceeding 0.4. That said, based on the posterior means, European countries still have relatively larger spatial correlations in their daily fluctuations, e.g., UK and Netherlands have a correlation of 0.33, UK and France have a correlation of 0.28, and Netherlands and Spain have a correlation of 0.22. The estimates show these European countries may experience more similar day-to-day changes, while countries in North America and Asia do not, as the correlation appears to be negligible between US and Canada (-0.087) and between China and India (-0.045), both of which are close to 0. The relatively low correlations between countries suggest that zero noise correlation between continents (i.e., block diagonal $\boldsymbol{\Sigma}$) is a reasonable simplifying assumption.

Figure 4 shows the posterior means and 95% credible intervals for the inferred cluster proportions in the different countries on each day t from Jan 7th, 2020 to Oct 14th, 2020. The posterior means of $\boldsymbol{p}_{\cdot,1}$ indicate that on Jan 7th, 2020, cluster I accounts for around 10% and cluster II accounts for around 90% of sequences. Nonetheless these initial proportions are quite uncertain due to the limited number of early cases, as seen in the wide credible intervals on the plot. The credible intervals narrow as we reach periods where a larger number of sequences are deposited. Overall, we see growth for cluster I accelerates during January to June but appears to rapidly fall off in July for many countries, while maintaining a substantive presence in the US, India and Australia. Cluster II clearly diminishes over time worldwide, and cluster III is fairly small but stable for all countries. Cluster IV estimates show small fluctuations for most countries except Australia, where it expands to over 90% from July to August and may remain as the dominant cluster. Cluster V is estimated to first appear in August and thereafter shows a rapid growth in all European countries.

Our Bayesian hierarchical model also allows prediction of changes in cluster proportions for countries with missing or very sparse GISAID sequence data. Canada, Spain and China only have deposited sequences up to August, while France only has sequence data deposited until mid-September; our model is nonetheless able to provide the point and interval estimates for their cluster proportions over the entire period. As suggested in Figure 4, both Canada and China are projected to have cluster I gradually decrease together with a possible rise in cluster V; the expanding credible intervals reflect the increasing uncertainty associated with the increasing number of days with missing data. Meanwhile, the main clusters present in France by mid-October might be IV or V (or their combination), while Spain is projected to be dominated by cluster V much like the rest of Europe.

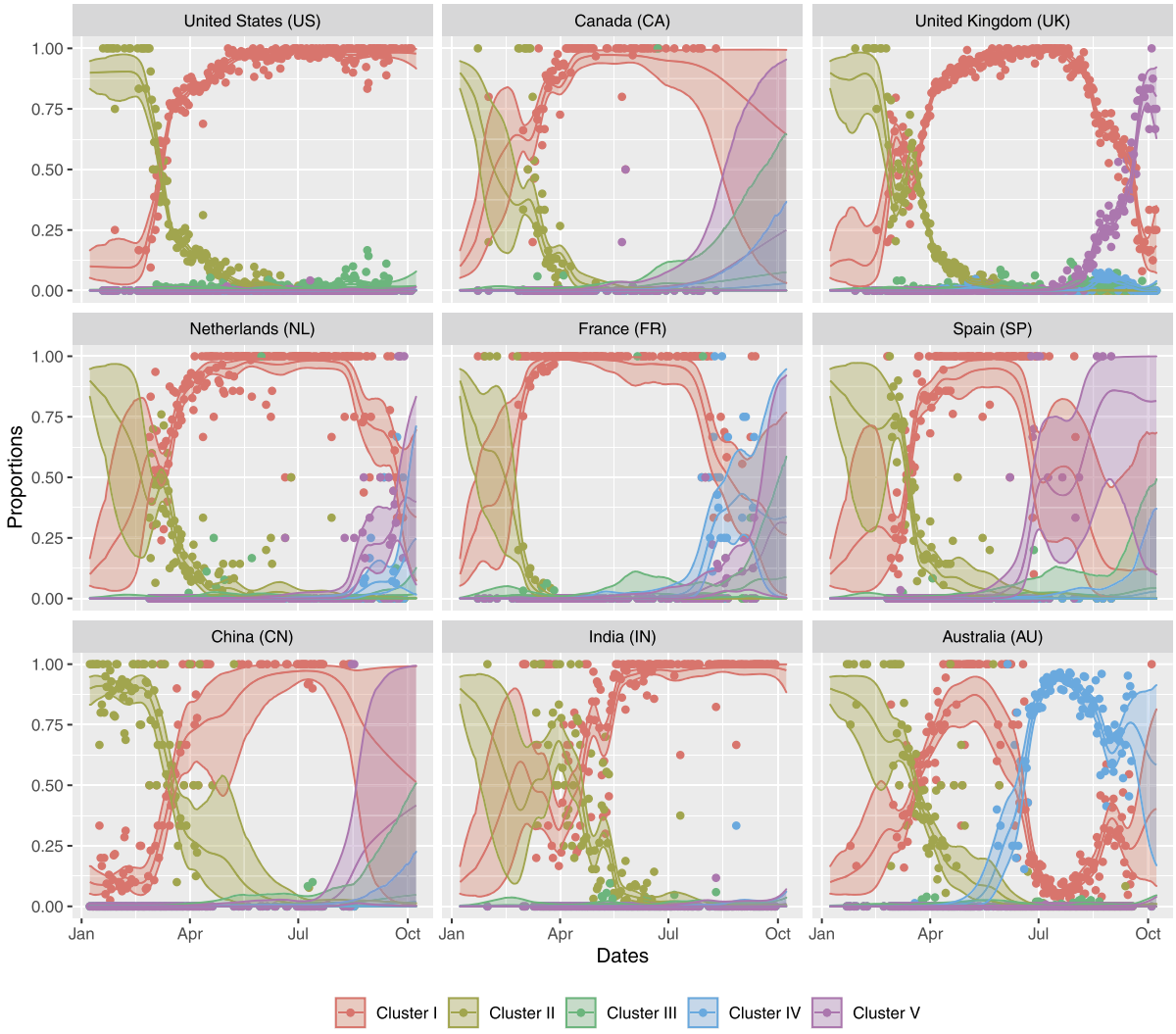

Figure 4: Posterior means and 95% credible intervals of inferred cluster proportions from the Bayesian hierarchical model for 9 countries. For each day, the points show the observed proportions, the middle solid lines indicate the posterior means of cluster proportions, and the bands indicate the 95% credible interval of cluster proportions.

To ensure that our results are robust, we performed a sensitivity analysis on the posterior parameters given different sets of priors, with a special focus on the sensitivity of $\boldsymbol{\alpha}$ and $\boldsymbol{p}_{i1}$. We created the following three scenarios to compare with our base scenario prior choices in Section 3.1, and the results are shown in Table 4. In scenario 1, we set the cluster proportions on the starting date as $\boldsymbol{p}_{i1} \sim \mathit{Dirichlet}(2.5, 38.5, 0.5, 0.5, 0.5)$, obtained by adding the pseudocounts to Dirichlet parameters corresponding to the Jeffreys prior, that is, $\mathit{Dirichlet}(0.5, 0.5, 0.5, 0.5, 0.5)$. In scenario 2, we set the priors for each growth parameter $\alpha_c \sim N(0, 1)$ independently for $c = 2, \ldots, 5$, instead of the uniform priors in our base scenario. In scenario 3, we combined both changes to the priors made in scenario 1 and 2. Compared with the main results in Table 3, all three scenarios in Table 4 show comparable posterior means and 95% credible intervals with

Table 4: Sensitivity analysis of the posterior parameters with different priors. Scenario 1 sets $\boldsymbol{p}_{i1} \sim Dirichlet(2.5, 38.5, 0.5, 0.5, 0.5)$. Scenario 2 sets the growth parameter $\alpha_c \sim N(0, 1)$, $c = 2, \ldots, 5$. Scenario 3 combines both changes to the priors from scenario 1 and 2.

	Scenario 1			Scenario 2			Scenario 3		
Parameter	Mean	2.5%	97.5%	Mean	2.5%	97.5%	Mean	2.5%	97.5%
α_2	−0.0492	−0.0655	−0.0351	−0.0494	−0.0656	−0.0345	−0.0503	−0.0664	−0.0352
α_3	0.0040	−0.0110	0.0188	0.0020	−0.0137	0.0173	0.0040	−0.0119	0.0195
α_4	0.0240	0.0007	0.0467	0.0177	−0.0069	0.0407	0.0242	−0.0000	0.0468
α_5	0.0341	0.0127	0.0552	0.0266	0.0050	0.0470	0.0338	0.0116	0.0548
$p_{\cdot,1,1}$	0.0839	0.0423	0.1418	0.1006	0.0528	0.1668	0.0850	0.0435	0.1447
$p_{\cdot,1,2}$	0.9157	0.8573	0.9575	0.8986	0.8319	0.9465	0.9145	0.8546	0.9563
$p_{\cdot,1,3}$	0.0004	0.0000	0.0016	0.0008	0.0001	0.0030	0.0004	0.0000	0.0017
$p_{\cdot,1,4}$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000
$p_{\cdot,1,5}$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000

the base scenario, which indicates our posterior parameters are fairly stable and robust to the different choices of priors.

4.2 Local Structural Impacts of Common Mutations

For each of the nine segments identified in Figure 3, we ran the SMC and DiSGro algorithms to sample conformations for the reference sequence and the mutated sequence. The specific segments, including the starting and ending positions, are shown in the first three columns of Table 5. Note that each of the segments considered contains a single mutation, for example, the length 15 segment 923–937 sampled for A930V does not overlap with 929–943 for D936Y since they occur in different clusters. For DiSGro (Tang et al., 2014), we used the program from the authors to generate 100000 conformations and kept the 5000 with the lowest energies as the representatives. For SMC (Wong et al., 2018), we ran the algorithm with 60000 particles as in Wong (2020), and also kept the 5000 with the lowest energies as the representatives.

To summarize the distributions of these sampled low-energy conformations in 3-D space, we used a similar metric as in Wong (2020), by computing all pairwise RMSDs between conformations. For each individual segment, let $x_R^{(1)}, \ldots, x_R^{(5000)}$ denote the 5000 conformations sampled for the reference sequence, and $x_M^{(1)}, \ldots, x_M^{(5000)}$ denote the 5000 conformations sampled for the mutated sequence. Then we computed three sets of RMSDs, defined via

$$\begin{aligned} d_{RR} &\doteq \left\{\text{RMSD}(x_R^{(k)}, x_R^{(l)})\right\} \text{ for } k, l \in \{1, 2, \ldots, 5000\} \text{ such that } k \neq l, \\ d_{MM} &\doteq \left\{\text{RMSD}(x_M^{(k)}, x_M^{(l)})\right\} \text{ for } k, l \in \{1, 2, \ldots, 5000\} \text{ such that } k \neq l, \\ \text{and} \quad d_{RM} &\doteq \left\{\text{RMSD}(x_R^{(k)}, x_M^{(l)})\right\} \text{ for all } k, l \in \{1, 2, \ldots, 5000\}. \end{aligned}$$

Thus the set d_{RR} approximately represents the distribution obtained by repeatedly sampling two random low-energy conformations from the reference sequence and computing the RMSD between those conformations; an analogous interpretation applies to d_{MM} for the mutated sequence. Meanwhile, d_{RM} considers pairwise RMSDs between one random conformation from the reference sequence and one random conformation from the mutated sequence. Visual differences

Table 5: Results for the lowest energy conformations sampled by the SMC and DiSGro algorithms on the reference and mutated sequence segments. The RMSD_R columns calculate the RMSD between the reference 3-D structure and the lowest energy conformation sampled for the reference sequence, thus measuring the accuracies of the algorithms for reconstructing each of these segments of the S-protein. The RMSD_{RM} columns calculate the RMSD between the lowest energy conformations sampled for the reference sequence and mutated sequence, thus measuring the extent to which the location of the energy mode may have shifted in 3-D space as a result of mutation.

			RMSD_R		RMSD_{RM}	
Cluster	Mutation	Sampled segment	SMC	DiSGro	SMC	DiSGro
I, III, IV, V	D614G	607–621	2.98	2.56	1.99	1.93
III, V	A222V	215–229	1.19	1.66	1.04	1.63
III	D574Y	567–581	5.35	5.13	2.40	12.19
III	H655Y	648–662	1.41	1.71	1.48	1.54
IV	S477N	470–484	3.76	14.71	3.14	13.05
IV	T632N	625–639	2.91	4.44	2.71	4.86
V	P272L	265–279	0.64	0.95	0.92	1.90
IV	A930V	923–937	0.89	0.76	1.15	1.04
I	D936Y	929–943	4.17	1.72	6.33	2.44

between the histograms of d_{RR} and d_{RM} (or d_{MM} and d_{RM}) would thus suggest that the low-energy conformations for the reference and mutated sequences lie in distinct regions of 3-D space.

Plots for d_{RR}, d_{RM}, and d_{MM} for each of the nine segments are shown in the panels of Figure 5, normalized to be probability densities and smoothed via kernel density estimation (Botev et al., 2010), labeled as Reference/Reference, Reference/Mutated, and Mutated/Mutated respectively. These RMSD distributions are largely indistinguishable for most segments, suggesting that there is little discernible impact to the local conformational space of the protein as a result of the mutation, regardless of which sampling algorithm is used. Three of the nine panels do have some visible differences between these RMSD distributions when sampled via the SMC algorithm: D574Y, A930V, and D936Y. In each case, the Reference/Mutated probability density visually appears as a compromise between the Reference/Reference and Mutated/Mutated densities, which is sensible.

In addition to the overall RMSD distributions, we may also specifically examine the lowest energy conformation, as is often done in protein structure prediction applications. First, we considered conformations sampled for the reference sequences, where the true structure is known from the PDB. For both DiSGro and SMC methods, we computed the RMSD between the true structure and the lowest energy conformation sampled by the algorithm. These results are shown in the RMSD_R columns of Table 5, which may be interpreted as the prediction accuracy if the algorithms are tasked with reconstructing the 3-D structure for each of these segments. Overall, these results show reasonable accuracies, with the segments 567–581 and 470–484 being the most difficult to predict correctly for both algorithms. Second, we considered conformations sampled for the mutated sequences, again taking the lowest energy conformation sampled by both algorithms. Here, the true structures are unknown (except for D614G) so prediction accuracy cannot be assessed in general. Thus, in the RMSD_{RM} columns of Table 5, we instead compute

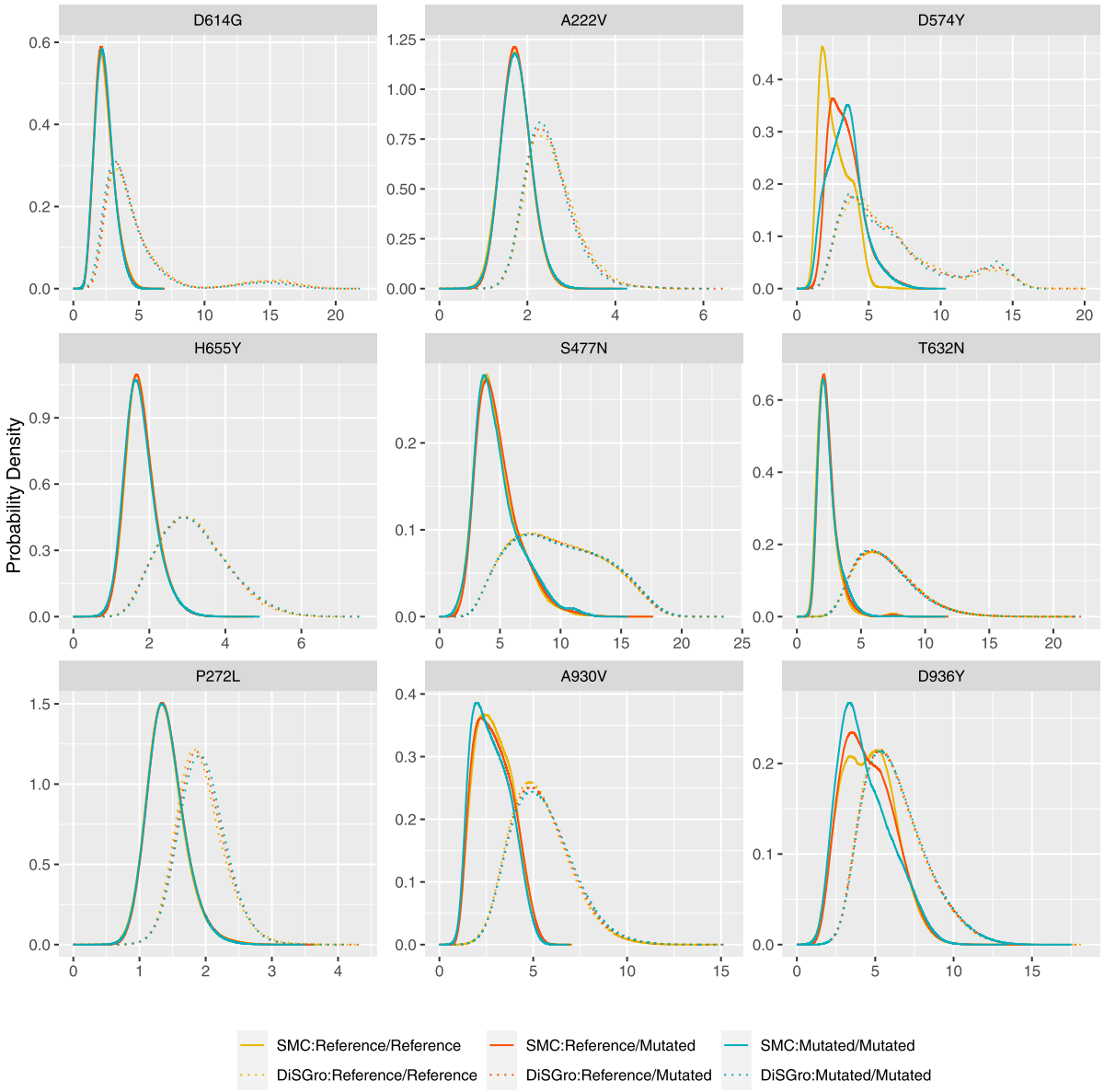

Figure 5: Probability densities of the pairwise RMSD distributions d_{RR} (Reference/Reference), d_{RM} (Reference/Mutated), and d_{MM} (Mutated/Mutated) for each of the nine segments in Table 5 based on the sampled conformations from SMC (solid lines) and DiSGro (dotted lines). The x-axes are RMSDs in units of Angstroms.

the RMSD between the lowest energy sampled conformations for the reference and mutated sequences, as a way to quantify whether the location of the mode of the energy distribution (as approximated by the samples) has shifted significantly after mutation. Here, both algorithms agree in predicting that mutations D614G, A222V, H655Y, P272L, and A930V result in relatively small local 3-D structural changes (RMSD < 2) in the lowest energy conformation, while larger local structural changes are predicted by one or both algorithms for D574Y, S477N, T632N, and D936Y.

For D614G, we may validate the sampling results as this mutation has been studied in the laboratory with a determined 3-D structure in the PDB (accession code 6XS6, Yurkovetskiy et al., 2020). The actual RMSD between the reference structure (6XM0) and 6XS6 computed over the positions 607–620 corresponding to the sampled segment is 0.38 (coordinates for position 621 are missing in 6XS6); in contrast, the RMSD when computed over the larger structural unit from positions 531 to 620 is 2.61. This result indicates that the local structural change as a result of the D614G mutation is indeed quite small, which is in agreement with the predictions of the sampling algorithms. The D614G mutation does however lead to more substantive global changes to the S-protein structure, which would be very difficult to predict computationally; general protein structure prediction remains a highly challenging problem, despite recent progress (Kryshtafovych et al., 2019).

5 Discussion

In this paper, we presented statistical approaches to tackle the challenges associated with the analysis of S-protein sequence and structure data. First, to better understand the evolution of S-protein sequences, we grouped the S-protein sequences into hierarchical clusters, and studied the spatial and temporal trend of these mutant clusters using a Bayesian hierarchical model. Second, we used sampling algorithms to investigate the possible changes in the local 3-D structure of the S-protein in the segments where the most frequent mutations occurred.

Based on our model estimates, we found that on average the reference sequence and its closely-related variants will diminish, while variants with the co-occurring mutations of D614G together with S477N or A222V tend to increase most strongly in prevalence over time. Our estimates of trend not only examined individual mutations as was analyzed in Korber et al. (2020), but also captured the prevalence of some co-occurring mutations that have so far received limited attention in the literature. Nonetheless, our findings on the reference sequence do align with Korber et al. (2020), where the authors showed that a transition of position 614 from D to G occurred in many regions around the world with varying levels of statistical significance. Our estimates of S477N and A222V are in agreement with the trends observed by the Los Alamos National Laboratory, and A222V in particular is also consistent with Hodcroft et al. (2020) where the authors reported its presence in the majority of sequences in Europe by the fall of 2020. In addition, we found spatial dependence in COVID-19 transmission across country boundaries to be low in general, but higher within Europe. This could be related to their relatively loose travel policies within EU members during COVID-19 (European Commission, 2020). Finally, a useful feature of our Bayesian approach is the ability to make projections of cluster proportions in countries where data is scarce or missing.

The result of the sequence analysis suggests potential fitness advantages or higher infectiousness for the co-occurring mutations D614G + S477N and D614G + A222V. In reality, while higher infectiousness may fully explain their growth in prevalence, other epidemiological factors may also play a role, for example, the characteristics of the infected population and the founder effect (Korber et al., 2020). Although Li et al. (2020) confirmed that D614G combined with other mutations (e.g., L5F, V341I, K458R, etc.) are more infectious than the reference sequence, the infectivities of D614G + S477N or D614G + A222V have not yet been mentioned and examined. Therefore, further experimental evidence is needed to confirm the increased infectivity of these co-occurrent mutations.

Having identified differences in the relative growth rates of the five mutant clusters considered, we examined whether the most common sequence mutations in these clusters were associated with changes in the 3-D structure of the protein near the mutation location. Based on two different sampling algorithms, we conclude that evidence for large local structure changes is generally weak. This computational result is consistent with the ground truth in the PDB for the one mutation (D614G) that has been studied in the laboratory thus far. For the mutations S477N and T632N which are associated with cluster IV, both algorithms agree that a shift in the local 3-D conformation with lowest energy might be possible (with change in RMSD greater than $\sim$3). Since protein structure determination experiments cannot keep pace with genome sequencing, we anticipate that computational approaches will continue to play an important role in understanding the possible structural impact of mutations.

Overall, S-protein sequence and structure datasets are a rich source of information that further research efforts can leverage for better understanding COVID-19, and we list some examples. First, the sequence data could be expanded to include other data sources; the sequences used in this paper were collected from GISAID only where the sequence deposition rates from some countries is low. Second, the sequence data could be combined with data on COVID-19 testing and case counts to estimate the actual prevalence of mutant clusters in different countries, in addition to their proportions. Third, many laboratories have separately contributed S-protein structures to the PDB and these could be further analyzed to quantify uncertainties associated with structure determination efforts.

Supplementary Material

The processed data, R code, and instructions for reproducing the results in this paper are provided in a supplementary .zip file.

Acknowledgement

This work was partially supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada.

References

Aitchison J (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B, Methodological*, 44(2): 139–160.

Aitchison J (1999). Logratios and natural laws in compositional data analysis. *Mathematical Geology*, 31(5): 563–580.

Amanat F, Krammer F (2020). SARS-CoV-2 vaccines: Status report. *Immunity*, 52(4): 583–589.

Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, et al. (1977). The protein data bank. *European Journal of Biochemistry*, 80(2): 319–324.

Botev ZI, Grotowski JF, Kroese DP, et al. (2010). Kernel density estimation via diffusion. *The Annals of Statistics*, 38(5): 2916–2957.

Cargnoni C, Müller P, West M (1997). Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models. *Journal of the American Statistical Association*, 92(438): 640–647.

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. (2017). STAN: A probabilistic programming language. *Journal of Statistical Software*, 76(1): 1–32.

Chen AT, Altschuler K, Zhan SH, Chan YA, Deverman BE (2020a). COVID-19 CG: Tracking SARS-CoV-2 mutations by locations and dates of interest. bioRxiv preprint: https://doi.org/10.1101/2020.09.23.310565.

Chen J, Wang R, Wang M, Wei G-W (2020b). Mutations strengthened SARS-CoV-2 infectivity. *Journal of Molecular Biology*, 432(19): 5212–5226.

Diehl WE, Lin AE, Grubaugh ND, Carvalho LM, Kim K, Kyawe PP, et al. (2016). Ebola virus glycoprotein with increased infectivity dominated the 2013–2016 epidemic. *Cell*, 167(4): 1088–1098.

Dong E, Du H, Gardner L (2020). An interactive web-based dashboard to track COVID-19 in real time. *Lancet. Infectious Diseases*, 20(5): 533–534.

Duffy S (2018). Why are RNA virus mutation rates so damn high? *PLoS Biology*, 16(8): e3000003.

European Commission (2020). Coronavirus: Commission proposes more clarity and predictability of any measures restricting free movement in the European Union. https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1555. Last checked on Dec 20, 2020.

Hodcroft EB, Zuber M, Nadeau S, Comas I, Candelas FG, Stadler T, et al. (2020). Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. medRxiv preprint: https://doi.org/10.1101/2020.10.25.20219063.

Huddleston J, Barnes JR, Rowe T, Xu X, Kondor R, Wentworth DE, et al. (2020). Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution. *eLife*, 9:e60067.

Korber B, Fischer W, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. (2020). Tracking changes in SARS-CoV-2 Spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell*, 182(4): 812–827.

Krammer F (2020). SARS-CoV-2 vaccines in development. *Nature*, 586(7830): 516–527.

Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J (2019). Critical assessment of methods of protein structure prediction (CASP) — Round XIII. *Proteins: Structure, Function, and Bioinformatics*, 87(12): 1011–1020.

Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172(9): 577–582.

Lauring AS, Andino R (2010). Quasispecies theory and the behavior of RNA viruses. *PLoS Pathogens*, 6(7): e1001005.

Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S, et al. (2020). The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell*, 182(5): 1284–1294.

Ning T, Nie J, Huang W, Li C, Li X, Liu Q, et al. (2019). Antigenic drift of influenza a (H7N9) virus hemagglutinin. *The Journal of Infectious Diseases*, 219(1): 19–25.

Onuchic JN, Luthey-Schulten Z, Wolynes PG (1997). Theory of protein folding: The energy landscape perspective. *Annual Review of Physical Chemistry*, 48(1): 545–600.

Phan T (2020). Novel coronavirus: From discovery to clinical diagnostics. *Infection, Genetics and Evolution*, 79: 104211.

Schaefer C, Rost B (2012). Predict impact of single amino acid change upon protein structure. *BMC Genomics*, 13(S4): 1–10.

Sedova M, Jaroszewski L, Alisoltani A, Godzik A (2020). Coronavirus3d: 3d structural visualization of COVID-19 genomic divergence. *Bioinformatics*, 36(15): 4360–4362.

Tang K, Zhang J, Liang J (2014). Fast protein loop sampling and structure prediction using distance-guided sequential chain-growth Monte Carlo method. *PLoS Computational Biology*, 10: e1003539.

Toyoshima Y, Nemoto K, Matsumoto S, Nakamura Y, Kiyotani K (2020). SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. *Journal of Human Genetics*, 65: 1075–1082.

Wan Y, Shang J, Graham R, Baric RS, Li F (2020). Receptor recognition by the novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS coronavirus. *Journal of Virology*, 94(7): e00127-20.

Ward JH (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301): 236–244.

Webb B, Sali A (2017). Protein structure modeling with MODELLER. In: *Functional Genomics* (M Kaufmann, C Klinger, A Savelsbergh, eds.), 39–54. Springer.

WHO (2020a). Coronavirus disease (COVID-19) situation dashboard. https://who.sprinklr.com/. Last checked on Dec 19, 2020.

WHO (2020b). Draft landscape of COVID-19 candidate vaccines. https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines. Last checked on Dec 20, 2020.

Wong SW (2020). Assessing the impacts of mutations to the structure of COVID-19 spike protein via sequential Monte Carlo. *Journal of Data Science*, 18(3): 511–525.

Wong SW, Liu JS, Kou S (2018). Exploring the conformational space for protein folding with sequential Monte Carlo. *Annals of Applied Statistics*, 12(3): 1628–1654.

Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, et al. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, 367(6483): 1260–1263.

Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798): 265–269.

Yurkovetskiy L, Wang X, Pascal KE, Tomkins-Tinch C, Nyalile TP, Wang Y, et al. (2020). Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell*, 183(3): 739–751.

Zhang J, Kou SC, Liu JS (2007a). Biopolymer structure simulation and optimization via fragment regrowth Monte Carlo. *Journal of Chemical Physics*, 126(22): 06B605.

Zhang J, Lin M, Chen R, Liang J, Liu JS (2007b). Monte Carlo sampling of near-native structures of proteins with applications. *Proteins: Structure, Function, and Bioinformatics*, 66(1): 61–68.

Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. (2020a). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798): 270–273.

Zhou T, Tsybovsky Y, Olia AS, Gorman J, Rapp M, Cerutti G, et al. (2020b). Cryo-EM structures delineate a ph-dependent switch that mediates endosomal positioning of SARS-CoV-2 spike receptor-binding domains. bioRxiv preprint: https://doi.org/10.1101/2020.07.04.187989.