Insights into Viral Evolution Dynamics: Mapping Mutations and Understanding the Emergence of the Omicron Variant in COVID-19

1st Tristan Larkin
University of New Mexico, Department of Computer Science
Albuquerque, United States of America
trlarkin@unm.edu

2nd Jack Wickstrom
University of New Mexico, Department of Computer Science
Albuquerque, United States of America
jwickstrom@unm.edu

***Abstract*—The COVID-19 virus pandemic has permanently altered the understanding of viral biology, particularly with regard to mutation mapping and and related techniques. This paper explores a number of methods to map viral mutations and analyze their implications. We begin with an examination of the antigenic properties of the SARS-CoV-2 virus, by performing mutation calculations to understand the genetic diversity and antigenic/functional consequences of mutations. Next, we construct a neutral network to visualize the genetic neighborhood of the virus and investigate its implications for viral evolution and robustness. Additionally, we create an antigenic map generated from mutation data converted into titer data. This provides insights into the antigenic relationships between different viral strains. Finally, we investigate the strange circumstances surrounding the emergence of the Omicron variant and assess possible scenarios for its origin using scientific literature and analysis from an argument based on neutral networks. We find that the development of the Omicron variant in the same population as the Delta variant is very unlikely.**

I. Introduction

COVID-19 caused a lot of changes to the way people do things, and something it changed is how biologists understand viruses. Thanks to the large amount of data collected on the Sars-Cov2 virus, we have been able to learn a lot about viruses. This paper will discuss a few ways we can map viral mutations and quantify their effect on the virus.

There are billions of virions inside the body during an infection, and many of them are replicating, with a chance of mutation. While most mutations have negative effects on the virus [1], the ability to mutate is what makes viruses like SARS-CoV-2 so dangerous, as they can continue to evolve to fight our attempts to kill it off. We are going to strictly consider mutations where one amino acid is swapped with another amino acid due to a single nucleotide change in the genetic code during replication. There are many nucleotide mutations that do not change how the proteins form, which is a form of robustness. However, even if the protein changes many mutations are essentially neutral, and do not meaningfully change the fitness of the virus.

Understanding how SARS-CoV-2 mutates to become a more fit virus can help us learn how to predict its evolution. Computer scientists have also based many computational models on biological systems, and this could be a path to developing new computational methods or analysis techniques.

II. Mutation Calculations

Question 1

How many different genomes are 1 nucleotide mutation away from the RBD of the original strain?

The RBD has 194 amino acids, and since each amino acid is encoded by 3 nucleotides, it has $194 * 3 = 582$ nucleotides. Since there are 4 possible nucleotides, each nucleotide can mutate into 3 different variations. This means there are $582 * 3 = 1746$ different genomes 1 nucleotide away.

a. **What fraction of those mutations are silent?** First, we took the original string of 582 nucleotides and generated all the 1746 mutations. Then, for each of these mutations, we converted them into the corresponding string of amino acids. If this string matched the string of original amino acids in the RBD, then the mutation was counted as silent. This resulted in 384, or 22.0% silent mutations.

b. **What fraction of those mutations are antigenically neutral?** To answer this question, we used Jesse Bloom's COVID antibody escape calculator [2]. This calculator contains a helpful Python module named `escapecalculator.py`.
This module allows users to input a list of amino acids edited in RBD and returns the fraction of antibodies that are still binding. If more than 99% of the antibodies were still binding, we considered the mutation to be antigenically neutral. This procedure resulted in 952, or 54.5% antigenically neutral mutations.

c. **What fraction of those mutations are functionally neutral?** To answer this question, we used Jesse Bloom's COVID spike amino-acid fitness calculator [3]. The code repository contains a CSV file named `aa_fitness.csv`. This file contains the fitness change resulting from changing one amino acid to another in the RBD. However, this file isn't exhaustive, i.e. it doesn't contain every possible mutation. Therefore, for each mutation, we simply took the average fitness of mutations across that particular site.

If this average fitness change had a magnitude less than 1, then we considered it functionally neutral. This procedure resulted in 1066, or 61.0% functionally neutral mutations.

d. **How many different genomes are 3 nucleotide mutations away** With more than 1 mutation, the problem becomes combinatoric. The formula is $\binom{n}{k} * 3^k$, with k = 3 and n = 1746. Solving this results in 882550620 mutations.

Question 2

How many different genomes are 1 amino acid mutation away from the RBD of the original strain? The RBD has 194 amino acids, each of which can mutate to 19 different amino acids. Therefore, the RBD has $194*19 = 3686$ different genomes 1 amino acid mutation away.

a. **What fraction of those mutations are silent?** Since every mutation involves the change of an amino acid, none of them are silent.

b. **What fraction of those mutations are antigenically neutral?** We used the same strategy as in question 1: generate all the mutations then call them in the Python module. This resulted in 1539, or 41.8% antigenically neutral mutations.

c. **What fraction of those mutations are functionally neutral?** Using the CSV file from the fitness calculator, we found 1862 or 50.5% functionally neutral mutations.

d. **How many different genomes are 3 nucleotide mutations away** We can use a similar combinatorial formula from question 1: $\binom{n}{k} * 19^k$. Note that 3^k was replaced by 19^k, since there are 19 different acids to mutate to. In this scenario, n = 194 and k = 3. This results in 8218069696 genomes that are 3 amino acid mutations away.

All the code used to compute these calculations can be found in our repository [4], in the file `part1.py`.

	a	b	c	d
Q1	22.0%	54.5%	61.0%	882550620
Q2	0%	41.8%	50.5%	218069696

III. Neutral Network

Neutral Network Background

For viruses to increase fitness they have to evolutionarily "search" for mutations that are beneficial to the virus. Generally, "weakly [harmful] mutations are more abundant than neutral mutations" [1]. There are billions of virions inside an infected host, so a virus can spare some poor mutations if it means that some lead toward a helpful mutation. In the context of a virus's genetic code, a neutral network is a graph where the vertices are sequences of genetic code for the virus, connected by edges representing the mutation that takes the virus from one vertex to another. Specifically, neutral networks are these graphs where all the edges produce little to no phenotypic changes, and by extension do not change the fitness of the virus (see Fig 1).

Two interesting things that neutral networks show are how a virus can explore the space of mutations and how certain traits can become robust to change. As mentioned, when a virus mutates it is generally going to become less fit and die off before it is able to reproduce a lot, losing out to the more fit viruses. However, it is beneficial to have an active population that is diverse, essentially not putting all your eggs in one basket. Neutral networks allow viruses to have a level of diversity in their population by having many equally fit, but genetically different, virions that can actively compete with each other. In Fig. 1 there are many white and slightly blue bubbles that are likely to survive together in a host each searching for good mutations in a different space, speeding up that search exponentially from just having a single dominant genetic virus.

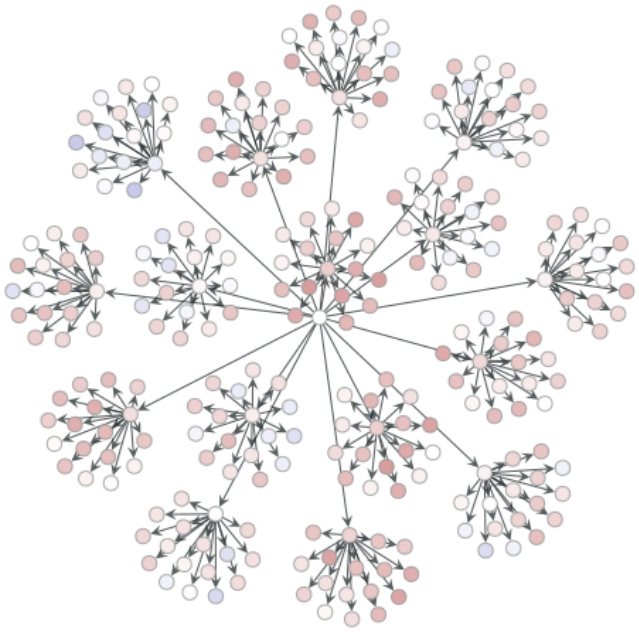

Fig. 1. Neutral network graph based on the data in [3]. Any edge represents a single amino acid change that affects fitness up to ± 0.1 fitness. The vertices are colored based on their relative fitness to the center starting node, where red is worse and blue is better. As expected there are more mutations that negatively impact the fitness than positively, 82% bad in this image. There exist cases where both mutations are negative, but some show that while the first mutation was poor, the second ended up bringing the overall fitness above the starting fitness. *Note that all mutations in this graph are small (hence this entire graph is a neutral fitness network) and the length of the edge does not have a meaning.*

On the other end of the spectrum, neutral networks can show how robust a virus is to change. Certain virus configurations are less volatile and will be able to have stray mutations that do not really change the trait. When looking at a neutral network that only includes a subset of the genetic code that encodes a vital part of a virus or other life form it will be expected for most major mutations to be multiple mutations away, so as to not risk mutating in a way that devastates the organism.

Neutral Network Creation

Since a neutral network is based on the fitness of potential mutations, we opted to use the same CSV file from Jesse

Bloom's fitness calculator that contains the fitness change of various amino acid mutations along the spike protein. Additionally, we also opted for the same threshold of neutrality from earlier: ± 1.

To create the neutral network, we started by filtering the fitness data down to only amino acids within the spike protein, and to those that were in our neutrality threshold. Then, we created a root node, shown in the center of figure 1, which represents the base unmutated virus. From this root node, we chose 16 variants 1 mutation away that were in our neutrality threshold, then attached them as outgoing nodes from the root node.

Unfortunately, as far as we could tell, calculator [3] contains no way to assess the fitness changes for a novel, mutated virus. After 1 mutation, the numbers of fitness changes for a subsequent amino acid mutation were no longer accurate, since they are referring to the base unmutated virus. However, we imagine that the difference in fitness change between two viruses that are one amino acid apart is very small, so we used the same fitness values from the CSV file for the second mutation.

Given the assumption from the last paragraph, we constructed the next layer of mutations as follows: for each of the 16 earlier mutations, we created another 16 mutations within the neutrality threshold. However, these mutations were somewhat different from the first layer, because we computed the fitness by adding together the fitness from the first and second mutation. We only considered the second layer mutation neutral if this "net" fitness was still below our neutrality threshold. Just as in the first layer, we set these 16 novel mutations as outgoing nodes from the mutation they were created from. All the code to create the neutral network can be found in the file `part2.py` from our repository, [4].

IV. Antigenic Map

After building the neutral network, we used it to generate many different mutations and visualized these mutations using an antigenic map. An antigenic map is a graphical representation used in immunology and virology to visualize the relationships between different strains or variants of a virus based on their antigenic properties.

A. Antigenic Map Creation

To create the antigenic map, we started by generating 270 mutations using our neutral network. For each mutation, we first perform a random walk to an outer edge on the neutral network, giving us a neutral mutation two amino acids away from the original. Then, from this mutation, we randomly choose 1 two 3 more amino acids to mutate. These mutations aren't limited to neutral mutations, so these mutations might have a dramatic effect on the fitness of the mutation.

This process leaves us with a mutation that is 3 to 5 amino acids away from the original spike protein, which may or may not be neutrally mutated. After generating these mutations, we must process them into a titer data format that can be used to generate an antigenic map.

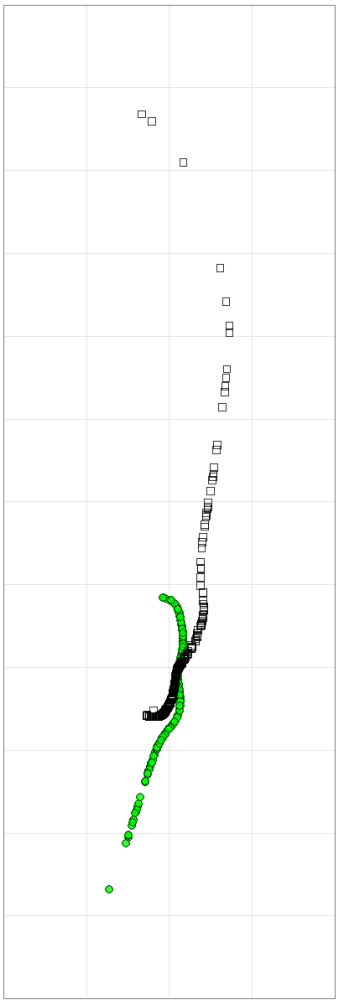

Fig. 2. An antigenic map of 270 different variations of the COVID spike protein. The strains are colored circles and antisera are uncolored open squares. "The spacing between grid lines is 1 unit of antigenic distance—corresponding to a twofold dilution of antiserum in the HI assay. Two units correspond to fourfold dilution, three units to eightfold dilution, and so on." [5]

To generate the titer data, we first begin by computing the escape amount for each mutation using Bloom's calculator [2]. Then, we compute antigenic distance with the formula $\text{distance}(ij) = |1 - \text{escape}_i/\text{escape}_j|$. Then, we can use this distance value to generate titer data with the formula: $\text{titer}(ij) = 2^{10-\text{distance}(ij)}$

Once titer data has been generated, we write it to a CSV file, then feed that file to the RACMACS tool written in R to generate the final antigenic map [5].

The code used to generate the CSV titer data can be found in our repo at [4] in the file `part2.py`. The code used to consume this file and actually generate the antigenic map can be found in `part3.r`.

B. Antigenic Map Discussion

Figure 2 shows the map resulting from the methodology outlined in IV-A. In a typical antigenic map, one can see different virus strains condensed into clusters throughout the

map. However, our map doesn't appear this way. This is because the antigenic distances between the mutations are much smaller than typical titer data used to generate an antigenic map.

One interesting observation of the antigenic map is that almost all the variations seem to lie along the same line, which shows that they are closely related.

Using the Bloom fitness calculator [3], we also calculated the average fitness of the strains shown in this antigenic map. The average fitness of the mutations was -3.97. This is unsurprising since most mutations in the wild are not advantageous. Because these mutations are on average unfit, it is unlikely to see them in the wild. If mutations on average improved fitness, then viruses would evolve extremely quickly, making countermeasures against them very difficult.

V. Part 4

Despite all the data that exists on the SARS-CoV-2 variants, especially from 2020 to 2022, there is an interesting mystery that persists: how did the Omicron variant come to be? Unlike the prior dominant variants, Omicron does not appear to be a descendent of another dominant variant before it ([6], [7]). There are three major theories for how Omicron's suddenly appeared:

1) It mutated in a certain population, but was either not tested for in that population, or maintained a low enough percentage to avoid being noticed.
2) It developed in an immunocompromised person, where the virus kept surviving for months or years, continuously mutating as a poor immune system failed to fight it off completely.
3) After developing in a non-human host, it was transmitted to a human where it spread like any other variant.

Among these options, 1) appears the least likely. Avoiding detection would be difficult since Omicron is over 50 mutations from the original Wuhan strand of SARS-CoV-2 [8]. Based on research regarding how multiple variants in the same location come to be and how viruses are expected to evolve to best survive, we want to find evidence to help determine if Omicron could have developed along with the same variants that gave Delta.

If Omicron developed along with the other variants but was able to remain hidden, then it would have to be possible for multiple variants to exist in the same community. First, it is necessary to determine whether or not contracting a Delta variant protects against Omicron. Paper [10] suggests that contacting Delta provides protection against Omicron for a considerable period. It is still possible that the two viruses were living together. This is similar to behavior in the neutral network, where there might exist multiple equal fitness versions of the virus where neither is going to kill off the other via competitive exclusion (this point is further explored in Fig. 3). The way that COVID-19 tends to have one dominant strand is supported mathematically by Wang [11]. However, another paper that looks at the rate of viral production in the body and the rate of infection shows that two viruses can coexist if the

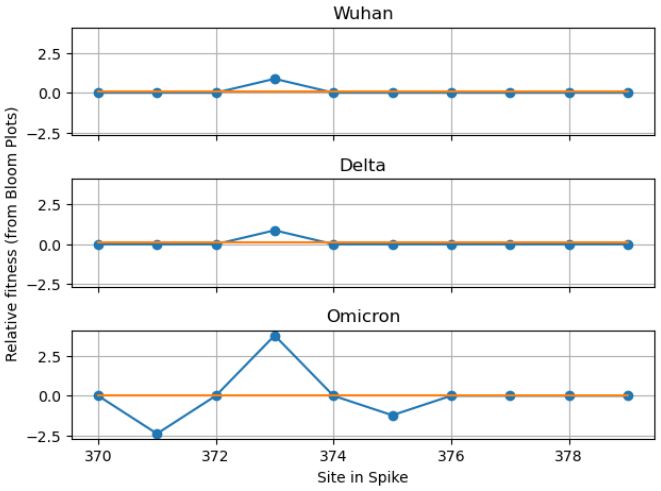

Fig. 3. Data presented from [3]. These plots show the sites 370 to 379 in the SARS-CoV-2 spike protein. Looking at each individual site and the affected fitness, these are not individually neutral mutations, and would not be able to be connected on a neutral network. However, the overall fitness change (in orange) is similar. This supports that Omicron did not evolve from Delta, but since they would not even connect on a neutral network it shows that it is unlikely they developed together in the same population since these large changes in fitness would have likely caused one to become dominant before coming back to being neutral. One interesting fast is that the delta variant did not change anything in this region, while Omicron had multiple neaby mutations.

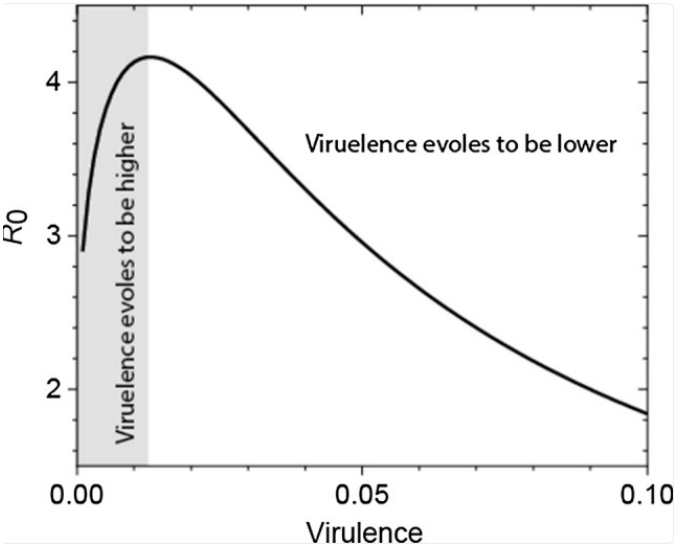

Fig. 4. Reproduced from [9]. A classical view of viruses where the danger posed by a virus decreases when R_0 decreases. This is not a closed question for COVID-19 since it is not entirely clear what the relationship between the number of virions produced and the physical impacts of the infection on the infected individual is. This is one of many assumptions that were made before COVID-19 that are not entirely accurate.

rate of infection is larger in one, and the rate of production in the body is larger in the other [12], but that paper is older than COVID and might not be as accurate due to the amount we have learned since the COVID-19 pandemic. Omicron was a very infectious disease that tended to not cause as strong of symptoms, which might lead us to believe that it does not produce as many virions in the body, but for COVID it is shown that producing more virions does not necessarily make infected more sick [9]. Overall the evidence here points to Omicron not developing along with the Delta variants.

While the origins for Omicron were quite mysterious, it did follow the expected progression that scientists imagined for COVID: becoming more infectious and less dangerous. Kun et. al. explore whether or not these are reasonable assumptions to make about how COVID will evolve in the future [9]. While Omicron was less deadly than previous variants, that was not necessarily because the virus was less dangerous. By the time Omicron was a major force, much of the population was vaccinated, many people changed behaviors to increase their safety, and most people had contracted some variant of COVID-19. Their paper concludes that many assumptions that are used to predict virus behavior do not apply to COVID-19.

Here we present a short look at a single possibility to answer the question: where did Omicron come from? Based on the literature available, Omicron developing along with Delta (i.e. in the same population) appears to be very unlikely. Given that COVID has broken some of the historical assumptions held pre-pandemic, it is hard to predict where the next variant will come from and how it will act, even with the extensive data we have on past variants. This shows that even reducing the scope of a question to a small, well-documented piece of the COVID puzzle it is still a difficult topic and warrants the many researchers working to understand the force of nature that is viral diseases.

VI. Contribution Statement

Jack Wickstrom:

- Co-wrote the code for the neutral network generation
- Wrote code and paper section for II
- Wrote code and paper section for IV
- Contributed to sections I, III. and V

Tristan Larkin:

- Co-wrote the code for the neutral network generation
- Contributed to sections I, III. and V
- Researched for section V

VII. Acknowledgments

Used Chat GPT for brainstorming ideas for part 4, the definition of an antigenic map, and the title. A large portion of this project uses the data and calculators created by Jesse Bloom.

References

[1] Andreas Wagner. The role of robustness in phenotypic adaptation and innovation. *The Royal Society*, 2012.

[2] Jesse Bloom. Escape calculator for the sars-cov-2 rbd. https://github.com/jbloomlab/SARS2-RBD-escape-calc/, 2024.

[3] Jesse Bloom. Fitness effects of sars-cov-2 amino-acid mutations estimated from observed versus expected mutation counts. https://github.com/jbloomlab/SARS2-mut-fitness, 2024.

[4] Jack Wickstrom and Tristan Larkin. Neutral_networks. https://lobogit.unm.edu/jwickstrom/neutral_networks, 2024.

[5] Sam Wilks. Racmacs. https://github.com/acorg/Racmacs/, 2024.

[6] Lok Bahadur Shrestha, Charles Foster, William Rawlinson, Nicodemus Tedla, and Rowena A. Bull. Evolution of the sars-cov-2 omicron variants ba.1 to ba.5: Implications for immune escape and transmission. *Reviews in Medical Virology*, 32(5):e2381, 2022.

[7] Gao G. F. Wang Q. Du, P. The mysterious origins of the omicron variant of sars-cov-2. *Innovation*, 2022.

[8] Yamin Sun, Wenchao Lin, Wei Dong, and Jianguo Xu. Origin and evolutionary analysis of the sars-cov-2 omicron variant. *Journal of Biosafety and Biosecurity*, 4(1):33–37, 2022.

[9] Ádám Kun and et al. Do pathogens always evolve to be less virulent? the virulence-transmission trade-off in light of the covid-19 pandemic. *Biologia Futura*, 74(1-2):69–80, 2023.

[10] Mariana Baz, Nivedita Deshpande, Caroline Mackenzie-Kludas, Florence Mordant, Danielle Anderson, and Kanta Subbarao. Sars-cov-2 omicron ba.1 challenge after ancestral or delta infection in mice. *Emerging Infectious Diseases*, 28(11):2352–2355, 2022.

[11] Wei Wang. Competitive exclusion of two viral strains of covid-19. *Infectious Disease Modelling*, 7(4):637–644, 2022.

[12] M.N. Burattini, F.A.B. Coutinho, and E. Massad. Viral evolution and the competitive exclusion principle. *Bioscience Hypotheses*, 1(3):168–171, 2008.