Time Series Prediction Of SARS-CoV-2 Spike Glycoprotein Mutations With LSTM Neural Networks

Ritchie Yu
McGill University

Introduction

- Mutation of SARS-CoV-2 may lead to viral structural changes which challenge the efficacy of vaccines and antivirals
- Pre-emptively identifying next generation mutations would aid novel drug development

- SARS-CoV-2 spike glycoprotein is an important potential drug target
- This project proposes using LSTM networks to predict next generation mutations in the spike glycoprotein amino acid sequence.

Strategy

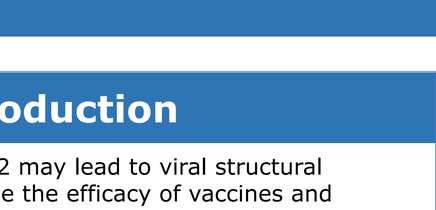
1. **Multilabel binary classification problem**
   - Predict if mutation occurs next month, at every site
2. **Single site binary classification problem**
   - Predict if mutation occurs next month, at specific site

Data

Represent amino acid 3-grams as vectors (ProtVec)

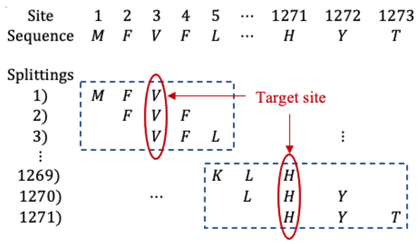
Multilabel classification

- Sum all 3-gram splittings to represent entire amino acid sequences

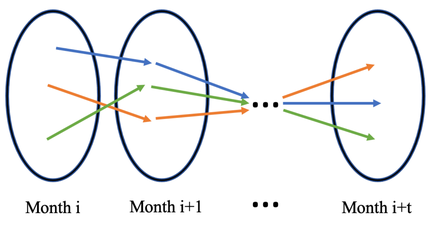


Site classification

- Sum neighbouring 3-gram splittings to represent target site

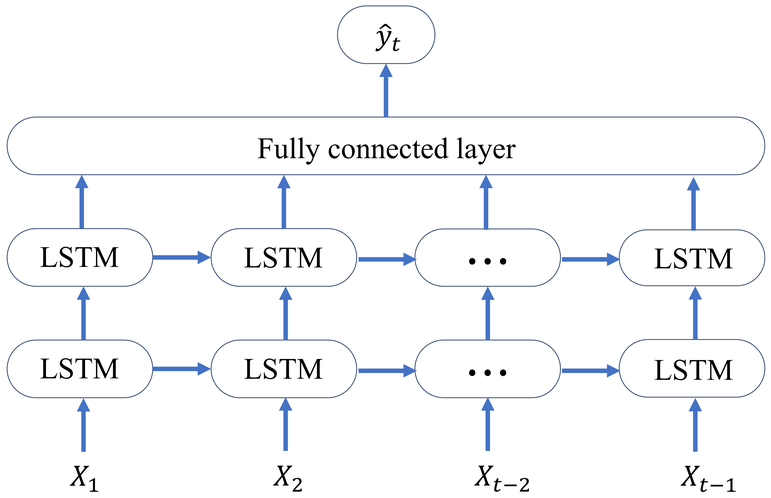


Random time-series linkage to estimate evolutionary history



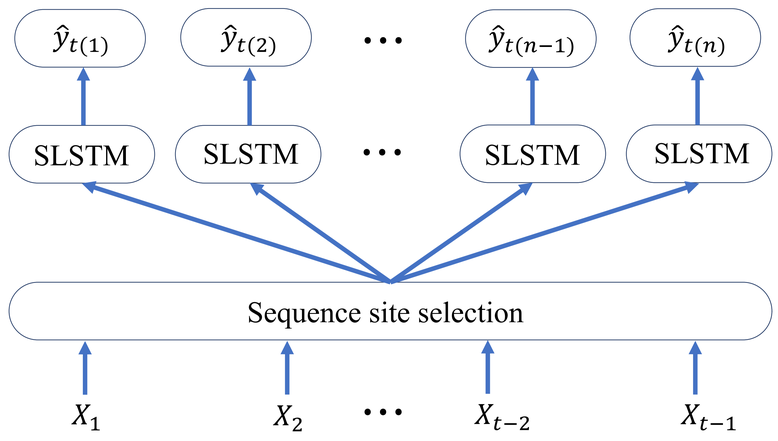
Network architecture

Multilabel binary classification



Each X_i represents a single embedded amino acid sequence sampled from one month, ordered temporally. $\hat{y}_t$ is a vector representing which sites at time t are predicted to mutate.

Single site binary classification

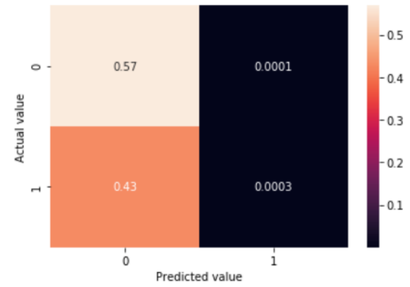


Each X_i represents a single embedded amino acid sequence sampled from one month, ordered temporally. A target site for prediction is selected and passed through a dual stacked LSTM and a fully connected layer. $\hat{y}_{t(i)}$ is a binary scalar representing whether a mutation at the site is predicted to occur at time t.

Results

Present results for both the multilabel and single site binary classification problems are inconclusive. Both models suffer from unchanging loss, validation and training accuracy, as well as precision and recall. Accuracy on test data for both models is essentially equivalent to guessing the labels.

Example: Normalized confusion matrix for single site mutation prediction at site 222. The model accurately predicts negative instances, but not positive ones.



Conclusions

It has previously been shown that LSTM neural networks combined with attention can effectively predict mutations in influenza A virus (Yin et al., 2020). This suggests that while the present results are inconclusive, the problem likely is not with the model.

Possible reasons for poor results:

- Given the COVID-19 pandemic has only recently begun, the evolutionary history of the virus, captured by the data used, may be too short for the LSTM networks to properly learn.
- Random sampling of global sequences may poorly capture evolutionary history. If so, clustering methods could be a solution.

Further exploration of these reasons outlined is needed to improve the accuracy of the proposed models in predicting SARS-CoV-2 mutations.

Citations

Yin, R., Luusua, E., Dabrowski, J., Zhang, Y., & Kwoh, C. K. (2020). Tempel: time-series mutation prediction of influenza A viruses via attention-based recurrent neural networks. *Bioinformatics (Oxford, England)*, *36*(9), 2697–2704. https://doi.org/10.1093/bioinformatics/btaa050

Asgari, E., & Mofrad, M. R. K. (2015). Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLOS ONE*, *10*(11), e0141287. https://doi.org/10.1371/journal.pone.0141287

Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell*, *181*(2), 281-292.e6. https://doi.org/10.1016/j.cell.2020.02.058

We gratefully acknowledge the Authors from the Originating laboratories responsible for obtaining the specimens and the Submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which this research is based.