MAIS 202 Deliverable 2 – Progress report
Ritchie Yu

Problem

As stated in Deliverable 1, this project aims to predict next-generation mutations in SARS-CoV-2 using historical data. This is useful for anticipating potentially dangerous mutations as the virus evolves, which would aid vaccine development. I have made some changes since Deliverable 1, which I outline below.

Adjustments and notes

Original plan

Predict mutations occurring in the full SARS-CoV-2 nucleotide sequence. For that, I proposed to use the GISAID database. I also proposed to use global data without discriminating by geographic location

Updated plan

The GISAID database continues to be my source of data. However, I have realized that my original plan to use entire genomes is very computationally expensive. A single SARS-CoV-2 genome contains ~30 000 letters, and I have over 100 000 of them. So, to more quickly produce preliminary results, I have made three temporary adjustments:

i) I've decided to try to predict *amino acid* mutations before moving on to the genomic mutations. Amino acid sequences are significantly shorter than genomic sequences, so training is less expensive.

ii) These amino acid sequences constitute the spike glycoprotein of SARS-CoV-2, as depicted in the diagram below. By choosing to predict mutations occurring in one particular protein of the virus, I am also reducing the computational power needed.

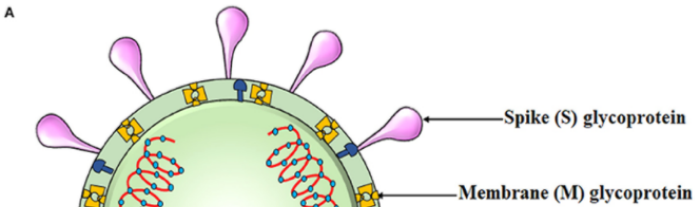

Figure 1. Spike glycoprotein of SARS-CoV-2 (Shah et al., 2020). This is a very important protein for the virus, as it is used for cellular entry and infection. It will also likely be the antigenic target of vaccines.

Combined, i) and ii) reduce a single sequence length to ~1200 letters

iii) I'm now using data obtained from labs in the USA rather than global data, since I believe it may be easier to capture evolutionary changes in a smaller region. I chose the USA because they have produced the most SARS-CoV-2 data.

Important note

During data preprocessing, I noticed that the number of mutations that have occurred in the amino acid sequence of the spike glycoprotein is much smaller than I expected. If it becomes the case that there have not been enough mutations since the pandemic began for my model to work, I may need to use influenza virus data instead. The data that would be needed for this is also available on GISAID.

Data preprocessing

Like I mentioned earlier, the GISAID database is still the database I am using. However, for now I am using amino acid sequences in addition to genomic sequences. These sequences were all collected in the USA. All sequences are downloaded in FASTA format, and fortunately, GISAID already pre-aligned the sequences, so I did not need to do multiple sequence alignment myself.

The FASTA files were parsed in Biopython. Each sequence contains five labels, but only two of the labels are relevant for this project: "ID" and "Seq". Example sequences are shown below.

```
ID: Spike|hCoV-19/Wuhan/WIV04/2019|2019-12-30|EPI_ISL_402124|Original|hCoV-
Name: Spike|hCoV-19/Wuhan/WIV04/2019|2019-12-30|EPI_ISL_402124|Original|hC
Description: Spike|hCoV-19/Wuhan/WIV04/2019|2019-12-30|EPI_ISL_402124|Origi
Number of features: 0
Seq('MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDL...HYT')
```

Figure 2. Spike glycoprotein amino acid sequence of the original SARS-CoV-2 virus (Wuhan). "ID" contains information such as where and when the virus was collected. "Seq" is the full amino acid sequence.

```
ID: hCoV-19/Wuhan/WIV04/2019|EPI_ISL_402124|2019-12-30|China
Name: hCoV-19/Wuhan/WIV04/2019|EPI_ISL_402124|2019-12-30|China
Description: hCoV-19/Wuhan/WIV04/2019|EPI_ISL_402124|2019-12-30|China
Number of features: 0
Seq('-----------------------------------------------------...---')
```

Figure 3. Full nucleotide sequence of the original SARS-CoV-2 virus (Wuhan). Again, "ID" contains information on when and where the virus was collected. "Seq" is the entire virus genome. The dashes are simply there for multiple sequence alignment purposes. The actual nucleotides can be seen if the sequence is expanded.

At this preliminary stage, only the amino acid sequences have been preprocessed. Three steps were taken:

1. Clean the FASTA file by removing sequences with "X"s. The "X"s mean that the sequence is incomplete, so those sequences were removed. As precaution, a snippet of code was also inserted to remove sequences with duplicate IDs.
2. With the clean FASTA file, sequences collected in countries besides the USA were removed. I will attempt to train global data at a later stage in this project.
3. The remaining sequences were then sorted by collection date

4. Finally, each amino acid sequence was converted into a 100 x 1 vector representation using pre-trained vector representations of 3-gram amino acids (Asgari et al., 2015). These protein embeddings are similar in principle to word embeddings. This data was publicly available on Harvard Dataverse so I did not need to train the representations myself.

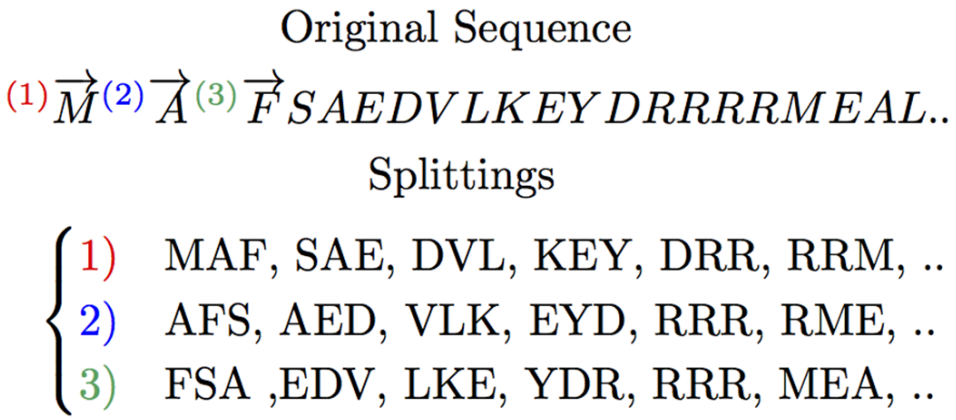

Figure 4. Amino acid sequence 3-gram splitting. Each 3-gram is embedded into a 100 x 1 vector representation. By summing all of the 3-gram vector representations in (1), (2), and (3), a 100 x 1 vector representation of the entire sequence is obtained. The effect of summing overlaps is to preserve order.

Machine learning model

(Will be added in the following days)

Works cited

Shah, V. K., Firmal, P., Alam, A., Ganguly, D., & Chattopadhyay, S. (2020). Overview of Immune Response During SARS-CoV-2 Infection: Lessons From the Past. *Frontiers in Immunology*, *11*. https://doi.org/10.3389/fimmu.2020.01949

Asgari, E., & Mofrad, M. R. K. (2015). Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *Plos One*, *10*(11). https://doi.org/10.1371/journal.pone.0141287