Supplementary Document for "Statistical challenges in the analysis of sequence and structure data for the COVID-19 spike protein".

Shiyu He and Samuel W.K. Wong

This supplementary document describes the data and software used in this study, and provides the steps to reproduce the analyses presented in the paper.

1 Sequence dataset

1.1 Obtaining the data

The S-protein sequence data are publicly available from GISAID, accessible at https://www.gisaid.org/. The dataset analyzed in the paper was accessed on Oct 14th, 2020.

1.2 Obtaining the software

The sampling for the Bayesian hierarchical model is implemented through the rstan package (the Stan interface in R), which is free to install in R.

1.3 Clustering the sequences

The script "cluster.R" inside the R folder implements the hierarchical clustering of S-protein sequences. It imports data from "sprotein_reference.csv" and "sprotein_unique.csv", which contain the reference sequence, and all unique sequences extracted from the GISAID dataset (after incomplete sequences were removed), respectively. Figure 1 and Figure 2 show the excerpts of these two files which must be supplied. The output file "top_mutations.csv" is created and saved inside the output folder, producing the results of Table 1 in the paper.

column
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNII

Figure 1: An excerpt of the "sprotein_reference.csv" file.

column
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNII
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNII
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNII
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNII

Figure 2: An excerpt of the "sprotein_unique.csv" file.

The script also imports data from "cluster_top_seq.csv", which shows the top three unique sequences in each cluster ranked by frequency. The mutations in these top three sequences along with the information in "cluster_top_seq.csv" are saved in "top_sequences.csv" inside the output folder, producing the results of Table 2 in the paper.

After the clusters of the unique sequences are identified, the number of sequences in each cluster are calculated for each country, and are presented in Figure 1 of the paper.

1.4 Exploratory data analysis

The script "eda.R" inside the R folder carries out the exploratory analysis for the cluster proportions. The script imports data from "country_clusters.csv" files which list the IDs, names, sequences, dates, and clusters of sequences for each country. The dates are extracted from the sequence names, and the

ID	Name	Seq	Cluster	Date
Spike|hCoV	Spike|hCoV-19/USA/CA-CDPH-UC11/2020|	MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGV	2	2020-03-05
Spike|hCoV	Spike|hCoV-19/USA/MN1-MDH1/2020|20	MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGV	2	2020-03-05
Spike|hCoV	Spike|hCoV-19/USA/WA-UW95/2020|202	MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGV	2	2020-03-10
Spike|hCoV	Spike|hCoV-19/USA/WA-UW106/2020|20	MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGV	2	2020-03-11
Spike|hCoV	Spike|hCoV-19/USA/WA-UW116/2020|20	MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGV	2	2020-03-11
Spike|hCoV	Spike|hCoV-19/USA/WA-UW117/2020|20	MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGV	1	2020-03-11

Figure 3: An excerpt of the "US_cluster.csv" file.

clusters are obtained from their sequence matches with the unique sequences in the five hierarchical clusters. Figure 3 shows an excerpt of the "US_cluster.csv" file to illustrate how to supply these files.

The script file produces Figure 2 in the paper. The input data array for the Bayesian hierarchical model is also generated under the name "array_allcont.rda" in the output folder.

1.5 Sampling for the Bayesian hierarchical model

The stan file "bayes_hierarchical_model.stan" inside the R folder generates samples for the Bayesian hierarchical model and produces the output used for posterior inference.

The script "bayes_hierarchical_model.R" inside the R folder loads data from "array_allcont.rda", implements the Stan file and generates the posterior means and 95% credible intervals for parameters and the daily cluster proportions. The script also produces Figure 4 in the paper.

1.6 Sensitivity analysis for the Bayesian hierarchical model

The stan file "model_sensitivity_1.stan", "model_sensitivity_2.stan", "model_sensitivity_3.stan" inside the R folder generates the samples used for scenarios 1-3 in the sensitivity analysis for the Bayesian hierarchical model.

The script "test_sensitivity_1.R", "test_sensitivity_2.R", "test_sensitivity_3.R" inside the R folder implements the Stan files for sensitivity analysis and generates the posterior parameters for scenarios 1-3. These script files are also used to generate Table 4 in the paper.

2 Structure dataset

2.1 Obtaining the data

The 3-D protein structure data are publicly available from the Protein Data Bank (PDB), accessible at https://rcsb.org. The list of S-protein structures is obtained by navigating to "COVID-19/SARS-CoV-2 Resources" and following the link to "Spike protein and spike receptors". The specific PDB codes used for the sampling study are deposited by these authors:

- 6XM0 (Consensus structure of SARS-CoV-2 spike at pH 5.5)
  - Cryo-EM Structures Delineate a pH-Dependent Switch that Mediates Endosomal Positioning of SARS-CoV-2 Spike Receptor-Binding Domains. Zhou, T., Tsybovsky, Y., Olia, A.S., Gorman, J., Rapp, M.A., Cerutti, G., Chuang, G.-Y., Katsamba, P.S., Nazzari, A., Sampson, J.M., Schon, A., Wang, P.D., Bimela, J., Shi, W., Teng, I.T., Zhang, B., Boyington, J.C., Sastry, M., Stephens, T., Stuckey, J., Wang, S., Friesner, R.A., Ho, D.D., Mascola, J.R., Shapiro, L., Kwong, P.D. *bioRxiv*.
- 6XS6 (SARS-CoV-2 Spike D614G variant, minus RBD)
  - Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. Yurkovetskiy, L., Wang, X., Pascal, K.E., Tomkins-Tinch, C., Nyalile, T.P., Wang, Y., Baum, A., Diehl, W.E., Dauphin, A., Carbone, C., Veinotte, K., Egri, S.B., Schaffner, S.F., Lemieux, J.E., Munro, J.B., Rafique, A., Barve, A., Sabeti, P.C., Kyratsous, C.A., Dudkina, N.V., Shen, K., Luban, J. (2020) *Cell* 183: 739-751.e8.

2.2 Obtaining the software

The methods used for sampling protein segments are freely available as follows:

- SMC
  - Package for Linux systems available from: https://swong.ca/downloads/petals-smc.tar.gz
- DiSGro
  - Package for Linux systems available from: http://tanto.bioe.uic.edu/DiSGro/

2.3 Sampling 3-D structures using SMC and DiSGro

This supplementary package also includes the 3-D structures prepared for sampling using SMC and DiSGro. These are located in the "data" folder. The PDB structure "ref-6XM0A.pdb" corresponds to the reference sequence. The other PDB structures correspond to the nine mutations listed in Table 5 of the main paper. For example, "A222V.pdb" substitutes the amino acid V into the PDB structure at position 222, where the reference sequence had the amino acid A.

The script "sample3D.R" inside the R folder carries out the segment sampling using SMC and DiSGro, extracts energy and RMSD statistics from the samples for Table 5, and computes the pairwise RMSD matrices for Figure 5.

The kernel density estimates for Figure 5 are computed using the method by Botev et al. (2010). R code implementing that method can be obtained at the link https://faculty.missouri.edu/~kaplandm/code/kde.R and place "kde.R" inside the R folder. To then reproduce Figure 5, run "make-smc-densityplot.R" after executing "sample3D.R".