Monte Carlo Simulations of SARS-CoV-2 evasion from T-Cells

Overview

SARS-CoV-2's spike (S) glycoprotein—the crown-shaped surface protein that docks the virus onto human cells—is also the main signal that our body's cytotoxic T cells use to identify and kill infected cells. These T cells scan short nine-amino-acid "signatures" (epitopes) from the spike protein; if an epitope matches a receptor the cell has learned from prior infection or vaccination, the infected cell is destroyed. These epitopes essentially mark viruses as targets, and by default the virus has around 200 of them in its sequence. The purpose of this project was to simulate the virus mutating via the Monte Carlo method, and see if the virus could gain more T-cell epitopes (becoming more susceptible to T-cell attack and less infectious) or lose the T-cell epitopes it already has (becoming less susceptible to T-cells, evasion) after mutating (**Figure 1**).

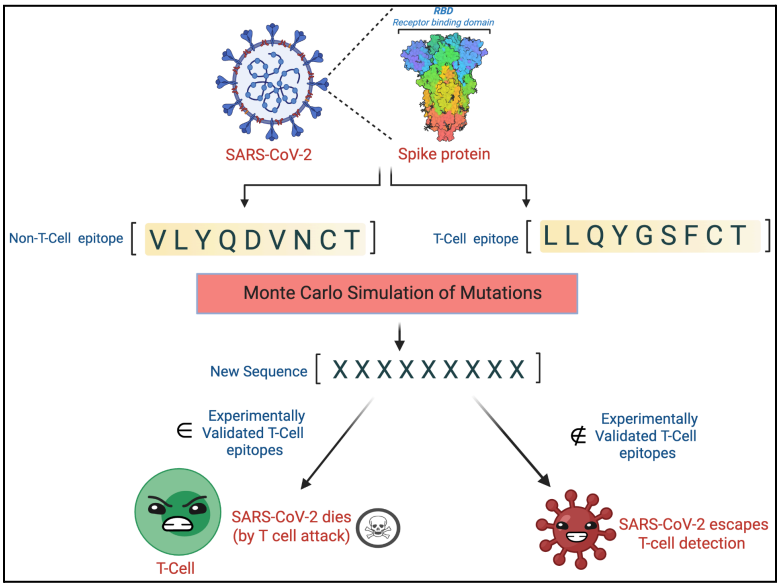


Figure 1: Project Overview

Methodology

Experiment 1:

- Slide a 9 mer window across the coronavirus spike protein sequence S_{wt} of length 1273 amino acids.
- For N = 100,000 independent trials:
  - Draw K = Binom(L, μ) mutations
  - At each of the positions in K, substitute a different amino acid with equal probability for all 19
  - Count # new epitopes found after mutating
  - Count # old epitopes disappeared after mutating

- Calculate final probabilities and the net change in epitopes

In addition to this experiment, I performed two others. Experiment 2 is similar to Experiment 1 but we count the number of mutational cycles until at least one epitope is gained and the first epitope loss independently. We then repeat the experiment 100 times and graph the number of cycles required (Figure 3 & 4). If we don't achieve a result by 100,000 cycles, we simply return 100,000 to signify a larger amount needed. While Experiment 1 focuses on identifying probabilities based on the frequentist view, Experiment 2 takes the geometric/exponential "trials until success" approach.

Finally, Experiment 3 is similar to Experiment 1 but instead of assuming all substitutions are equally likely, I utilized a BLOSUM62 substitution matrix of probabilities to substitute more likely amino acids (for more details see Appendix).

Results

	$E = 279$ Epitopes only from SARS-CoV-2	$E_2 = 4222$ Epitopes from any virus	Utilizing BLOSUM62 matrix and E_2 dataset	Increasing μ (mutational rate) to 0.01
p_gain	0.00002	0.00002	0.00004	0.01875
p_loss	0.00987	0.00996	0.01024	0.99998
average epitope change (gain - loss) for 100,000 trials	-0.02087	-0.02111	-0.0213	-19.8846

Table 1: Key Probabilities for each Test - Probabilistic Monte Carlo Simulation

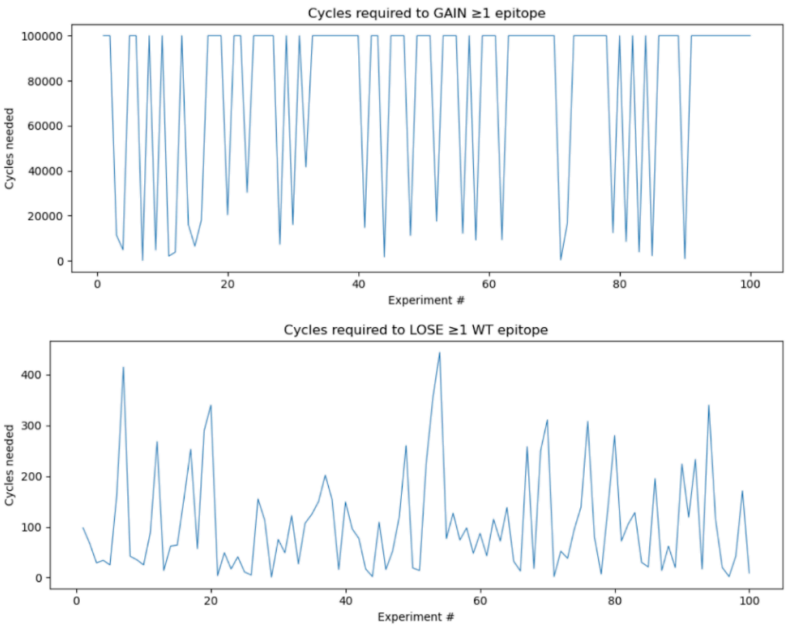


Figure 3: All substitutions equally likely

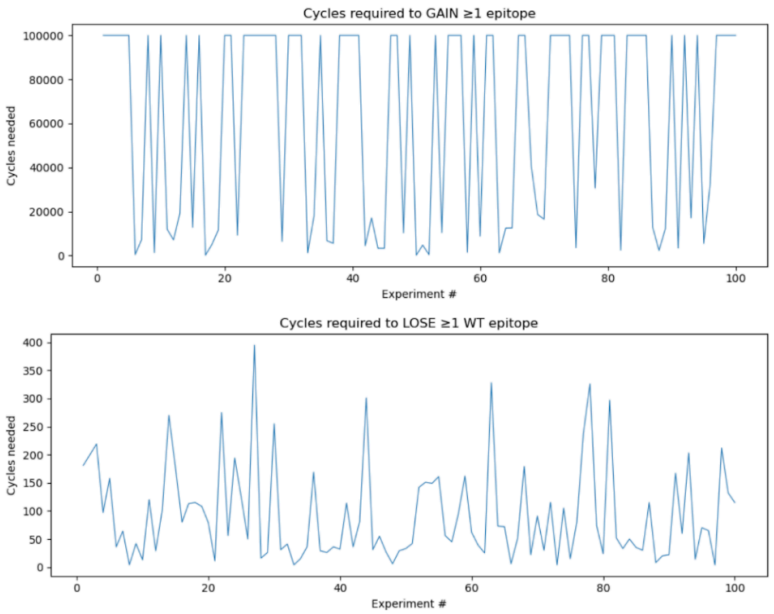


Figure 4: Substitutions Weighted by BLOSUM62

Discussion

Initially, I tested a set of T-cell epitopes for SARS-CoV-2 exclusively. However only 279 epitopes existed in the database, and only about 50 of these epitopes were not already present in the spike protein. The odds of the virus mutating to one of these new 50 epitopes are very low, shown by p_gain = 0.00002 (**Table 1**). The scenario of a net gain (mutating to a new epitope without losing an existing one) is a zero probability and did not occur once in 100,000 trials, also exemplified by the mean net change in epitopes being slightly negative (-0.02087 - Table 1). Another reason for the low probability is the low mutation rate governed by binomial K. We expect E[x] = np or 1273 * 0.00001 or 0.01273 mutations occur per iteration, so most iterations will not see a mutation. Increasing the mutational rate shown in the last column of Table 1 greatly increases both the probability of gaining and losing epitopes (0.01875 and 0.99998). Using a larger dataset of epitopes slightly increases the probability of gaining an epitope (**Table 1, Col 2**), and utilizing the BLOSUM62 matrix rather than uniform random substitutions results in the highest probability of gaining epitopes (**Table 1, Col 3**).

When running Experiment 2 with cycles required to find our first gain/loss (**Figure 3 and Figure 4**), we notice that a significantly higher number of cycles are required to gain our first epitope versus losing the first epitope. Lots of experiments resulted in a cycle count of 100,000, indicating that none of the cycles saw a new epitope gained. When comparing Figure 4 to Figure 3, it seems that on average, less cycles were required to gain the first epitope, and less experiments were hitting the 100,000 mark.

In terms of future work for this project, one avenue is exploring external tools to identify the binding affinity of certain mutated 9-mer sequences, since some mutated sequences can still elicit an immune response without matching a known T-cell epitope exactly. For example, a tool called NetMHCPan returns a binding affinity to HLA alleles, and we can use a cutoff in the monte carlo simulation to signify a good/bad epitope. Another way to extend the project is modeling cycles as a gamma or beta distribution, with the posterior as an update from Monte Carlo trials, something like Beta(num success trials, total - num success).

Overall, these results suggest it is significantly harder for the coronavirus to gain new epitopes than lose existing ones upon mutation. This implies that the virus, when it mutates, will have a much easier time evading T-cells, making it much more deadly. While earlier studies have mapped individual spike epitopes (Grifoni et al., 2020) or catalogued escape mutations in circulating variants (Korber et al., 2020), they mostly take a retrospective lens—collecting mutations after they arise in nature and then reporting how immunity responds In contrast, I start at the other end: I let every residue mutate according to BLOSUM-62 and a uniform random, run thousands of Monte Carlo trials, and measure the raw chance that any mutation will add or erase a T-cell target. This simplistic probability-first approach shows that epitope loss outpaces gain by orders of magnitude, showing how random drift by itself is enough to help a virus evolve to become more deadly.

Appendix

Full Methodology:

Experiment 1:

- Slide a 9 mer window across the coronavirus spike protein sequence S_{wt} of length 1273 amino acids.
    - Let $W = \{S_{WT}[i: i + 9] \mid 0 \leq i \leq 1273\} \cap E$ where E is a curated catalogue of experimentally validated T-cell epitopes known to trigger an immune response from IEDB Epitope database. Thus W represents all 9 window epitopes the virus already has.
- For N = 100,000 independent trials:
    - Draw K = Binom(L, μ) mutations where L = 1273 and μ is the mutation rate of coronavirus, defined as 10^{-5} in literature. K represents the number of mutations to perform across the whole sequence
    - At each of the positions in K, substitute a different amino acid with equal probability for all 19
    - Extract $H = \{S_{Mutated}[i: i + 9] \mid 0 \leq i \leq 1273\}$
    - **Gain set:** G = H \ W (# new epitopes found after mutating)
    - **Loss set:** L = W \ H (old epitopes disappeared after mutating)
    - increment variables G, L, and add to an array storing net changes in epitopes G - L
- Find $p_{gain} = \frac{G}{N}$, $p_{loss} = \frac{L}{N}$, $mean\ \Delta\ in\ epitopes = \frac{1}{N}\sum_{t=1}^{N}(|G_t| - |L_t|)$

Experiment 2:

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	**9**	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	**4**	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	**4**	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	**7**	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	**4**	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	**6**	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	**6**	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	**6**	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	**5**	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	**5**	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	**8**	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	**5**	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	**5**	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	**5**	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	**4**	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	**4**	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	**4**	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	**6**	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	**7**	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	**11**

Figure 2: BLOSUM62 Matrix

Scientists built the BLOSUM62 matrix by comparing many naturally related protein sequences, counting which amino-acid swaps show up most or least often. For every pair X, Y of amino acids, the matrix provides an integer $S(X,Y) \in \{-4, -3, ..., +11\}$ where S > 0 indicates X->Y is more likely. Each number is a base-2 log odds ratio, so to get a probability I calculated:

$w(X,Y) = 2^{S(X,Y)}$, $P(Y|X) = \frac{w(X,Y)}{\sum_{Z \neq X} w(X,Z)}$ where Z iterates over the 19 possible substitutions. I used this probability as a weight when choosing which amino acid to substitute.

Distribution/Frequency Tables:

0	99003
-4	60
-1	376
-5	23
-2	339
-3	186
-8	8
-6	2
-7	3

NET CHANGE

0	99998
1	2

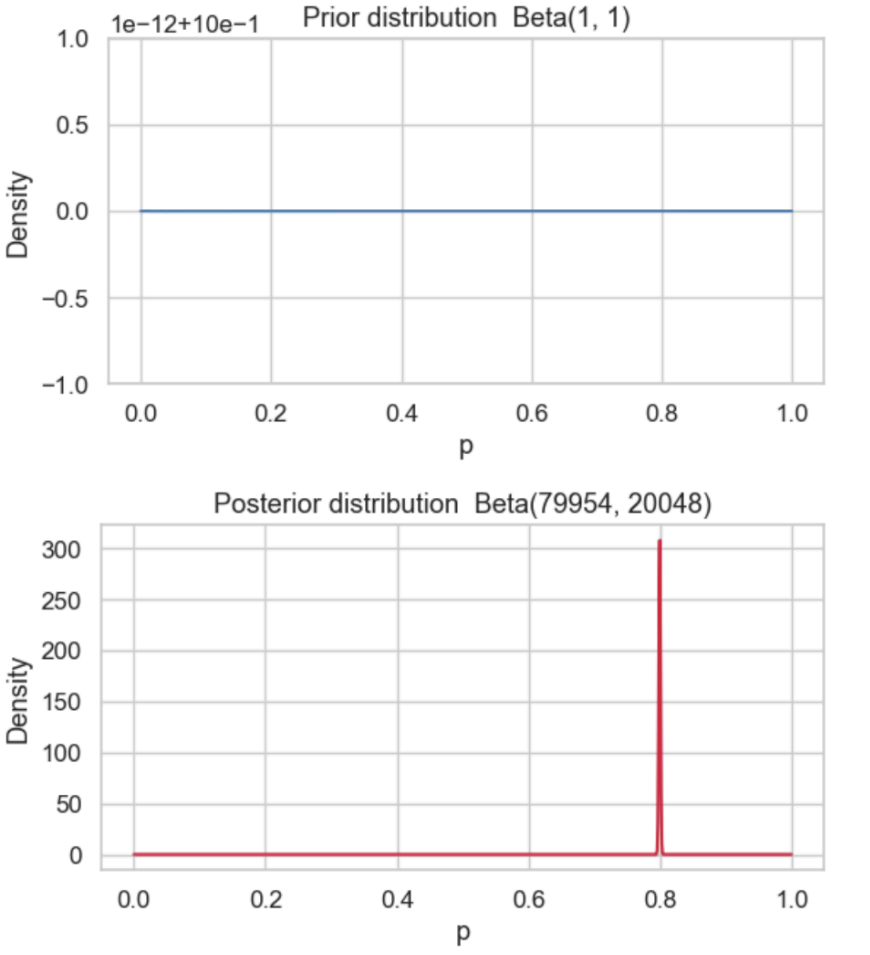
 # GAIN

0	99003
4	60
1	375
5	23
2	339
3	187
8	8
6	2
7	3

 # LOSS

The left column represents the number of epitopes and the right column represents the number of trials out of 100,000 which represent that number. The first table is the net change (gains - losses), and we can see that not a single trial resulted in more epitopes gained than lost as the highest value is 0 and the rest are negative. In 8 of the trials, 8 epitopes were lost, representing the max of the frequency data. The other two tables represent independent gains and losses. I

also briefly modeled the number of cycles (out of 100,000) required to elicit the first epitope gain as a Beta(a,b) shown below, though this part requires further analysis/interpretation for its usefulness which is why it wasn't included in the main writeup.



References:

1. **Browne, H. J., Gupta, S., & Smith, D. J.** (2023). *BLOSUM-guided deep mutational scanning reveals constraints on SARS-CoV-2 spike evolution* (Version 1). bioRxiv. https://doi.org/10.1101/2023.01.17.524472

2. **Callaway, E.** (2023). *How SARS-CoV-2 continues to evolve and evade immunity*. *Nature Reviews Microbiology,* 21, 459-460. https://doi.org/10.1038/s41579-023-00878-2

3. **Grifoni, A., Weiskopf, D., Ramirez, S. I., et al.** (2020). *Targets of T-cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. Cell, 181*(7), 1489-1501.e15. https://doi.org/10.1016/j.cell.2020.05.015

4. **Korber, B., Fischer, W. M., Gnanakaran, S., 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.e19. https://doi.org/10.1016/j.cell.2020.06.043

5. **Nielsen, M., Lundegaard, C., & Lund, O.** (2004). *Prediction of MHC binding and antigenic peptides. Nature Biotechnology, 22*(8), 1035-1036. https://doi.org/10.1038/nbt0804-1035

6. **Tarke, A., Sidney, J., Methot, N., et al.** (2021). *Negligible impact of SARS-CoV-2 variants on $CD4^+$ and $CD8^+$ T-cell reactivity in COVID-19 exposed donors and vaccinees. Journal of Experimental Medicine, 218*(6), e20210211. https://doi.org/10.1084/jem.20210211
7. **IEDB database**: https://www.iedb.org/