SARS-Cov-2 New Variants Prediction & Analysis

Yasmin Heimann, hyasmin, yasmin,heimann@mail.huji.ac.il **/ Shon Cohen**, seanco, Shon.cohen@mail.huji.ac.il **/ Tom Eliassy**, tom_e91, Tom.Eliassy@mail.huji.ac.il

Preface: Problem & Background

Covid-19 Pandemic and Project Definition

The novel Covid-19 disease, caused by Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-Cov-2), is a highly contagious disease that emerged in 2019 in Wuhan, China, and was declared as a pandemic by the World Health Organization.

The disease rapidly spreads from one person to another, as it has a high transmission and infection rate. That leads to an enormous number of virus replication in the host cells (such as human cells). Each replication of the virus has the potential to create mutations which are the result of random mistakes in the replication process. These mutations change the virus's RNA strain (i.e., it's genetic material), and can emerge a new covid-19 variant which has an evolutionary advantage on previous strains. Each new variant that has the potential to be fixed in the population is a variant of concern that should be investigated, as it might be more transmissible or dangerous to it's hosts - which in our concern are humans.

In our project, we aim to investigate new artificially generated mutations in the spike protein of the SARS-Cov-2 which resides on its surface and binds to the human ACE-2 receptor.

Biology Background

Biologically, SARS-Cov-2 enters our cells using the ACE-2 receptor, which is a protein on the surface of many cell types including the lungs, heart, blood vessels, kidneys, liver and gastrointestinal tract.

The SARS-CoV-2 virus uses its spike-like protein on its surface, and more specifically its **RBD** (receptor binding domain), to bind to the ACE2 receptor on human cells. The ACE-2 receptor acts as a cellular doorway for the virus that causes COVID-19.

After the virus enters the cell cytoplasm, it uses the cell's resources to replicate itself - causing a disease in the human body.

A better bond that is formed between the spike and ACE-2, creates an easier ability for the SARS-Cov-2 to infect human cells.

Problem Description

As new variants in this pandemic occur every few months, it is a high issue of concern.

We would like to explore and evaluate the risk of possibly new mutated SARS-Cov-2 variants, which are mutated in the RBD part that is crucial for the binding of SARS-Cov-2 to human cells.

In our project, we will focus on mutations that occur in the spike protein and use syntheticaly generated variants of the S1 domain of the Spike protein (i.e., RBD).

We will encounter the following research questions:

- How can we find new relevant SARS-Cov-2 variants?
- Which mutated strains of the SARS-Cov-2 virus may be a variant of concern?
  - What areas of the virus is more significant?
  - Which mutations might produce better interaction with the human receptor ACE-2 and thus be more dangerous?

1 Project Data

We have used 2 different data sources in our project.

Original SARS-Cov-2 strains of the 7 variants collected from the WHO. The Wuhan strain was used to generate new mutated sequences, whereas the other 6 strains where used in the project evaluation. The data is formatted in a FASTA file format - a text-based format for representing proteins or nucletides sequences. A sequence in FASTA format begins with a single-line description (i.e., name and accession number), followed by lines of sequence data (single-letter codes).

Number of records	Size	Source
7	few KB (a fasta file with $\sim 7 \cdot 1250$ characters)	The WHO site (world health oraganization)

Synthetic data: Given the original Wuhan strain that was common in March 2020, we took the sequence of its RBD part of the spike (the part that binds to the ACE-2, 200 AA in positions 597 to 792 in the spike that has 1200 positions of amino acids). Then, we generated new RBD strains, using random mutations in the sequence (i.e., changing amino acid A to amino acid B). Considering computational challenges, we have decieded to create new sequences with 1 to 5 mutations. That is because we saw that all the variants besides the omicron had up to 5 mutations in the RBD, and each run on a new sequence is time consuming, which makes it impossible to run a full analysis in a manner of less than weeks to months. We have generated 10K new mutated sequences.

Number of records	Size	Source
10K	~60MB	Synthetic self-generated by our pipeline

2 Solution

We aim to explore variants by generating new mutations in the RBD (receptor binding domain) of the spike protein, and evaluate them by determining a binding score to the human ACE-2 receptor using state-of-the-art computational biology and data science algorithms.

For that, we aim to find and explore new spike-protein variants - mutated in the RBD part - and that have binding scores that show **better binding compared to the Wuhan original strain**.

This will mean that these variants will probably interact better with the ACE-2 receptor, and thus be **more infectious or dangerous**.

The workflow we have executed is aiming to determine how the binding likelihood of each new generated sequence to the ACE2 receptor. A lower binding score - means greater free energy, that will imply on a better bond (a comfortable state to be in energetically). A higher score means that the proteins reject each other and are less likely to bond.

From the data of the binding strength we can conclude wether the generated mutations can be a possible variant of concern or not.

2.1 Project Workflow

For finding the best new variants, we have executed the following workflow.

1. **Generate new mutations** in the RBD of the Wuhan strain sequence (A 1D sequence of amino acid letters representation determining the order of the protein's amino acid).
   (a) We took the RBD sequence and randomised up to 5 mutations in each generated sequence (i.e., chose random position and changed the amino acid randomly).
   (b) **Relevant code** for this part is the **mutation_generator.py** file.
2. **Finding 3D Structure of the generated strains using SCRWL4,** i.e., fold the sequence to it's 3D structure (the representation of the protein in the cell).
   (a) From the PDB (Protein Database) we extracted the solved 3D structure of the wuhan SARS-Cov-2 strain bonded to the human ACE-2 receptor, found here https://www.rcsb.org/structure/6LZG. This is an accurate 3D model of the wuhan strain spike-s1 domain (RBD) bound to the ACE-2 human receptor, that was reported in the following **paper**:
      i. https://www.sciencedirect.com/science/article/pii/S009286742030338X?via%3Dihub
   (b) For each generated strain sequence, we have predicted its new 3D structure with its point mutations, based on the known structure using SCRWL4 - a protein side-chain conformations prediction software.
      i. **paper**: http://dunbrack.fccc.edu/SCWRL3.php/SCWRL4Paper.pdf
      ii. **software site:** http://dunbrack.fccc.edu/SCWRL3.php/
   (c) The SCRWL method uses the known structure we found in (a) to predict the 3D representation of the generated sequences with the 1-5 mutations in it. The outcome of SCRWL represents more accurately the linkage between the RBD and the ACE-2, as it is based on a well-known and accurate model. Thus, building the whole new structure from scratch with alphaFold for 1-5 changes will be a source for more mistake, as in SCRWL we use the same known and solved structure with prediction of only few point mutations (1-5 changes in the strain out of 200 positions in the strain).
   (d) **Relevant code** for this part is in the file **scwrl_model_generator.py** that is pre-processing the data and running the software using a path from Dr. Dina Schneidman's lab '~dina/software/progs/scwrl4/Scwrl4'. The path uses the installed software in cs computers under the cs licence from Dina's lab.
3. **Evaluate the binding** score of the mutated sequence 3D structure and ACE-2 3D structure.

(a) We used the **SOAP** software for calculating the binding score of each 3D representation of mutated RBD and ACE-2. SOAP scoring method is based on calculations of the energy between amino acids participating in the bond, where a lower score means the structure is more stable and more likely to bond easily and create the RBD-ACE2 complex which means the virus will enter the cell more easily.

(b) We used various visualisations such as scores histograms, clusterings and the PyMol software that we used to look on the 3D representation of the sequences with the best score and analyze the change in binding by looking at the structures.

(c) **Relevant code** for this part is in the files:

i. **scwrl_model_generator.py** that is pre-processing the data and running the software using a pearl script from Dr. Dina Schneidman's lab called 'run_SOAP.pl'. The script uses the installed software in cs computers under the cs licence from Dina's lab.

ii. **data_analyzer.py** that is plotting statistics on the scores of the generated sequences.

4. **Validation**: we used the 6 variants that we know has emerged from the Wuhan strain, and compared the mutations in the sequences with the best scores we found (i.e., lowest scores - more energetically favored). **We validated our method** by analysing whether the generated mutations that met with our criteria (best SOAP score), has been found in any of the known variants, or in the same region.

(a) We used MSA (multiple sequence alignment, JALVIEW software) that uses the editing distance we learned in class to analyze changes between the best generated sequences and the real SARS-Cov-2 variants.

(b) Clustering the best and worst results as a phylogenetic tree (a representation of an evolutionary clustering) to see if we found sequences that are evolutionary close to the real variants that emerged from the Wuhan strain.

5. **Conclusions** on the findings using score and sequences visualisations and PyMol software for the 3D structures visualisations.

(a) In the results section we will discuss our findings after looking at the best and worst generated sequences scores.

The figures bellow show:

- A scheme of the representation from a 1D sequence of amino acids to the protein 3D structure and to the binding of two proteins, where the score is based on the 2 proteins binding.
- A scheme of the full workflos of the project

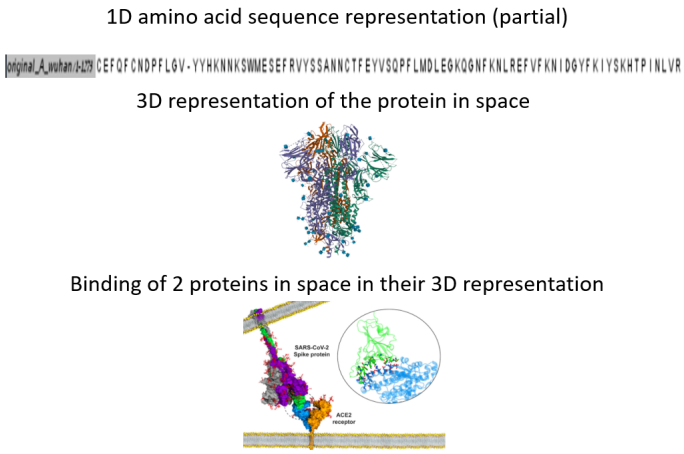

Figure 1: The flow of proteins representation in an amino acid sequence, a 3D representation, and a binding of two proteins

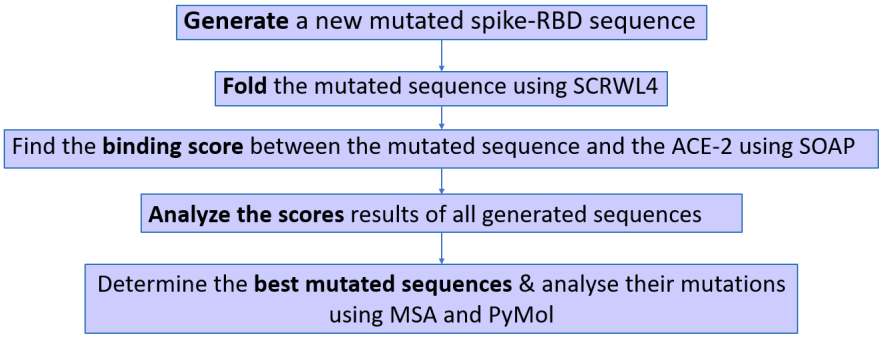

Figure 2: A scheme of the project workflow

3 Evalutaions

For evaluating the score results of our project workflow, we have extracted multiple statistics and data visualizations, to get insights and conclusions on the results.

3.1 Evaluation criteria

We use the SOAP algorithm as the scores prediction model, which is a well-known and common algorithm in the field of protein-protein interactions. The algorithm is based on bio-physical knowledge of experts in the field, and calculates the potential energy anf forces of a given 3D structure. We can learn from this score, when used in comparison with multiple stuctures' scores.

Our **evaluation criteria** is based on combining the best SOAP score of the generated sequences with the real variants data.

A **success** of our method will be granted if using our score criteria (the SOAP score) will detect variants that are similar in their properties to the real variants.

In order to make sure that our findings are not by chance, and that we cover enough sequences that are significantly different, we extracted the distribution of their scores (Figure 3) and made sure they cover the scores distribution of the real variants (Table 1).

3.2 Setup

Our experiments are done using computational methods only, on known and sythetic biological data.

We ran all of our algorithms on the cs-computers using the cluster, to make it computationaly possible.

As for avoiding bias, we mentioned in section 3.1 that we have validated that the generated sequences are broadly distributed, and form a gaussian of scores that it's center covers the real variants distribution.

3.3 Results & Visualisations

In the following results and visualisation steps, we chose to demonstrate our results using multiple methods and visualisations that shows the interaction of the sequences and their scores.

Scores Histogram

We wanted to see the distibution of scores and validate that our sequence generation step was significantly broad.

The following histogram shows the binding scores of all the generates sequences, with the ACE-2 receptor.

We can see that most of the generates sequences are around the same score of the baseline - the Wuhan strain sequence, which makes sense and validates our workflow mechanism.

Moreover, we can see that we still have some interesting sequences that caused higher and lower binding score.

In our next evaluations, we will focus on the best and worst sequences' score.

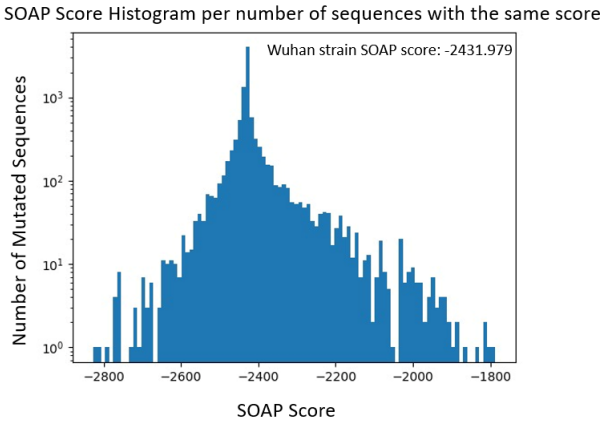

Figure 3: All generated sequences SOAP score

Sequences Dimensionality Reduction Analysis

We wanted to understand whether there is a correlation between the sequences (by using their embeddings which we extracted using SeqVec, https://github.com/rostlab/SeqVec) and the SOAP binding score.

We can see in the following figure, that the embedding does not catch any correlation between the scores, which can be a result of the difficulty to differently represent sequences with 1-5 changes in the positions. Thus, it seems that SeqVec lacked the sensitivity required for point mutation changes, and no further understanding can be made from this visualisation.

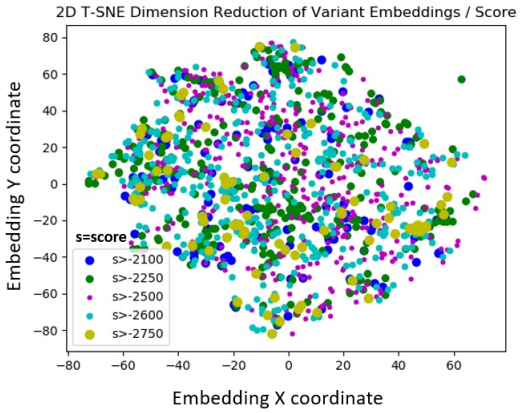

Figure 4: T-SNE on the generated sequences embeddings in 2D and their score (by color)

Number if Mutations Diversity

As we generated up to 5 mutations, we wanted to explore whether there was a diversity in the sequences with 1 mutations and more.

For that, we extracted the scores distribution per number of mutations in the figure below.

We can see that the more mutations we have, the higher the variance.

Moreover, **interesting to see is that even with only 1 mutation in the RBD, we can make a significant change to the binding score with ACE-2.**

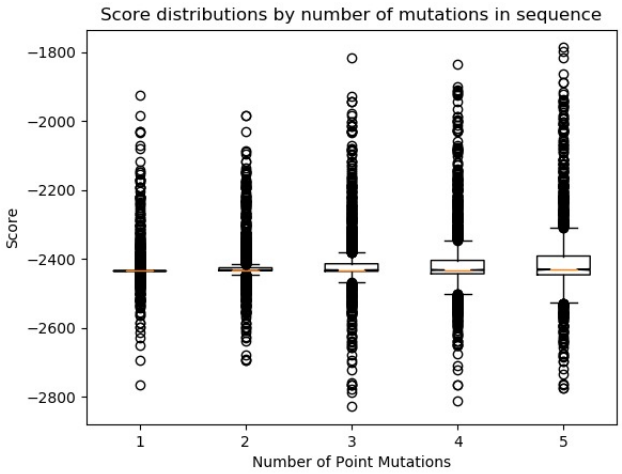

Figure 5: Sequences scores distribution with variance box, for each group with the same number of mutations

Mean Score per Mutated Position

To continue and validate our conclusion on that, we have extracted the mean score of sequences with the same mutation in each position in the sequence.

The figure below shows the sequence position in the x axis and the mean score of all sequences that had mutations in this position.

We see that the variance is greater from 1 to 5, but more interestingly we see that the location of the **most significant mutations are in the same region of positions 75, 125 and 150-175** of the RBD. If we look at the full spike sequence, the positions are: 672, 722 and 747-772. (Remember: we took positions 597-792 which are the RBD, receptor binding domain, that is crucial for ACE-2 binding).

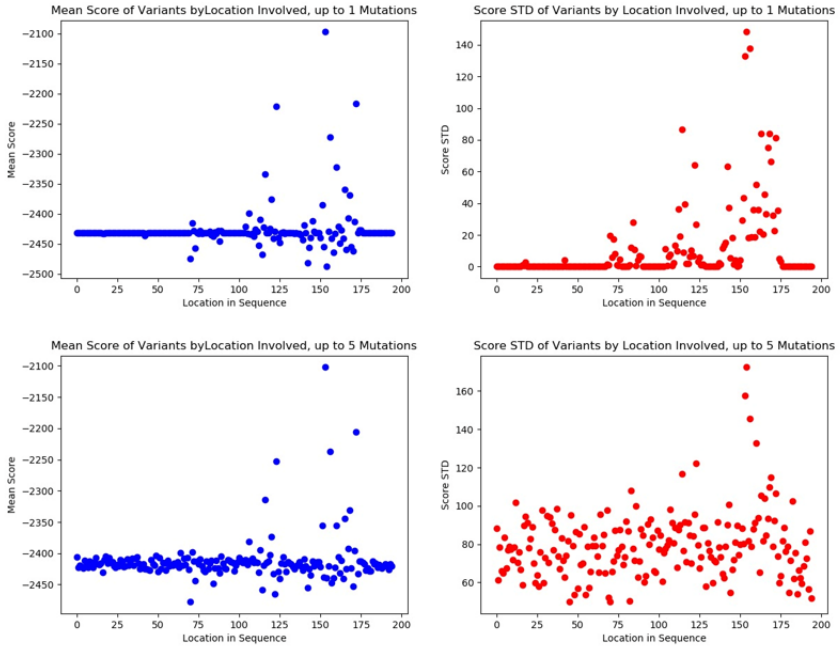

Figure 6: The scores mean (left) and variance (right) per mutated position of the groups of sequences with the same number of mutation 1 or 5

Single Mutations Analysis

To understand if our evaluation criteria was met, we first extracted the generated sequences with the best scores, which are around -2800.

We found some interesting similarities between the generated sequences :

1. Position number 750 has changed from amino acid N (asparagine) to F (phenylalanine) in **most of the 10 best scores**.

(a) Asparagine (N) is a polar amino acid, where Phenylalanine is aromatic that can create hydrogen bonds and interact with other aromatic amino acids.

2. Position number 749 has changed from amino acid F (phenylalanine) to K (lysine) in **most of the 10 worst scores**.

   (a) Phenylalanine (F) is a non-charged, aromatic amino acid that tends to create hydrogen bonds with other aromatic amino acids, where Lysing (K) is a positively charge amino acid. A change from charged to non-charges in biology is significant in the bonds and structure the amino acid can create.

In the following image taken from the PyMol software, which is used to visualize the 3D structure of proteins, we can see that this region around 750 where this mutations happened, is right on the binding area of the two proteins, spike RBD domain and ACE-2.

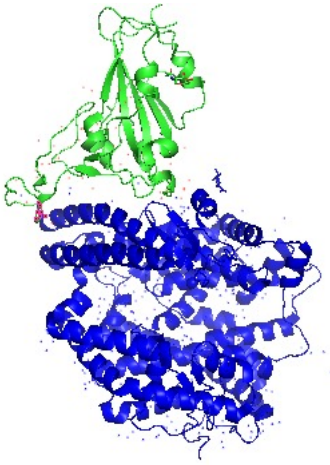

Figure 7: The Wuhan strain RBD sequence in green, ACE-2 protein in blue. The red part is the region around position 750.

Next, we zoomed in at the structure, showing it's side chains (which indicates the specificity of the amino acid), excatly at the 750 and 749 locations.

- We saw that the 750 mutation created a bond between the spike and the ACE-2 that weas'nt there before in the original sequence.
- Moreover, changing the 750 position to be aromatic, could interact with the aromatic side chain placed in the original 749 position. That could explain why the 750 mutation score was much better than other mutations, and that the worst scored sequences were mutated at 749 position - breaking this bond.

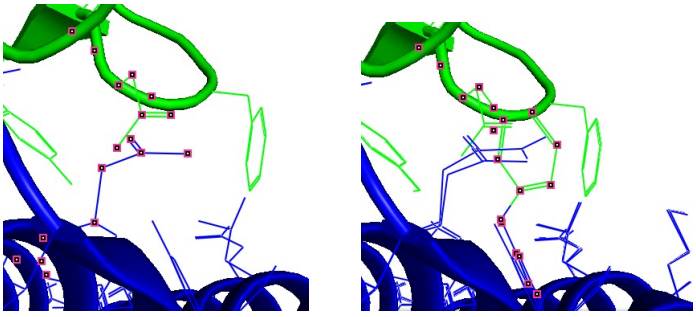

Figure 8: **Left:** Binding area with the original, not mutated sequence zoomed in. **Right:** Binding area with one of the best scored sequences, with mutation on 750 zoomed in. The interacting amino acids are selected with the red squares.

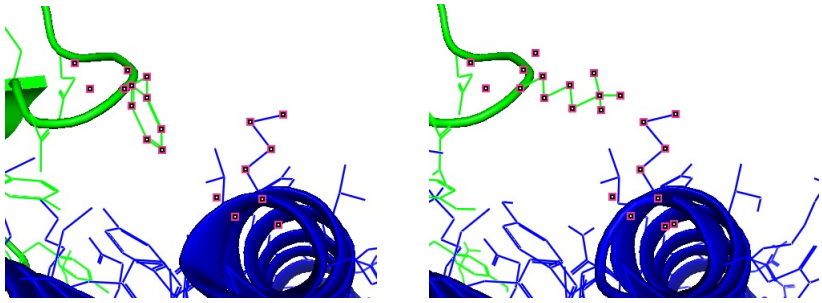

Figure 9: **Left:** Binding area with the original, not mutated sequence zoomed in. **Right:** Binding area with one of the worst scored sequences, with mutation on 749 zoomed in. The interacting amino acids are selected with the red squares.

Validation & Real variants analysis

We wanted to analyse the main real variants, from alpha to omicron, for the purpose of validating our method's results.

We looked at how many mutations appeared in each variant in its RBD region of the spike, and the SOAP score of each variant, shown below:

Variant name	number of RBD mutations	SOAP score
Wuhan	0	-2431.979
Alpha	1	-2519.612
Beta	3	-2529.631
Kappa	2	-2373.085
Lambda	2	-2467.97
Delta	2	-2402.579
Omicron	15	-2417.18

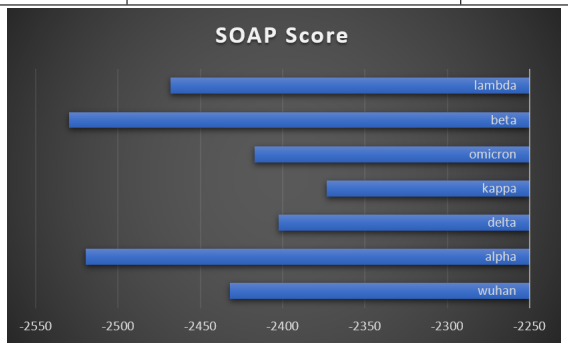

Table 1: Main existing variants statistics

To evaluate our ability to find possible future variants with our method, we performed the following validation steps:

1. We have extracted a list of the most common mutations and positions that appearead in the best scored, worst scored and existing variants.
2. We performed MSA (**multiple sequence alignment** using editing distance), to see the changes in the positions in the RBD area, between the best scored variants and the real variants sequences, to find more easily areas of similarity.
3. We performed an **evoltionary clutering algorithm** (i.e., UPGMA), to create the phylogenetic tree of the real variants with the best and worst scored sequences we generated, to evaluate the evolutionary emrging possibility from any of the real virus's variants.

From this evaluation steps, we have detected that the regions of **740-770** of the spike protein (the red area in the full PyMol 3D representation) were commonly mutated both in the best scored, worst scored and real variants sequences, leading to the conclusion that **meaningful mutations around this binding area can significantly change the binding affinity and infection ability of the virus.**

We have also detected the area around 600-640 to show significant mutations - for better or worse binding, in all 3 cases discussed.

((original AA, new AA, location), Number of appearances in

best 100 / worst 100 / real variants)

((('N', 'F', 750), 12)
(('N', 'Y', 750), 11)
(('Y', 'W', 752), 11)
(('G', 'R', 759), 9)
(('L', 'H', 718), 9)
(('Q', 'R', 756), 9)
(('G', 'W', 710), 9)
(('N', 'W', 750), 8)
(('G', 'Y', 710), 7)
(('F', 'Y', 719), 4)
(('N', 'L', 633), 2)
(('T', 'F', 693), 2)
(('A', 'R', 738), 2)
(('F', 'H', 778), 2)
(('R', 'S', 720), 2)
(('C', 'A', 751), 2)
(('Q', 'E', 677), 2)
(('V', 'Q', 773), 2)
(('S', 'Q', 638), 2)
(('A', 'K', 738), 2)
(('K', 'R', 619), 2)
(('P', 'M', 675), 2)
(('L', 'Q', 718), 2)
(('P', 'D', 754), 2)
(('F', 'K', 640), 2)
(('I', 'H', 665), 2)
(('L', 'G', 650), 2)

(('K', 'K', 749), 16)
(('F', 'Q', 749), 11)
(('F', 'E', 749), 9)
(('F', 'G', 749), 9)
(('F', 'P', 749), 7)
(('F', 'N', 749), 7)
(('F', 'D', 749), 6)
(('F', 'R', 749), 6)
(('Y', 'L', 768), 5)
(('F', 'S', 749), 4)
(('Y', 'C', 752), 3)
(('A', 'S', 615), 3)
(('F', 'A', 749), 3)
(('Y', 'S', 632), 2)
(('Y', 'G', 712), 2)
(('F', 'H', 749), 2)
(('Y', 'C', 768), 2)
(('Y', 'I', 768), 2)
(('D', 'A', 690), 2)
(('L', 'V', 598), 2)
(('A', 'E', 682), 2)
(('A', 'K', 660), 2)
(('Y', 'A', 752), 2)
(('T', 'S', 608), 2)
(('G', 'P', 602), 2)
(('V', 'K', 630), 2)
(('W', 'S', 699), 2)
(('E', 'I', 728), 2)
(('Y', 'S', 614), 2)

(('N', 'Y', 764), 3)
(('K', 'N', 680), 2)
(('T', 'K', 741), 2)
(('L', 'R', 715), 2)
(('E', 'K', 747), 1)
(('L', 'Q', 715), 1)
(('F', 'S', 753), 1)
(('G', 'D', 602), 1)
(('S', 'L', 634), 1)
(('S', 'P', 636), 1)
(('S', 'F', 638), 1)
(('N', 'K', 703), 1)
(('G', 'S', 709), 1)
(('S', 'N', 740), 1)
(('E', 'A', 747), 1)
(('Q', 'R', 756), 1)
(('G', 'S', 759), 1)
(('Q', 'R', 761), 1)
(('Y', 'H', 768), 1)
(('E', 'Q', 747), 1)

Figure 10: A list of the most common mutations and their positions in the best scored, worst scored and existing variants sequences

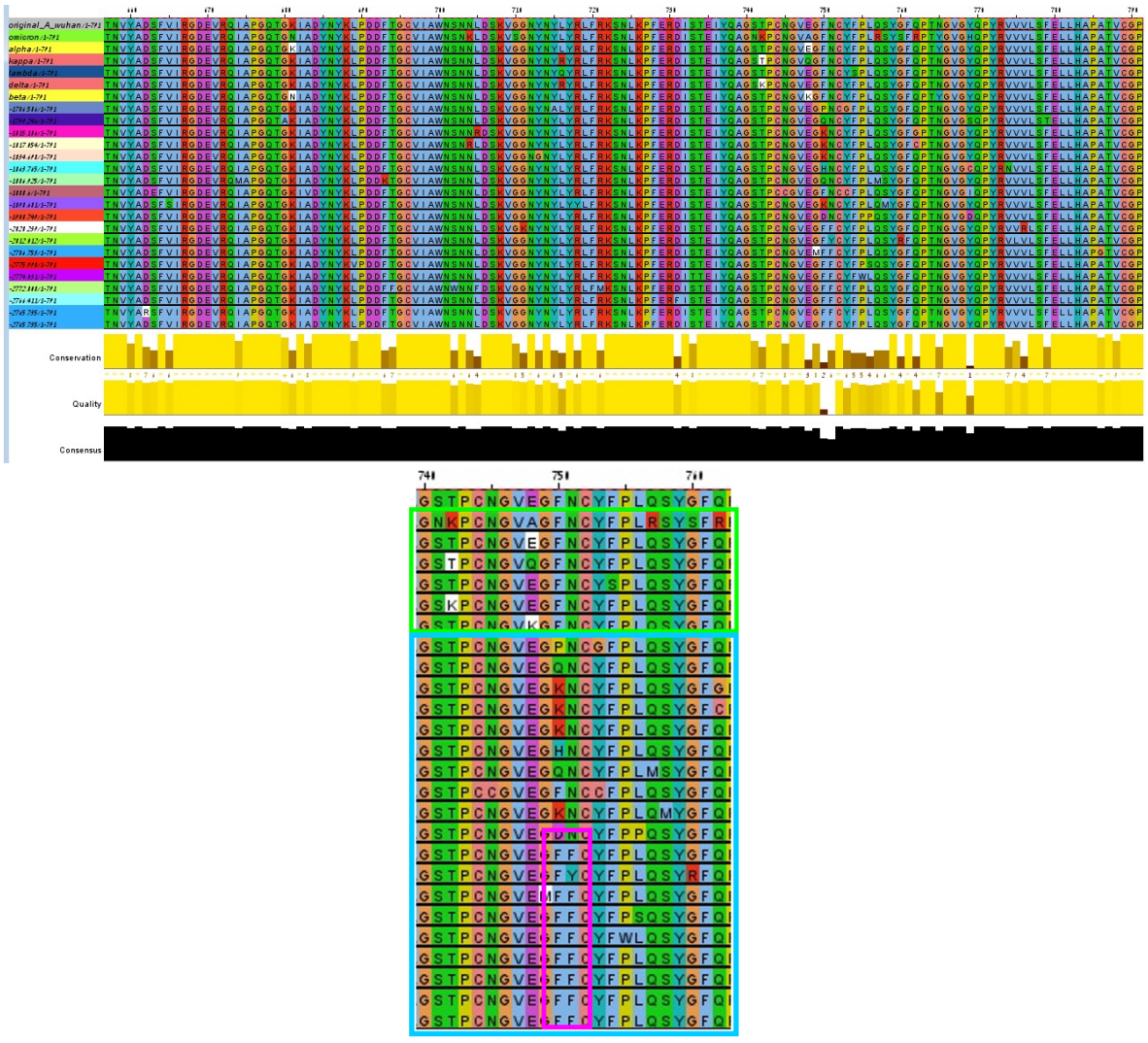

Figure 11: MSA of the original Wuhan strain, best scores sequences and the real existing variants. The MSA was performed on the full spike protein, and the screenshot shows the RBD area strating from 579 to 792. (Note: due to the alignment process, gaps were made and the 750 position in the sequences is now the 752 position in the MSA image).**Upper image:** full MSA. **Lower image:** the area of 740-760: first row in the Wuhan original strain, green is the real variants and blue is the best scored sequences. In pink we can see the 750 mutation $N \rightarrow F$ (in the MSA, 752 position).

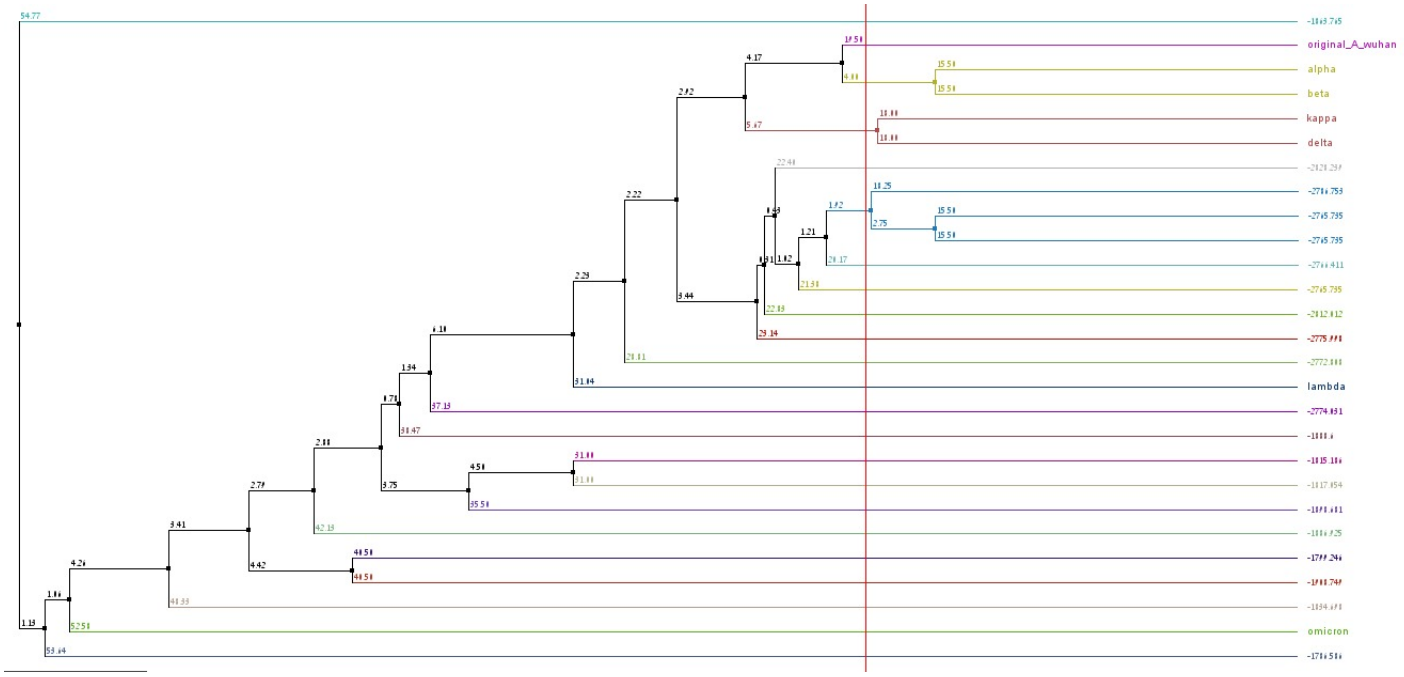

Figure 12: Evolutionary clustering: A phylogenetic tree of the real variants, best (lowest) and worst (highest) scored variants. The generated variants are noted by their score.

3.4 Impediments

Computational Complexity – Since there are a lot of possible mutations to check (Combinatorially, there are 20^{1274} optional sequences), we had to define a criteria for sequence elimination.

First, we tried to deploy a sophisticated model which is based on deep-learning techniques (https://github.com/anuprulez/clade_prediction) but unfortunately, after addressing the author, we got to know that the repository is still under development.

Afterwards, by getting further biology knowledge, we deducted that we should focus only on the RBD area (the S1 domain of the spike protein (i.e., sub-sequence) with less than 200 Amino-Acids).

In addition, we figured out that except the Omicron, a variant which has 15 changes (i.e., mutations) from the original Wuhan strain sequence, all the other variants have at most 3 changes. Therefore we limited our data to contain mutations that have at most 5 changes compared to the original Wuhan strain sequence.

SeqVec might not be sensitive for point mutations – Which makes it difficult to produce visualizations that have clear and meaningful distribution. In order to overcome it, we ran a PCA anlysis instead of the T-SNE, but found it to be even worse. That can lead to the conclusion that we SeqVec method is not a sensitive sequence embedding method, and another approach might be needed in order to visualise the results using T-SNE or PCA.

Folding strategy has changed during the project. We first wanted to use AlphaFold, a DeepMind software that predicts 3D structures of protein sequences, to fold all the generated sequences to their 3D structure from scratch. We found it to be inaccurate and time-consuming for doing it to all of the generated sequences. Thus, advised by Dr. Dina Schneidman, we chose to use a homology approach using SCRWL software, that can take a fixed structure (which we have had for the original Wuhan strain), and change its structure based on the new given point mutation positions. This has lowered our computing resources usage by a few folds, from computing 200 3D locations for each position, to 1-5 positions prediction in each new generated sequence, which also made the prediction more accurate, as it is based on a known 3D structure.

4 Further Work

- **Generating variants which have more than 5 mutations** - Check by the same manner which was done in our project how these variants may interact with the human ACE-2 receptor. The creation of such synthetic data may take a lot of time due to the combinatorial nature of the generation method and the computational complexity that the following algorithms in our method pipeline requires.
- **Mutations favoring sequence generation -** Generating mutations which have high probability to appear in nature, by using advanced NLP models, Probability Matrix etc.

- **Examine the top 100 generated variants** - In our project, we focused only on the top 1-10 generated variants due to time limitations. As a future work we would like to analyse more variants and possibly create an automatic analysis mechanism that can handle greater masses of generated variants.
- **Additional scoring method** - As the scoring method is given by a specific energy equation, it will benefit the results to use other scoring methods and compare the results between these methods.
- **More mutations types** - in our project we used only a mutation type called base substitution that causes an amino acid change in that position. In biology we can find more mutation types that should be condisered - deletion (i.e., delete an nucleotide which can change the amino acids) or insertion (which inserts a nucleotide which can change the amino acids). This could mimick the natural behaviour of mutations more accuratly, and thus be more expressive and lead to more real life possible results.

5 Conclusion

After various anlysis steps and evaluations of the results we can conclude that there are specific regions in the RBD sequence that is prone to changes that are significant.

These regions were also shown to be common in all real variants, which our method was successfully able to catch.

The regions of **740-770 and 600-640** of the spike protein were commonly mutated both in the best scored, worst scored and real variants sequences, leading to this conclusion, which is that **meaningful mutations around this binding area can significantly change the binding affinity and infection ability of the virus.**

In addition, we found out that there are **two positions - 749 and 750 - that are crucial** for the binding, and makes the difference between the best scored sequences mutations and the worst ones. The original 749 is aromatic, which bonds perfectly with other aromatic amino acid - thus when 750 becomes aromatic, the bond is tighter, and when we change the amino acid in 749 to be non-aromatic, we break this bond and might create rejection in the area that bonds with the human ACE-2.

Overall, our method showed significant results which can teach us more about the nature of new variants that might emerge, and even supply a baseline to future variants that are probable to come.

6 Supplements

The code is attached in a zip file, and can be run only on cs computers.

A video of the run can be found in a link in the README file which is located in the zip code file.