Modeling SARS-CoV-2 substitution processes: predicting the next variant - Supplementary information

Keren Levinstein Hallak[†] Saharon Rosset[*†]

June 24, 2021

NCBI phylogenetic tree

We perform the same analysis on the phylogenetic tree reconstructed by NCBI [1] [1] as was done in the main paper for the tree we reconstructed by applying the sarscov2phylo method by Lanfear [2] (see main paper, Table 1, Table 2 and Figure 2). The NCBI dataset contains 38,277 sequences that passed quality control out of the 61,835 sequences available on the site [2] on February 8th, 2021. The results are given in Table S1, Table S2 and Figure S1. The 3 highest-ranking models for the NCBI tree are different from those obtained in the main paper. However, the ranks of the NCBI highest-ranking models according to the phylogenetic tree reconstructed by the sarscov2phylo method are relatively high (585, 395, and 51 out of 43,254) and also vice versa, the ranks of the highest-ranking models in the main paper are relatively high according to NCBI's models ranking (1317, 88 and 2853 out of 43,254).

The results in Table S2 and Figure S1 highly resemble these from the main paper and follow the same analysis therein, confirming the robustness

[*]Corresponding author

[†]Department of Statistics and Operations Research, School of Mathematical Sciences, Tel-Aviv University, 6997801, Tel-Aviv, Israel

[1]https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/precomptree

[2]https://www.ncbi.nlm.nih.gov/sars-cov-2/

- Omission + Inclusion / Division	Gene	Nucleotide	Amino Acid	Codon	Codon position	Mature Peptide	Stem Loop	CG Pair	Right Neighbor	Left Neighbor
First models ranked by AIC	-	/	/	-	+	/	-	-	+	-
	-	/	-	/	-	+	-	-	+	+
	-	/	/	-	/	/	-	-	+	+
First models ranked by Poisson AIC	+	-	-	/	+	/	-	-	+	+
	+	-	-	/	+	/	+	-	+	+
	+	-	-	/	+	/	/	-	+	+

Table S1: **Top-scoring models for the training dataset for NCBI's reconstructed phylogenetic tree.** The first three rows correspond to the top-scoring models when NB regression is applied. The next three rows correspond to the top-scoring models when Poisson regression is used. Each explaining factor is either (−) omitted from the model, (+) used as an explanatory factor, or (/) used to split the GLM into sub-models.

of our method.

References

[1] DA Benson, M Cavanaugh, K Clark, I Karsch-Mizrachi, DJ Lipman, J Ostell, and EW Sayers. Genbank nucleic acids res 41 (d1). *D36–D42*, 2013.

[2] R. Lanfear. https://github.com/roblanf/sarscov2phylo, 2021.

	Model #	Non-synonymous amino acid substitutions						Synonymous amino acid substitutions					
		Poisson			Negative Binomial			Poisson			Negative Binomial		
		AUC	3% Lift Vs.		AUC	3% Lift Vs.		AUC	3% Lift Vs.		AUC	3% Lift Vs.	
			Random model	Base model		Random model	Base model		Random model	Base model		Random model	Base model
All genes	1	0.833	5.408	2.298	0.815	3.205	1.362	0.861	3.747	1.500	0.859	3.236	1.326
	2	0.832	5.308	2.255	0.815	3.004	1.277	0.865	3.804	1.523	0.863	3.634	1.488
	3	0.834	4.506	1.915	0.808	2.053	0.872	0.861	3.747	1.500	0.859	3.293	1.349
Spike gene	1	0.825	4.062	2.667	0.788	3.046	2.000	0.860	3.798	2.667	0.854	2.374	1.667
	2	0.821	4.569	3.000	0.786	3.046	2.000	0.880	4.273	3.000	0.874	3.798	2.667
	3	0.814	4.062	2.667	0.759	1.015	0.667	0.859	3.798	2.667	0.853	2.849	2.000

Table S2: **Prediction results for the top three models for NCBI's reconstructed phylogenetic tree.** We use the top three Poisson and Negative Binomial models from Table S1 for prediction on the test dataset. Results for the entire genome are in the first three rows, for the spike protein only in the last three. Results are shown separately for predicting amino acid substitutions (left half) and predicting synonymous substitutions (right half, these results are not discussed in the text). The first column in each quarter of the table shows the area under the ROC curve (AUC) for the corresponding prediction task and modeling approach. We highlighted the top-scoring model for every (substitution type, locus, approach) combination. Overall we obtained high AUC scores, showing the models successfully predicted many of the substitutions. The second and third columns in each quarter are 3% lift scores of each model versus the random model and the more elaborate base model (see text and Online Methods). The top models significantly outperform both baselines stressing the benefits of our approach over more naive statistical predictions. The model presented in Figure S1 (third Poisson model for non-synonymous amino acid substitutions) is also red-framed.

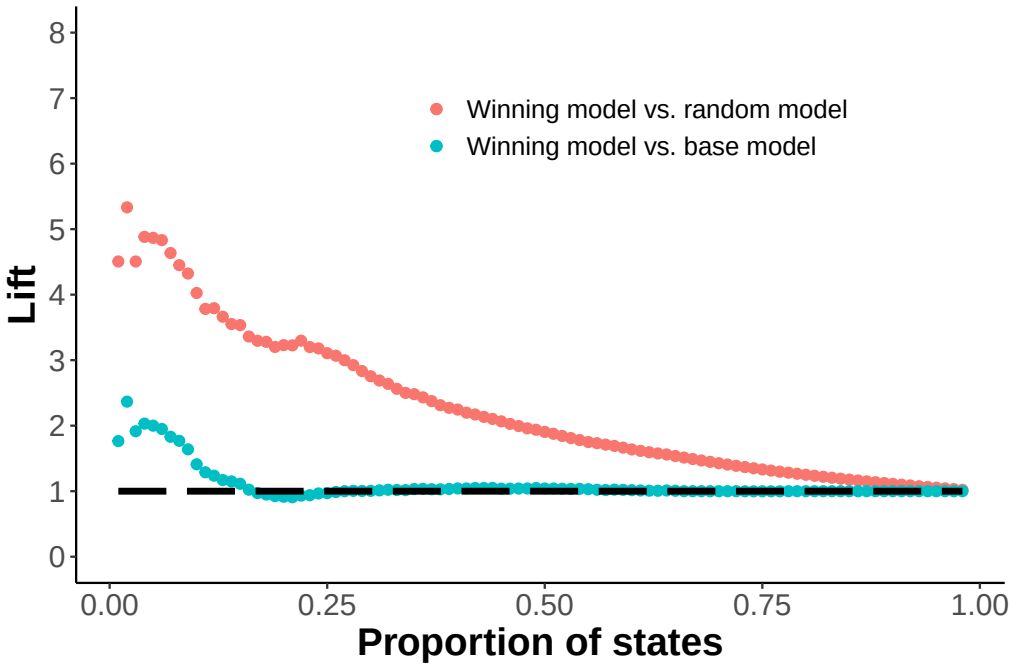

Figure S1: **Lift curves of the winning model versus the random (cyan) and base models (red) for NCBI's reconstructed phylogenetic tree.**