Predicting Mutations in SARS-CoV-2 with a Deep Learning Model

Annabelle Hsieh | hsiehannabelle2024@gmail.com

SHINE Lab
La Canada High School, Class of 2024
USC Viterbi Department of Electrical Engineering, SHINE 2023

Introduction

With the recent pandemic, SARS-CoV-2 has been an important topic of research. Due to its rapid mutation rate and contagious nature, it is extremely difficult to create a vaccine that can work effectively against the virus and continue to be used for a long period of time. In my lab, my teammate and I used a pre-made deep learning model to predict mutations in SARS-CoV-2 and tested its accuracy and learning loss with different sequence lengths.

Figure 1. An example of how the deep learning model works as well as its objectives.

Objective & Impact of Professor’s Research

The Cyber-Physical Systems lab focuses on using machine learning algorithms and mathematical models to analyze complex networks and process the rules and patterns that define the relationships in those systems. Networks play an important role in much of our everyday life with some of the most prominent examples being the Internet, our social network (i.e. relationships with other people), and our biological network. Understanding these systems can help lead to a greater and more efficient understanding about how these networks work and what they will do in the future.

Acknowledgements

I would like to thank Professor Bogdan for giving me this research opportunity as well as my mentors Xiongye Xiao and Qi Cao for guiding me through my project. Additionally, I would like to thank Marcus Gutierrez (my Center Mentor), Vela Benedicto (my teammate), and ChatGPT (a large language model) for supporting me through this research process.

Research & Learning Process

To complete this project, there were lots of outside information and skills that we needed to know beforehand.

1. Data collection

<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>Position</th>
<th>Type</th>
<th>Sequence Cut</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA297</td>
<td>241 C + T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A99G</td>
<td>405 A + T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TN306</td>
<td>687 T + A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TN182</td>
<td>279 C + T</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

   Figure 2a. (above) A picture of the csv file I coded.

   Figure 2b. (to the left) The company we collected data from.

2. Understanding network science and deep learning

   Figure 3a. (above) A visual of how deep learning works.

   Figure 3b. (to the right) The textbook we read to learn the basics of network science.

3. Learning how to use PyTorch + PyMOL

   Figure 4. A model of the nine predicted amino acids within the D614G mutation.

4. Understanding and adapting the code

   Figure 5. An image of the code written to produce a confusion matrix that depicts the model’s accuracy.

Methods & Results

Overall, the model performed best when cut into sequences with a length of 27 bases. As seen in the confusion matrix below, the model was correct more than 70% of the time for all bases and reached over 80% accuracy for bases C and T.

Figure 7. A confusion matrix of the model's performance with a sequence length of 27.

Results Analysis

In the future, I would like to look more into the application of network science and deep learning, especially in relation to mental health as it is a subject that I feel is very relevant today. When I initially started, the work seemed overwhelming and almost impossible to do given my little experience in research. However, with time, I definitely picked up many important skills from my mentors. To future SHINE students, I would recommend not being afraid to ask questions and to come in with the mentality that you are here to learn, so it is ok if you cannot exactly match the work the Ph.D. students are doing.

Citations


