CS523 Complex Adaptive System Assignment 2: Viral Evolution

1st Alana Chigbrow
Dept. of Computer Science
University of New Mexico
Albuquerque, USA
achigbrow@unm.edu

2nd Jason Stewart
Dept. of Computer Science
University of New Mexico
Albuquerque, USA
jastewart@unm.edu

3rd Dawud Shakir
Dept. of Computer Science
University of New Mexico
Albuquerque, USA
dawud@unm.edu

***Abstract*—This paper explores the concepts of mutation, neutral networks and antigenic evolution using various computational tools and methods and the genome of the original strain of the SARS-CoV-2 virus. We perform generalized calculations, create neutral networks and antigenic maps based on Bloom's antigenic escape calculator, and finally generate a likely probability of synonymous mutation in light of a single mutation and compare various viral amino acid frequencies.**

***Index Terms*—SARS-CoV-2, Antigenic Map, Neutral Network, Mutation**

I. Introduction

Viruses have one of the highest mutation rates, second only to viroids [12]. Their mutation rate is so high that each offspring of a parent virus is likely to have at least one to two mutations that make them distinct from their parent [12]. Mutations are most often deleterious [12], though some are either beneficial or synonymous. When a mutation is synonymous, it indicates that the offspring have a different genotype than the parent but the same phenotypic expression of that genotype. The work of Andreas Wagner indicates that this robustness of phenotype in the face of genotype change creates neutral networks that help an organism bridge the gap to the next best-suited phenotypic variation, all while maintaining its current level of successful adaptation to the fitness landscape [13]. These next-best-suited phenotypic variations in viruses are often variations that allow the virus to escape the immune response of its host. This is called antigenic evolution. It occurs when the mutation causes a change in the virus that makes it more difficult, or impossible, for the host's immune system to recognize the antigen despite previous exposure [13]. In the case of SARS-CoV-2, there is significant antigenic mutation in the Receptor Binding Domain (RBD) of the Spike protein because this is where the host antibodies bind. Changes in this region make it more difficult for the immune response to recognize the virus, impeding the immune response. High viral mutation rates combined with the ability to generate and exploit large neutral networks give viruses a metaphorical leg up in their ability to generate antigenically different variants. In this project, we seek to explore how expansive phenotypic and antigenic neutral networks can grow using the original strain of SARS-CoV-2, Wuhan-hu-1, as our base model [1].

In Part I, we consider the full Wuhan-Hu-1 genome with approximately 29,903 nucleotides first and the RBD, with approximately 669 nucleotides, second. We perform very generalized calculations to determine the impact of accumulated mutations on the size of the genome, phenotypic neutral network and antigenic neutral network.

In Part II, we discuss our neutral network code and how it can be used to generate antigenic neutral networks and antigenic maps for the SARS-CoV-2 virus. We also investigate how epistatic mutations effect the evolution of SARS-CoV-2.

In Part III, we expand on the calculations performed in Part I by introducing host-specific amino acid frequencies and probabilities of amino acid change given a single mutation. We experimentally confirm our method for deriving the approximate percentage of phenotypically neutral mutations given a genome size and specified number of mutations. We further perform analysis of the SARS-CoV-2 amino acid distribution and compare it to human and vertebrate-host expected frequency values.

II. Part I - Calculations

In this section, we estimate the fraction of mutations that are synonymous during viral and antigenic evolution, explain our methods, and provide an estimate of the expected number of synonymous mutations.

Questions A 1-7 consider synonymous mutations in the original Wuhan strain.

QA1: How many genomes are one basepair mutation away from the original strain?

The full genome of the Wuhan-Hu-1 strain is 29,903 nucleotides [1]. The number of genomes that are one basepair mutation away from the original strain is given by the equation:

$$\binom{29,903}{1}3^1 = \frac{29,903!}{(1)!(29,903-1)!}3^1 = 89,709.$$

Therefore, there are 89,709 genomes that are one base pair away from the original strain.

QA2: How many genomes are synonymous and non-synonymous?

An estimate for synonymous substitutions is the number of possible amino acids divided by the number of possible basepair permutations. There are 20 amino acids. Each amino acid is coded by 3 basepairs, so there are $4^3 = 64$ uniquely ordered combinations of A, C, G, and T. In addition, there are roughly 3 basepair combinations for each amino acid. That means, out of all possible genomes, roughly $\frac{20}{64} = 0.3125$ mutations code for the same amino acid and are synonymous. Since there are 89,709 possible genome combinations one mutation away from the original strain, $\frac{20}{64} \times 89,709 = 28.034$ genomes are synonymous and conversely, $(1 - \frac{20}{64}) \times 89,709 = \frac{44}{64} \times 89,709 = 61,675$ genomes are non-synonymous. This is an upper bound on the amount of synonymous and non-synonymous genomes because in reality not each basepair will create an amino acid if it exists between a stop and a start codon in the genome.

QA3: How many genomes are in a neutral network of 3 mutations? (Don't count reversions back to previous genomes as new)

We use the same equation from QA1 to calculate this number, adding up the amount of genomes that are 1, 2, and 3 mutations away from the original genome:

$$\sum_{i=0}^{3} \binom{29,903}{i} 3^i = 1.20 \times 10^{14}$$

Therefore, there are about 1.20×10^{14} genomes in a neutral network with 3 mutations.

QA4: How many genomes are one mutation away from the neutral network in QA3?

We can find the difference between a 3-mutation network and a 4-mutation network (one mutation away) by calculating the difference in size between the two neutral networks. There are

$$\binom{29,903}{4} 3^4 = \frac{29,903!}{4!(29,903-4)!} 3^4 = 2.6980 \times 10^{18}$$

genome combinations with four mutations. Therefore, there are 2.6979×10^{18} genomes that are one mutation away from the 3-mutation network.

QA5: How many of these genomes (the ones 1 mutation away from the 3-step-neutral network) are synonymous and how many are non-synonymous?

We estimated in QA2 that $\frac{20}{64}$ of mutations are synonymous and $\frac{44}{64}$ are non-synonymous mutations. Following this same method for the neutral network with 4 mutations, there are an estimated $(\frac{20}{64})(2.6979 \times 10^{18}) = 8.4309 \times 10^{17}$ synonymous mutations and $(\frac{44}{64})(2.6979 \times 10^{18}) = 1.8548 \times 10^{18}$ non-synonymous mutations that are one mutations away from the 3-step-neutral network.

TABLE I

Growth rate of the **full** Wuhan-Hu-1 genome between k_i and k_{i+1} mutations for $n = 29,903$ basepairs

k	Number of Genomes	Est. Neutral	Increase from k-1
0	1	–	–
1	89709	28034	28034x
2	4.02×10^9	1.26×10^9	44812x
3	1.20×10^{14}	3.76×10^{13}	29850x
4	2.70×10^{18}	8.43×10^{17}	22500x

QA6: How fast are synonymous and non-synonymous mutations accumulated one mutation away from an n-step-neutral network?

To find how fast mutations accumulate, we computed the genome sizes for 0 to 4 mutations (k-values) and then used the synonymous ratio used in questions QA2 and QA5 to determine how quickly synonymous and non-synonymous mutations accumulate.

In Table I, we divide the estimated number of mutations for the k^{th} mutation and by the estimated number of mutations for the $(k-1)^{\text{th}}$ mutation in order to find the accumulation rate of synonymous growth after each mutation. The synonymous ratio remains constant at $\frac{20}{64}$ of the total number of genomes.

As we can see in the third column of Table I, the largest increase in synonymous (and by extension, non-synonymous), genomes is going from 1 to 2 mutations. This is because the number of *new* genomes increases the most from generation 1 to generation 2. After generation 2, there are fewer new genomes because we do not count genomes that mutate back to a genome in a previous generation.

QA7: What does exploration at the edges of neutral networks mean for viral evolution?

A virus that explores the edges of a neutral network is mutating into new genomes that have not yet existed in the wild. This is good for the virus because it is exploring new ways to evade immune systems and be better adapted to the environment it spreads in between hosts. On the other hand, this is not good for humans because it means the virus is more likely to find an adaptation that lets it escape our immune response. One goal of modern medicine that we discussed in Complex Adaptive Systems lecture [20] is forcing a virus to mutate in specific ways within its neutral network. If successful, we can make the virus evolve itself into a corner where no matter how its genome is mutated our immune system will still be able to recognize and eradicate the virus.

Questions B 1-7 consider antigenically synonymous mutations in the RBD of the original Wuhan strain.

QB1: How many genomes are ***one*** *basepair mutation away from the original strain's receptor binding domain (RBD)?*

There are 223 amino acids in the Wuhan spike protein and 669 necleotides. The number of genomes that are one basepair mutation away from the original RBD strain is given by the equation:

$$\binom{669}{1}3^1 = \frac{669!}{(1)!(669-1)!}3^1 = 2,007$$

Therefore, there are 2,007 genomes that are one basepair away from the original RBD strain.

QB2: How many genomes are synonymous after <u>one</u> mutation? How many are non-synonymous?

We used the Bloom calculator to identity sites that were clearly antigenically active. We chose 5 sites — 346, 444, 484, 486, and 490 — in the Wuhan protein spike because they showed the strongest antigenic activity. These 5 sites we considered to be antigenically non-neutral. The 218 remaining sites we considered neutral. These 5 sites correspond to 15 nucleotides. If we assume that a nucleotide change in one of these 15 locations corresponds to a change in the amino acid (a broad assumption and certainly an over-count as some of the changes in nucleotides would generate a codon that would encode the same amino acid) and that each of the nucleotides can take up to 3 different values (A, G, C, T) that are not its original value, we can say that there are $15 * 3 = 45$ non-synonymous mutations (about 2.24%) and $2007 - 15 = 1,962$ synonymous mutations (about 97.76%).

QB3: How many genomes are in a neutral network with three mutations? Do not count reversions from previous genomes.

The Wuhan spike protein has 669 nucleotides. After three mutations, there will be

$$\sum_{i=0}^{3}\binom{669}{i}3^i = 1.3413 \times 10^9$$

total genomes, not all of which will be neutral. Because we are assuming that even a single nucleotide change in the antigenically non-neutral sites will generate a non-synonymous genome, we can determine how many genomes are in the neutral network by subtracting the number of nucleotides associated with those sites (15) from the total number of nucleotides in the RBD (669) and then we have $\binom{654}{3} * 3^3 = 1.2530 \times 10^9$ total genomes in the neutral network containing 3 mutations.

QB4: How many genomes are <u>one</u> mutation away from the neutral network calculated in QB3?

We have 1.2530×10^9 neutral genomes available for mutation in the neutral network. Each of these contains 669 nucleotides. We need to accumulate 1 more mutation in order to arrive at the 4^{th} generation. That additional mutation needs to account for all 669 nucleotides in order to get an accurate understanding of the number of synonymous vs. non-synonymous mutations in this final generation. We can do this in broad strokes and what is definitely an over-count by multiplying the total neutral genomes by the number of genomes produced by 1 mutation to the original genome. We have approximately:

$\binom{669}{4} = 6.70 \times 10^{11}$ total genomes in the 4^{th} generation of genomes resulting from the neutral network. Again, this is very much an over-count as there will likely be duplicates and reversions in the genomes.

TABLE II
ACCUMULATION OF ANTIGENICALLY NEUTRAL MUTATIONS AFTER K MUTATIONS IN N=669 BINDING SITES

k	Number of Genomes	Est. Neutral	Increase from k-1
0	1	–	–
1	2,007	1962	28034x
2	2.01×10^6	1.97×10^6	1001x
3	1.34×10^9	1.31×10^9	667x
4	6.70×10^{11}	6.55×10^{11}	500x

QB5: How many genomes one mutation away are synonymous? How many are non-synonymous?

Using the estimate of antigenically non-synonymous mutations from QB2, we have $\frac{\text{218 neutral sites}}{\text{223 total sites}}(6.70 \times 10^{11} \text{ genomes}) = (0.9776)(6.70 \times 10^{11} \text{ genomes}) = 6.55 \times 10^{11}$ antigenically synonymous genomes and $\frac{\text{5 non-neutral sites}}{\text{223 total sites}}(6.70 \times 10^{11} \text{ genomes}) = (0.02242)(6.70 \times 10^{11} \text{ genomes}) = 1.50 \times 10^{10}$ antigenically non-synonymous genomes.

QB6: How fast are synonymous and non-synonymous mutations accumulated one mutation away from an n-step-neutral network?

To find how fast antigenically neutral mutations accumulate, we computed the number of genomes for 0 to 4 mutations (k-values) and then used the synonymous percentage calculated in QB2 to determine how quickly synonymous and non-synonymous mutations accumulate.

In Table II, we divide the estimated number of mutations for the k^{th} mutation and by the estimated number of mutations for the $(k-1)^{\text{th}}$ mutation in order to find the accumulation rate of synonymous growth after each mutation. The synonymous ratio remains constant at 97.76% of the total number of genomes.

The results in Table II mirror the results in Table I. This is not a surprise since both tables were calculated using the same formulas. The only difference is the ratio used to estimate the number of neutral mutations. Again, the largest increase in synonymous (and by extension, non-synonymous), genomes is going from 1 to 2 mutations. This is because the number of *new* genomes increases the most from generation 1 to generation 2. After generation 2 there are fewer new genomes because we do not count genomes that mutate back to a genome in a previous generation.

QB7: What does exploration at the edges of neutral networks mean for antigenic evolution?

When we reduce the scope of non-synonymy from overall phenotype to antigenically important change, we reduce the number of variants and subsequently increase the number of synonymous genomes. This means that the neutral network available to the genome, that is the evolutionary space it has to explore in order to either find a genome that is better suited to the current fitness landscape or that is best suited to an evolving fitness landscape, is much larger. The resulting genomes are phenotypically more variable because they have access to a broader number of ”novel phenotypes through mutation” [13].

It is tempting to think this amplified phenotypic variability is ”not a big deal” in the case of viruses because we are talking about antigenic stability. That is, the immune system’s antibodies can still recognize and opsonize the virus with the synonymous genome so that can be destroyed by macrophages [15]. That would mean that any antigenically synonymous mutations would be nixed by the host’s immune system and thus the virus would not have the opportunity to fully exploit its increased phenotypic variability in its search for the best possible phenotype for the fitness landscape. Unfortunately, that is not the case with regard to SARS-CoV-2. People who have this infection are generally most contagious before they even have symptoms of illness [16]. This means that it is still possible, despite the antigenic synonymy, to spread these more phenotypically variable genomes. This can lead to antigenic mutations that are more fit than their parent genome and that are one or more steps closer to a novel antigen that the host’s immune system has not encountered before.

QC: Assume 0.001 percent of all neutral mutations compensate for some deleterious effect of one mutation that escapes antibodies. What are the chances of antibody escape?

Non-functional mutations compensate for deleterious mutations and are epistatic, allowing mutations in antigenically active sites to persist. We can estimate the effect of epistatic mutations on non-neutral genomes by multiplying the number of genomes in column 2 in Table II by 0.0224, the ration of non-synonymous antigenic mutations to synonymous antigenic mutations calculated in question QB2, and then adding this number to the amount of neutral mutations in column 3 in Table II. However, without performing these calculations it is easy to see that if 0.001% of non-neutral genomes are now considered neutral due to epistatic changes, it is still orders of magnitude lower than the amount of neutral genomes that already exist. As a result, we can conclude that the effect of epistatic mutations is negligible when considered against the total number of neutral genomes.

We still cannot discount epistatic mutations entirely. A more thorough analysis of epistatic changes was performed in Part 2. Table III show that epistatic changes can indeed impact the neutral network the SARS-CoV-2 virus generates. We describe this phenomena in Section III-F.

III. Part II - Antigenic Neutral Network and Antigenic Map

We implemented an antigenic neutral network in python within the *src/neutral_network* directory in the project repository. The file *antigenic_neutral_network.py* contains the code for an AntigenicNeutralNetwork object and the supporting functions to build a neutral network in memory, display the neutral network graphically, and save a titer table used to build an antigenic map from the neutral network. The file *driver.py* is the driver script that builds a neutral network by calling the code in *antigenic_neutral_network.py*. In this section, a ”node” refers to a mutated virus in the neutral network. All nodes but the root node will have at least one of their antigenic binding sites mutated, except in the epistatic case when the root node will also contain mutations.

A. Variables

Within *driver.py* the following variables can be adjusted:

- Parameters for adjusting the neutral network and antigenic map:
  - **tolerance:** The threshold for which mutations are considered neutral. Any mutations with a Bloom escape value [9] above this tolerance are considered neutral.
  - **size:** The neutral network will be built until it reaches this many nodes in size.
  - **print_network:** Whether the neutral network should be displayed in your internet browser after it finishes building and saved as an HTML file in the data/ directory of the repository.
  - **only_mutate_neutral:** Whether only neutral nodes should be considered for further mutation while building the neutral network.
  - **save_titer_table:** Whether the titer table should be computed for use in Racmacs [17] for generating an antigenic map. The titer table is saved in data/titer.csv.
- Parameters for examining the effect of epistatic change:
  - **consider_epistatic_change:** Whether epistatic change, as defined in Part 1, Question C, in the assignment handout, should be considered while building the neutral network. If set to True, the program will calculate the effects of epistatic change and exit. The parameters in the previous section will not be used.
  - **starting_nodes:** How many normal and epistatic nodes should be mutated to produce the final probability of a mutated epistatic node being neutral compared to a non-epistatic node.
  - **times_to_mutate:** How many times each normal and epistatic node or its children should be mutated.
  - **root_escape:** How well should the root epistatic node(s) escape antibodies if they were not epistatic. In other words, this is the node’s maximum escape value from the Bloom calculator [9]. A lower value will mean the epistatic change is more antigenically powerful.

B. How the Neutral Network is Built

An antigenically-neutral network is built by calling the *build* function of an AntigenicNeutralNetwork object. The neutral network will be built according to the size and tolerance values described in Section III-A. The following is the pseudo code for the *build* function:

1) The root node is created and added to the neutral network. The root node has no mutations and an antibody escape of 1.
2) While the number of nodes is less than the desired size of the neutral network:
   a) A random neutral node is chosen from the neutral network.
   b) The node has one of its binding sites mutated and its new escape is calculated using the Bloom calculator.
   c) The node is identified as antigenically neutral or non-neutral according to the *tolerance*, as defined in Section III-A.
   d) If the neutral network is set to be printed, the x-y coordinates of the node are computed using a naïve method that attempts to cluster non-neutral nodes radially around the neutral node they mutated from. For simplicity, this method does not check for conflicting x-y coordinates or overlapping edges in the neutral network.
3) The number of total nodes and neutral nodes in the neutral network are printed to the console.

C. Computing the Antigenic Map

The antigenic map is computed according to the article in the Racmacs documentation titled ”Making an antigentic map from titer data” [18]. Since we did not directly obtain the titer data from experiments, the titer table is back-calculated using the distance table. This step is necessary because the Racmacs program [17] that creates the antigenic map can only read in titer data, not distance data.

According to Wilks [19], a distance table is a matrix that approximates the antigenic distance between each antigen-serum titer. In our case, this is how antigenically different each node in our neutral network is from all other nodes. The distance table is created in the *make_titer_and_distance_tables* function in *antigenic_neutral_network.py*. The antigenic distance between nodes i and j is computed as

$$distance_{ij} = |1 - escape_i/escape_j|$$

For example, if node i has an escape of 0.9 and node j has an escape of 0.7, the distance between the two nodes is

$$distance_{ij} = |1 - 0.9/0.7| = 0.286$$

. Once the distance table is created, the titer table is computed by reversing the procedure outlined by Wilks in [19]. First, a column base of 10 is assumed for all columns. We experimented with random column bases spread between 5 and 20 and column bases based off the maximum value of each column in the distance table, but none of these options made a noticeable difference in the resulting antigenic map. Therefore, we chose to use the simplest method of assuming a column base of 10 for all columns. Using a column base of 10, the titer table values are computed as follows:

$$titer_{ij} = 2^{10 - distance_{ij}}$$

This formula produces a higher titer value for nodes that have a smaller antigenic distances between them, following Wilks' definition in [19]: ”Higher values (higher titers) imply greater similarity between a serum and an antigen”. The final titer table is saved as *titer.csv* in the *data* directory of our GitHub repository. We used the Racmacs program [17] to generate an antigenic map from our titer table, following the instructions in the article ”Making an antigenic map from titer data” [18]. We display our antigenic maps in Section III-E.

D. Computing the effect of epistatic change on antigenic neutrality

We implemented our neutral network such that it can also be used to estimate how epistatic change has an effect on the virus' ability to escape antibodies. To perform this estimation, two neutral networks are built: one with an epistatic node as its root and another with a non-epistatic, non-mutated node at its root. The epistatic neutral network varies from a normal neutral network in the following ways:

- The root node is mutated until it has an antibody escape lower than the value of **root_escape**, defined in Section III-A. Then this escape is adjusted back to 1 by adding an ”escape adjustment” to the node. For example, if the root node has an escape of 0.6, its escape adjustment will be 0.4. More simply, the epistatic change compensates for a loss of 0.4 escape.
- Each new node in the epistatic neutral network inherits the epistatic change (i.e. escape adjustment) from the root node. The new node is classified as neutral if its Bloom escape value plus the escape adjustment is higher than the value of **tolerance**, defined in Section III-A.

Using the variables defined in Section III-A, **starting_nodes** epistatic and non-epistatic neutral networks are built until they reach a size of **times_to_mutate**. Then the average ratio(s) of neutral to non-neutral nodes are calculated for the **starting_nodes** epistatic and non-epistatic neutral network(s) and printed to the console.

E. Neutral network and antigenic map results

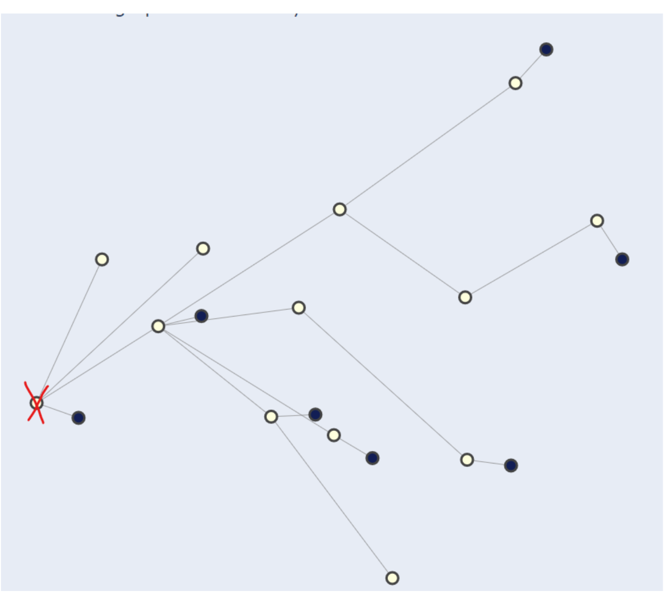


Fig. 1. A neutral network of size 20 and tolerance 0.95. Filled-in nodes are non-neutral. The root node is marked with the red ’X’.

Using *driver.py*, a neutral network of size 20 and tolerance 0.95 was created, shown in Figure 1. The root node is marked with the red ’X’ and filled-in nodes are non-neutral. Seven of the twenty nodes are non-neutral and the root node was able to mutate at least four times while still remaining neutral.

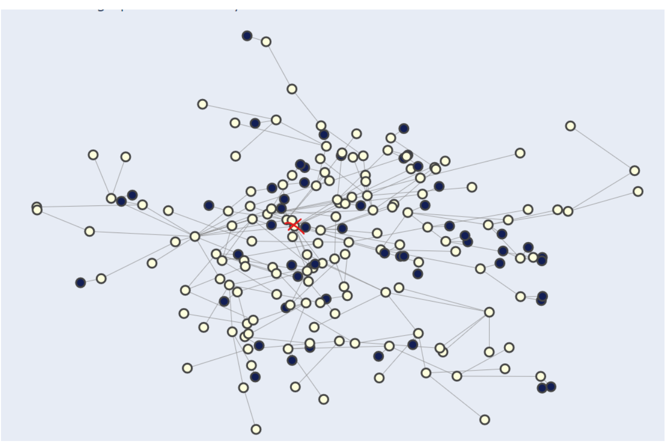


Fig. 2. A neutral network of size 200 and tolerance 0.95. Filled-in nodes are non-neutral. The root node is marked with the red ’X’.

Figure 2 shows a neutral network of size 200 and tolerance 0.95. Because our program does not check for conflicts in x-y coordinates, the network is messy; however, it can still be observed that the virus is able to accumulate a wide variety of mutations while still remaining antigenically neutral. In this network, 145 of the 200 nodes remain neutral.

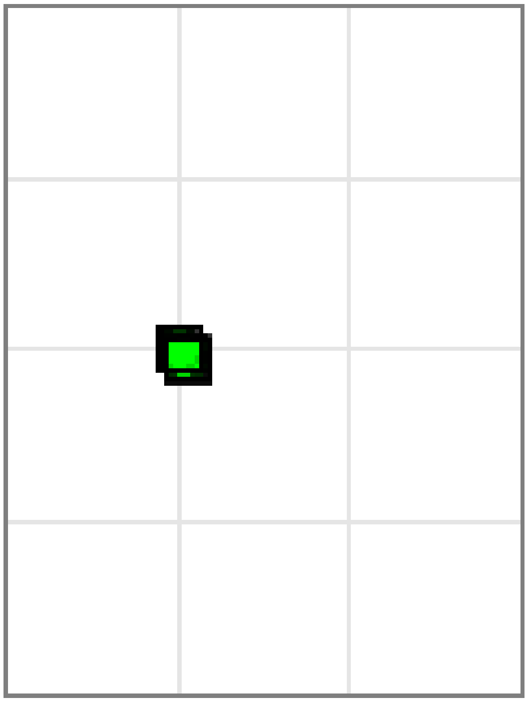


Fig. 3. An antigenic map created from the neutral network displayed in Figure 1.

Figure 3 shows the antigenic map created from the neutral network shown in Figure 1. Most of the nodes in this neutral network have similar antigenic escapes (the majority greater then 0.95); consequently, this map is not very interesting because all the nodes are clustered near each other in space. To get around this problem, we set the value of **only_mutate_neutral** in *driver.py* to False. This allows for non-neutral nodes to be mutated, creating more antigenic differences in the resulting ”semi-neutral” neutral network. The antigenic map in Figure 4 displays the greater antigenic distances that occur when non-neutral nodes are allowed to be mutated additional times. Figure 5 shows the semi-neutral network used to create the antigenic map in Figure 4.

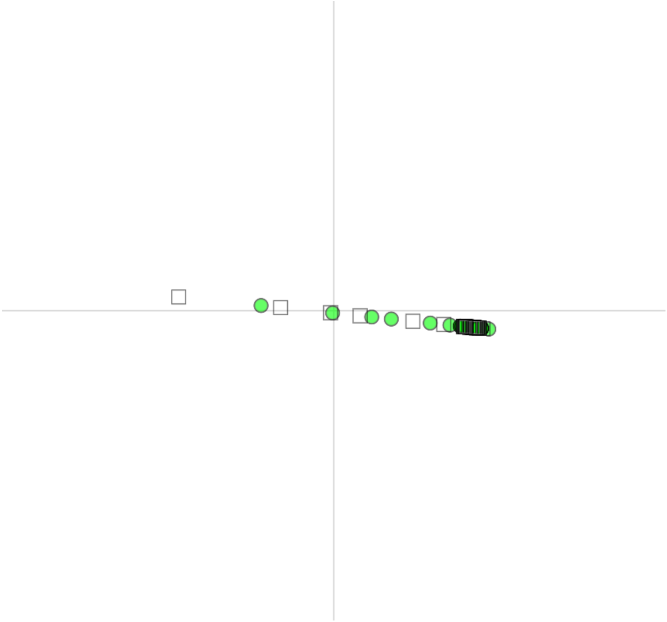


Fig. 4. An antigenic map created from a neutral network of size 50 where non-neutral nodes are allowed to be mutated additional times.

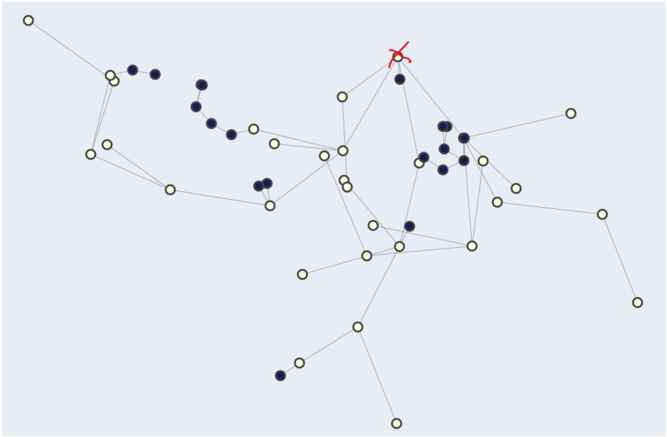


Fig. 5. The semi-neutral network used to generate the antigenic map in Figure 4

In Figure 4, the root node is in the far right portion of the map. As nodes are mutated they slowly spread to the left. This antigenic map suggests that many of the mutated nodes are antigenically similar to the root node with a few nodes on the far right displaying high antigenic differences. This hypothesis is supported in Figure 5; many of the nodes are neutral and antigenically similar to the root, but a few non-neutral nodes have been mutated multiple times to realize the greater antigenic differences that manifest themselves in the antigenic map (Figure 4).

F. Epistatic mutation results

Table III shows the results of how well epistatic change can compensate for deleterious mutations that would otherwise reduce the virus' ability to escape antibodies. Intuitively, as an epistatic change becomes more powerful (i.e. escape adjustment increases), more nodes should be neutral that would otherwise not be neutral. The results in Table III support this hypothesis - as the escape adjustment increases, more nodes in the network are neutral compared to the non-epistatic network. This means the virus is able to explore more genetic variations without scarifying its ability to escape antibodies. Although good for the virus, this has a negative effect for humans because it allows the virus to become more genetically diverse without the body recognizing it. A more genetically diverse virus is more robust and is less likely to evolve itself into a corner and become eradicated.

IV. Part III - Extended Analysis

The "back of the envelope" calculations in Part I of the project showed that the number of potential genomes created by accumulating 3 or more mutations in the Wuhan-Hu-1 genome (29903 total nucleotides [1]) is astronomical. Even using the 669 nucleotides found in the RBD of the S1 unit of the Spike protein creates a near overwhelming number of potential mutated genomes. The in-exact nature of the calculations in Part I begged the question of whether we could be more precise in the estimation of phenotypically synonymous genomes. To answer this we combined the probabilities of amino acid self-preservation (i.e. remaining unchanged) given a Single Nucleotide Substitution (SNS) found in [2] with different nucleotide frequency measures to include the frequency found in the RBD. We created a mutation function *mutate.py* (found in the mutation directory of the project repository) to experimentally verify the results of the more precise calculations.

A. Methods

Chan, Koukouravas, Yeo, Koh, and Gan calculated the probability of amino acid change given an SNS [2]. They performed the calculations for individual SNS mutations (e.g. when a T is substituted in place of another base) and for all 4 bases (A, C, G, T) together. Their findings for the combined calculations indicate a "pattern of self-preservation" [2]. That is, the probability that an amino acid will remain unchanged despite an SNS where any of the bases is substituted in is higher than the probability that it will become a different amino acid. This can be observed in the diagonal of the table they presented in their paper 6. We use this diagonal to generate a vector of amino acid self-preservation probabilities (see Table VI).

Next, we use the work of Mohanta, Mishra, Mohanta, and Al-Harrasi [3] in creating 2D maps of the viral proteome to identify frequency of amino acids in a viral proteome. They analyzed a variety of viruses and grouped them by host-type. We use the amino acid compositions for human-host virus proteome and vertebrate-host virus proteome in our calculations to provide an estimate of frequencies of the 20 possible amino acids. We choose to use human and vertebrate-host proteome in part due to curiosity as to which would be a closer match to the SARS-CoV-2 frequencies. We had to make

TABLE III

PERCENT CHANGE IN NEUTRAL NODES FOR VARYING ESCAPE ADJUSTMENTS; TOLERANCE = 0.95, STARTING_NODES = 10, AND TIMES_TO_MUTATE = 300

Escape adjustment	Non-epistatic network percent of nodes neutral	Epistatic network percent of nodes neutral	Difference
0.10	81.77%	83.77%	2.00%
0.20	80.37%	85.27%	4.90%
0.30	82.50%	87.03%	4.53%
0.40	82.50%	90.17%	7.67%
0.50	82.90%	91.87%	8.97%
0.60	81.30%	94.87%	13.57%

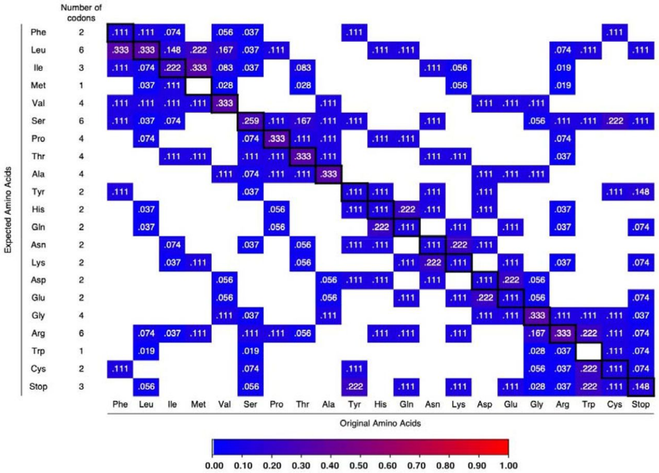


Fig. 6. Calculated probability of amino acid change given SNS [2]

two minor adjustments to the numbers reported in Mohanta, Mishra, Mohanta, and Al-Harrasi's supplementary materials. First, we divided all reported values by 100 to convert from percent of composition to decimal. Second, the sum of the frequencies of the human-host proteome was 0.0002 less than the 1.0 value it should have been likely due to rounding inaccuracy due to the number of digits reported for each value. To rectify this, we added $\frac{0.0002}{20}$ to each value so that it would sum to 1.0. The vertebrate-host proteome frequencies were likewise 0.00013 less than one and we corrected this in similar fashion. These values are found in Table VI.

We calculated two more sets of frequencies using a python script, *rbd.py*, we developed to create a histogram of the amino acid frequencies found in the Spike protein and its RBD of the Wuhan-Hu-1 viral genome [1]. These values are found in Table V.

We then calculated the overall probability of no amino acid changes for each frequency type (human-host, vertebrate-host, and RBD) by multiplying the self-preservation value for each protein ($p(x_i)$ by its frequency ($(f(x_i)$ and then summing across the entire set of weighted probabilities. We then divide this value by the number of mutations ($m = 1, 2, 3$) to find the expected percentage of synonymous mutations per generation.

$$P_{syn}(G_m) = \frac{\sum p(x_i) * f(x_i)}{m}$$

We use the above $P_{syn}(X_m)$ values to re-calculate the

TABLE IV

AMINO ACID: FREQUENCY IN THE WUHAN-HU-1 RBD — FREQUENCY IN THE FULL WUHAN-HU-1 SPIKE PROTEIN

AA	Self-Preservation	H-H Freq	V-H Freq
Ala	0.333	0.05054	0.06817
Arg	0.333	0.04901	0.05831
Asn	0.111	0.05869	0.04828
Asp	0.111	0.05532	0.05639
Cys	0.111	0.02181	0.02083
Gin	0.111	0.03646	0.03206
Glu	0.111	0.05650	0.05604
Gly	0.333	0.05148	0.05833
His	0.111	0.02222	0.02173
Ile	0.222	0.07077	0.05926
Leu	0.333	0.09374	0.09050
Lys	0.111	0.06532	0.05660
Met	0.000	0.02208	0.02516
Phe	0.111	0.04273	0.04261
Pro	0.333	0.04821	0.05219
Ser	0.259	0.07559	0.07448
Thr	0.333	0.06670	0.06020
Trp	0.000	0.01248	0.01199
Tyr	0.111	0.04059	0.03906
Val	0.333	0.05976	0.06790

TABLE V

AMINO ACID: PROBABILITY OF SELF-PRESERVATION GIVEN AN SNS — FREQUENCY IN HUMAN-HOST VIRAL PROTEOME — FREQUENCY IN THE WUHAN-HU-1 RBD — FREQUENCY IN VERTEBRATE-HOST VIRAL PROTEOME

AA	RBD Frequency	Spike Frequency
Ala	0.05381	0.06206
Arg	0.04933	0.03299
Asn	0.09417	0.06913
Asp	0.04036	0.04870
Cys	0.04036	0.03142
Gin	0.03139	0.03771
Glu	0.03139	0.04870
Gly	0.06726	0.06441
His	0.00448	0.01335
Ile	0.04036	0.05970
Leu	0.06278	0.08484
Lys	0.05381	0.04792
Met	0.00000	0.01100
Phe	0.07175	0.06049
Pro	0.05830	0.04556
Ser	0.07623	0.07777
Thr	0.05830	0.07620
Trp	0.00897	0.00943
Tyr	0.06726	0.04242
Val	0.08969	0.07620

TABLE VI
PROBABILITY OF SYNONYMOUS MUTATION GIVEN FREQUENCY OF AMINO ACID AND PROBABILITY OF SELF-PRESERVATION GIVEN AN SNS.

Frequency Type	G1	G2	G3	G4
Human-Host	0.2193	0.1097	0.0731	0.0548
Vertebrate-Host	0.2256	0.1128	0.0752	0.0564
Wuhan-Hu-1 RBD	0.2233	0.1117	0.0744	0.0558
Wuhan-Hu-1 Spike Protein	0.2251	0.1125	0.0750	0.0563

neutral network with three mutations by calculating each step in the network (generation 1, generation 2, and generation 3 having 1, 2, and 3 mutations respectively). Each generation accumulates one mutation in addition to any mutations inherited from the preceding generation(s). We only carry the synonymous, or neutral, genomes forward into the next generation to be mutated. Generations 2, 3, and 4 each have an experimentally generated new variable the represents the percentage of the calculated mutations (both synonymous and non-synonymous) that are "new" (i.e. they are not repeats of previous genomes). This value is $new = 0.855, 0.809, and 0.786$ for generations 2, 3, and 4 respectively. We arrived at these values by observing total mutations, duplicate, and reversion genomes across a number of runs of our mutation script, *mutation.py*. The first generation in the neutral network, $neutral1$, is calculated by accumulating a single mutation via the n choose k formula multiplied by 3 and then multiplying that by $P_{syn}(X_1)$. The second generation, $neutral2$, is calculated by multiplying $neutral1$ by the same n choose k formula multiplied by three and then multiplying by $P_{syn}(X_2)$. The third generation, $neutral3$, is calculated by multiplying $neutral2$ by the same n choose k formula multiplied by three and then multiplying by $P_{syn}(X_3)$. Finally, we arrive at the calculation of the fourth generation, G_4, by multiplying $neutral3$ by the n choose k formula multiplied by three. We calculate the total neutral genomes in G_4 by $neutral4 = generation4 * P_{syn}(X_4)$ and then subtract that value from G_4 to arrive at the total number of variants in $generation4$.

$$\lceil P_{syn}(G_1) * \binom{n}{1} * 3 \rceil = neutral1$$
$$\lceil neutral1 * P_{syn}(G_2) * \binom{n}{1} * 3 * new \rceil = neutral2$$
$$\lceil neutral2 * P_{syn}(G_3) * \binom{n}{1} * 3 * new \rceil = neutral3$$
$$neutral3 * \binom{n}{1} * 3 * new = G_4$$
$$\lceil P_{syn}(G_4) * G_4 \rceil = neutral4$$
$$G_4 - neutral4 = variant4$$

The mutation script that we created takes in a genome string and generates mutations of that string by spinning off 3 additional genomes for each nucleotide in the genome making the appropriate base (A, C, G, or T) substitutions for each. It uses the 'biopython' library to convert the genome into a protein and then compares that against the original Wuhan-Hu-1 virus which was also converted via 'biopython'. If the two protein chains are the same, it returns as synonymous else, it is a variant. We tested the script by using a "genome" of length 18, "TACCATGGAATTACTGCG", and confirmed that it returned the appropriate mutations and that the total number of genomes in each generation matched what we would expect given the $\binom{18}{k} * 3^k$ calculations.

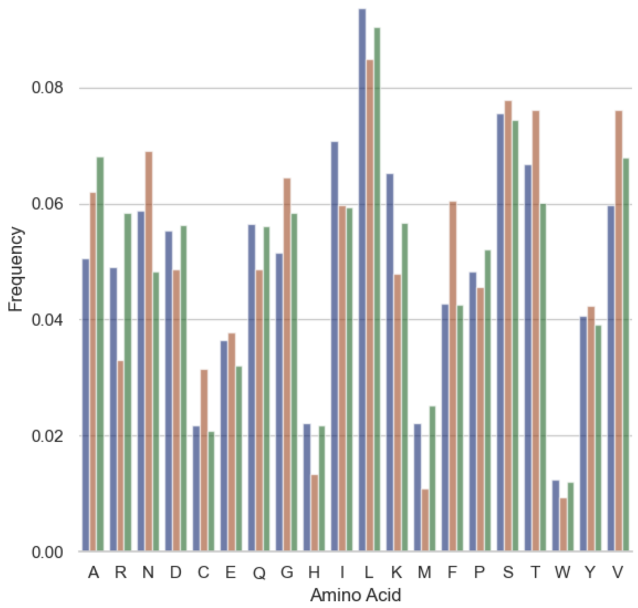
We use this script to mutate the test genome to generate a neutral network of 3 mutations. We then perform the same experiment on the "443-450 RBD Loop" [5] and the entire RBD genome. Note: We define the RBD genome as starting at position 22838 and ending at position 23516 of the Wuhan-Hu-1 genome [1]. We do not have an absolute reference for this decision. We know the RBD is ARG319-PHE541 ([4]). The S protein starts at position 21563. We confirm our RBD genome location by comparing the biopython protein translation of genome positions 2283-23516 to the amino acids in position 319-541 extracted from the spike protein in the FAST of the genome [1]. We generate all 3 mutations on the RBD Loop. We create only the first generation of the entire RBD and count its synonymous mutation. We do not complete all three generations of mutation on the entire RBD due to computing resource constraints.

B. Results

We compared the frequency of different amino acids in the spike protein to the frequencies expected of human and vertebrate-host viral proteome (figure 7). We note that while there are some differences in frequencies, the spike protein frequencies are fairly close to those expected in both human and vertebrate-host viral proteome.

We charted the same data using a line plot to get a better idea of how closely the amino acid frequencies aligned (see Figure 8). Visually, the expected frequencies for the human-host and vertebrate-host proteome are much more similar to each other than either are to the spike amino acid frequencies. This is confirmed mathematically by taking the absolute value of the difference between the frequency vectors and summing it (see Table VII). The human and vertebrate-host proteome differ in frequencies by only 0.1. The frequencies in the spike protein are closest to the vertebrate-host proteome, but that value is only 0.0019 away from the value found comparing the human-host proteome.

We compared the frequency of different amino acids in the RBD to the frequencies expected of human and vertebrate-host viral proteome (figure 9). The RBD frequencies show much greater variance when compared to the expected values for both the human and vertebrate-host viral proteome. Specifically the expected frequencies for Asparagine (N), Cystein (c), Leucine (L), Methionin (M), Phenylalanine (F), Tyrosine (Y), and Valine (V) differ by $\approx \pm 0.02$ or more. This is much more variance than the overall spike protein values which generally stay within ± 0.01 of either or both the human and vertebrate-host viral proteome expected values. Of the amino acids wherein the RBD frequencies differ greatly, only three



Source: **Color**
Human-Host: Blue
RBD: Red
Vertebrate-Host: Green

Fig. 7. Spike Protein AA Frequency Comparison - Bar Chart

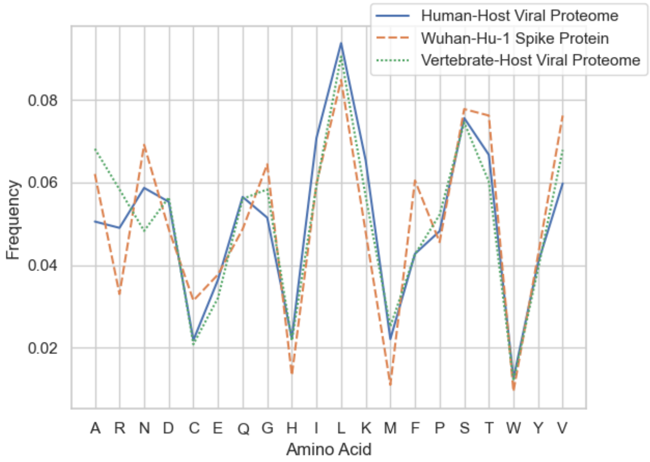
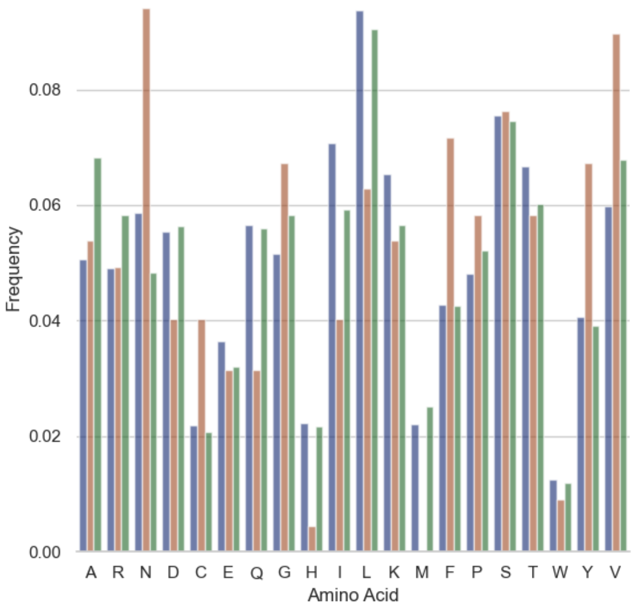


Fig. 8. Spike Protein AA Frequency Comparison - Line plot

have persistence values (likelihood of self-preservation) that are either the highest possible self-preservation probability (L and V each at 0.333) or very low (M at 0).

We first generated the expected and experimental neutral network values for the test genome of length 18, ”TACCATGGAATTACTGCG” (Phenotype: YHGITA). The results are found in Table VIII. The largest difference between the expected and actual values is 0.0214 (gap between the actual and the expected human-host calculation for the first generation).



Source: **Color**
Human-Host: Blue
RBD: Red
Vertebrate-Host: Green

Fig. 9. RBD AA Frequency Comparison

TABLE VII
TOTAL DIFFERENCE BETWEEN AMINO ACID FREQUENCIES

Difference Between	**Difference**
Human-Host & Vertebrate-Host	0.1000
Human-Host & Spike Protein	0.1869
Vertebrate-Host & Spike Protein	0.1850

Next, we calculated the expected results for the RBD 443-450 loop (Genotype: ”TCTAAGGTTGGTGGTAATTATAATTAC”; Phenotype: ”SKVGGNYNY”). The results are found in Table IX. We also took the absolute value of the difference between each of the expected percentages and the actual percentages at each generation of the neutral network (see Table X). The max difference between any of the expected values and the actual values is 0.0157. The minimum difference is 0.0074.

Finally, we examined the expected and experimental neutral network values for the first generation neutral network for the

TABLE VIII
CALCULATED AND ACTUAL RESULTS FOR THE FIRST, SECOND, AND THIRD NEUTRAL NETWORK MUTATIONS OF A TEST STRING

Calculation Type	**G1-N**	**G2-N**	**G3-N**
Expected-HH	0.2193 - 12	0.1097 - 61	0.0731 - 195
Expected-VH	0.2256 - 13	0.1128 - 68	0.0752 - 224
Expected-Spike	0.2251 - 13	0.1125 - 68	0.0750 - 223
Expected-RBD	0.2233 - 13	0.1117 - 68	0.0744 - 222
Actual	0.2407 - 13	0.1133 - 68	0.0656 - 195

TABLE IX
Calculated and Actual Results for the first, second, and third neutral network mutations of the RBD 443-450 Loop

Calculation Type	G1-N	G2-N	G3-N
Expected-HH	0.2193 - 18	0.1097 - 137	0.0731 - 657
Expected-VH	0.2256 - 19	0.1128 - 149	0.0752 - 735
Expected-Spike	0.2251 - 19	0.1125 - 149	0.0750 - 733
Expected-RBD	0.2233 - 19	0.1117 - 147	0.0744 - 718
Actual	0.2099 - 17	0.1023 - 124	0.0620 - 525

TABLE X
Difference between calculated and actual results for the first, second, and third neutral network mutations of the RBD 443-450 Loop

Actual difference between	G1-N	G2-N	G3-N
Expected-HH	0.0094	0.0074	0.0111
Expected-VH	0.0157	0.0105	0.0132
Expected-Spike	0.0152	0.0102	0.0130
Expected-RBD	0.0134	0.0094	0.0124

entire RBD genome found in Table XI. Again, we see a very strong connection between the expected and actual values. The difference between the the actual results for the RBD and the expected results for a human-host viral proteome is the minimum at only 0.0009.

TABLE XI
Calculated and Actual Results for the first, second, and third neutral network mutations of the RBD genome

Calculation Type	G1-N
Expected-HH	0.2193 - 441
Expected-VH	0.2256 - 453
Expected-Spike	0.2251 - 452
Expected-RBD	0.2233 - 449
Actual-RBD	0.2202 - 442

C. Discussion

Our experiments indicate it is possible to combine the probability of amino acid self-preservation in light of an SNS with the appropriate frequency of amino acid occurrence and arrive at a fairly accurate percentage of neutral mutations for any number of generations. This estimate would likely be higher than that actual values based on our experimental results. There are greater discrepancies (though still relatively small) with each successive generation. This method of calculating the percentage of neutral mutations per generation is most accurate when applied to an entire protein, rather than just a part of it, and when using frequencies for the appropriate host of the virus.

The high degree of correlation between the expected percentage of synonymous mutated genomes in the human-host proteome and the experimentally-derived percentage of synonymous mutated genomes in the RBD and its subsection is interesting in light of two factors. First, the fact that the experiments matched up better against the human-host proteome than they did the expected values derived for either the RBD or the Spike protein itself is odd but perhaps understandable in light of the fact that the mutations are performed on the genome so there is not a direct tie. The high correlation is also interesting in light of the fact that, while small, there is a closer connection between the amino acid frequencies in the vertebrate-host proteome and the spike frequencies than between the human-host and the spike.

Do these results say anything about the origin of SARS-CoV-2? We do not know. The fact that the amino acid frequencies more closely align with the vertebrate-host proteome lends itself to the interpretation that ”[t]he most probable explanation for the introduction of SARS-CoV-2 into humans involves zoonotic jumps from as-yet-undetermined, intermediate host animals at the Huanan market (34, 38, 39)” [8]. However, we cannot fully discount the lab leak theory especially in light of endorsement by now two U.S. intelligence services [9]. The fact that the mutation of the RBD and its sub-region more closely align with the expected values for a human-host proteome are expected, as the values we mutated were the genomic basis for a human-host proteome. However, Wuhan-Hu-1 is the first generation of SARS-CoV-2 in humans. Should we expect the results to be skewed more towards being somewhere between the two (human and vertebrate-host) expected values rather than so much closer to the human-host? Does the correlation between human-host and actual results lend itself to the interpretation that the virus had been manipulated during gain-of-function research? We do not have the answers to these questions, only the questions themselves. To begin to answer them, we would need to identify viral proteins in the first generation variants from known spillover events and analyze them in a similar fashion to what was done here. From there we could start to draw some conclusions about human versus vertebrate-host origins based on the experimental results.

Does the origin, wet market or lab, of SARS-CoV-2 really matter at this point? If we show that it is conclusively derived from the wet-markets in China, will this result in the removal of wet-markets as an institution? And, if so would that halt spillover events? Given the fact that ”[a]n estimated 60% of the 1400 species of infectious microbes known to be pathogenic in humans are transmitted by animals” [11], it is not likely that we will ever completely eliminate spillover events of this nature. If we show conclusively that there was a lab leak, does that fix anything? Again, not in our estimation. Virologists were already aware of the danger of this type of study and yet this did not halt anything. Laboratory study of viruses, to include gain-of-function research, will continue as it has in the past despite this danger. The best we could hope for in this hypothetical situation is improved laboratory containment procedures and an acknowledgement. Given the price we all, including the members of the Wuhan laboratory, have paid during this pandemic the improved laboratory procedures are likely already in place. That is the important part.

Perhaps a better use of research funds and tax-payer funded political action would be identifying ways to better communicate scientific data to the broader public and best practices for future events like the current one.

D. Future Work

While it is ”neat” that we can derive a very close approximation to the expected number of synonymous mutations in a viral genome given that we know its host, it is not enough. There are an overwhelming number of possible mutations out there and just providing the numbers does not tell us how to counteract the virus. It is a war of information. To win, you have to be able to sort out what is important and actionable. We have become deeply focused on the ”specific”. What specific mutations are possible and likely? What are the specific differences between each variation? Even the way we fight the virus is specific; we are injected with an mRNA copy of the spike protein from whatever variant was circulating at the time we received the vaccination. We have to figure out ways around the information overload and the ”original antigenic sin” of too much specificity [6]. In light of this, we believe one possible avenue for new research is to become more ”general” or ”big picture” in our approach. For example, we could analyze the spike protein using wavelet transform like the work of Trad, Fang, and Cosic [7]. If we can understand the larger view and then synthesize that with our more granular and specific view, then perhaps we can find a new way around this problem.

V. Contributions

- Alana Chigbrow completed the research, experimental work, and writing for Part III and wrote the body of the Abstract and Introduction. She also helped complete Part 1, Questions B 1-7 and Question C using the initial drafts written by Dawud.
- Jason Stewart completed the research, experimental work, and writing for Part II. He also wrote the function *src/utils.py* to check for protein mutation for Part III. Finally, he helped complete the final versions of Part 1, Questions A and B 1-7, and Question C using the initial drafts written by Dawud.
- Dawud Shakir completed the research, initial calculations and writeup for Part 1 Questions A and B 1-7.

References

[1] “Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-hu-1, co- nucleotide-NCBI,” National Center for Biotechnology Information, 18-Jul-2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore/NC$_0$45512. [Accessed: 07-Mar-2023].

[2] K.-F. Chan, S. Koukouravas, J. Y. Yeo, D. W.-S. Koh, and S. K.-E. Gan, “Probability of change in life: Amino acid changes in single nucleotide substitutions,” bioRxiv, 01-Jan-2020. [Online]. Available: https://www.biorxiv.org/content/10.1101/729756v5.full. [Accessed: 10-Mar-2023].

[3] T. K. Mohanta, A. K. Mishra, Y. K. Mohanta, and A. Al-Harrasi, “Virtual 2D mapping of the viral proteome reveals host-specific modality distribution of molecular weight and isoelectric point,” Scientific reports, 28-Oct-2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8553790/. [Accessed: 12-Mar-2023].

[4] J. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang, L. Zhang, and X. Wang, “Structure of the SARS-COV-2 spike receptor-binding domain bound to the ACE2 receptor,” Nature News, 30-Mar-2020. [Online]. Available: https://www.nature.com/articles/s41586-020-2180-5. [Accessed: 18-Mar-2023].

[5] A. Mittal, A. Khattri, and V. Verma, “Structural and antigenic variations in the spike protein of emerging SARS-COV-2 variants,” PLoS pathogens, 17-Feb-2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8853550/. [Accessed: 19-Mar-2023].

[6] A. Zhang, H. D. Stacey, and M. S. Miller, “Original Antigenic Sin: How First Exposure Shapes Lifelong Anti–Influenza Virus Immune Responses ,” American Association of Immunology, 15-Jan-2019. [Online]. Available: https://doi.org/10.4049/jimmunol.1801149. [Accessed: 19-Mar-2023].

[7] C. H. de Trad, Q. Fang, and I. Cosic, “Protein sequence comparison based on the wavelet transform approach,” Protein Engineering, Design and Selection, vol. 15, no. 3, pp. 193–203, Apr. 2002.

[8] J. E. Pekar, A. Magee, E. Parker, N. Moshiri, K. Izhikevich, J. L. Havens, K. Gangavarapu, L. M. Malpica Serrano, A. Crits-Christoph, N. L. Matteson, M. Zeller, J. I. Levy, J. C. Wang, S. Hughes, J. Lee, H. Park, M.-S. Park, K. Ching Zi Yan, R. T. Lin, M. N. Mat Isa, Y. M. Noor, T. I. Vasylyeva, R. F. Garry, E. C. Holmes, A. Rambaut, M. A. Suchard, K. G. Andersen, M. Worobey, and J. O. Wertheim, “The molecular epidemiology of multiple zoonotic origins of SARS-COV-2,” Science, vol. 377, no. 6609, pp. 960–966, 2022.

[9] ”Bloom, Escape calculator for SARS-CoV-2 RBD, https://jbloomlab.github.io/SARS2$_RBD_Ab_escape_maps/escape-calc/$

[10] C. Smith-Schoenwalder, “U.S. agencies divided over covid-19 ‘lab leak’ origin theory,” US News, 27-Feb-2023. [Online]. Available: https://www.usnews.com/news/health-news/articles/2023-02-27/u-s-agencies-divided-over-covid-19-lab-leak-origin-theory. [Accessed: 24-Mar-2023].

[11] N. Sankaran and R. A. Weiss, “Viruses: Impact on science and Society,” Encyclopedia of Virology, 01-Mar-2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7833661/. [Accessed: 24-Mar-2023].

[12] S. Duffy, “Why are RNA virus mutation rates so damn high?,” PLoS biology, 13-Aug-2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6107253/. [Accessed: 31-Mar-2023].

[13] A. Wagner, “The role of robustness in phenotypic adaptation and innovation,” Proceedings of the Royal Society B: Biological Sciences, vol. 279, no. 1732, pp. 1249–1258, 2012.

[14] D. J. Smith, A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, A. D. Osterhaus, and R. A. Fouchier, “Mapping the antigenic and genetic evolution of influenza virus,” Science, vol. 305, no. 5682, pp. 371–376, 2004.

[15] L. Sompayrac, “Lecture 1: An Overview,” in *How the immune system works*, Hoboken, NJ: Wiley-Blackwell, 2019, pp. 1–12.

[16] “Coronavirus disease (covid-19): How is it transmitted?,” World Health Organization, 23-Dec-2021. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted?gclid=Cj0KCQjwz6ShBhCMARIsAH9A0qXvHIEjp$_UQOd6pD071qIxN$ [Accessed: 02-Apr-2023].

[17] Sam Wilks, ”Racmacs computer code,” [Online]. Available: https://acorg.github.io/Racmacs/index.html. [Accessed: 29-Mar-2023]

[18] Sam Wilks, ”Making an antigenic map from titer data,” [Online]. Available: https://acorg.github.io/Racmacs/articles/making-a-map-from-scratch.html. [Accessed: 29-Mar-2023]

[19] Sam Wilks, ”An introduction to antigenic cartography,” [Online]. Available: https://acorg.github.io/Racmacs/articles/intro-to-antigenic-cartography.html. [Accessed: 29-Mar-2023]

[20] Melanie Moses, ”Complex Adaptive Systems,” [Lecture] Spring 2023 at the University of New Mexico