Amino Acids NN Report

Thomas Knickerbocker, Romell Padua

Abstract:

During the Coronavirus pandemic of 2020, the biggest issue of the new virus was the ease in which it was transmitted. With contagiousness taking the world by storm, it begs the question: “Is there a way to identify how contagious a virus is?” Identifying the factors that contribute to the infectiousness of a virus is integral to preventing the spread. One element that contributes to the attributes of a virus are the amount of amino acids their genome contains and the organization that they are in. A Multilayer Perceptron (MLP) is a neural network and a subset of machine learning. It uses many layers to learn and predict features of a dataset . Each of these layers provides a filter and the output is what will be used as the input for the next layer. There have been studies using machine learning looking at genetic sequences, but none using MLPs to analyze the features of a genome. Here we show that the amount of each amino acid present in a genome has some effect on the contagiousness of a Covid variant, but it is not a conclusive way to identify the potential infection rate. Based on prior studies it is theorized that Lysine is an amino acid that can help to prevent infection and arginine can potentially increase infection rate. Our findings demonstrate that the amount of amino acids present in a Covid variant’s genome sequence can be used as a factor when predicting outbreaks and the potential spread of variants, but other features will need to be present in order to conclusively identify how contagious a variant is. We believe that our model can be used to help prevent outbreaks by alerting scientists to how contagious a virus is. Machine learning models like ours can identify patterns in large samples of data that the human eye would not be able to recognize.

Summary of Previous Findings:

What we are doing has not been done exactly, but there are a number of studies that describe the impact of machine learning on genomes and the relationship between amino acids and contagiousness. One such study by Walaa Alkady, Khaled ElBahnasy, Víctor Leiva, and Walaa Gad compares different machine learning algorithms and their performances on identifying whether a genome is COVID-19. For their features they wrote a program to grab the dipole and volume values from the sequences and used those to train their models. The algorithms that they used were Bagging ensemble (BE), Decision trees (DT), Gradient boosting (GB), k-nearest neighbors (KNN), RF, and SVM. In this paper they briefly mentioned that machine learning can be used to predict infection rate of diseases which we decided to do. Since they proved the effectiveness of machine learning algorithms on genomic data, we expanded upon that by employing a Multilayer Perceptron (MLP) on genomic data to identify the contagiousness of a virus.

Additionally, we came across a study by Rayner Alfred and Joe Henry Obit where they did a study of different ai models and how well they are able to predict and detect various disease outbreaks. They looked at many models and found that deep neural networks such as FNN and BMA performed the best when compared to all other models tested such as ARMA,

LASSO, and MARS. This study would help to determine the type of machine learning model we would end up using. Since deep neural networks seemed to perform the best in this study when predicting and detecting potential outbreaks, we chose the neural network MLP as our model.

The first way we figured to calculate a contagiousness score to train our data on was the amount of uploads from the first upload date to directly a month afterward. One potential problem with this metric was that some variants of the strain were discovered after the first vaccine came out, leading to a lower contagiousness. One study by Seyed M. Moghadas, Thomas N. Vilches, Kevin Zhang, Chad R. Wells, Affan Shoukat, Burton H. Singer, Lauren Ancel Meyers, Kathleen M. Neuzil, Joanne M. Langley, Meagan C. Fitzpatrick, and Alison P. Galvani highlighted many ways that the vaccine affected the spread of Covid-19. One figure that we used to influence our contagiousness score was that approximately 81 percent of people in the US received the vaccine and it lowered the attack rate down to 40-50 percent.

Results:

We expanded upon the question: “Is there a way to identify how contagious a virus is?” to get: “What aminos contribute most to the contagiousness of a virus. To answer this, we chose to use a Multilayer Perceptron (MLP)to predict the contagiousness of a certain variant of the Covid virus. For the data that we fed into the model, it became a little more complicated. In the data we downloaded, the variant uploads did not come with a calculated contagiousness score that we could use as a target for our MLP. However, it did come with a collection data for each upload. Our contagiousness score for each unique variant was calculated by the amount of uploads within the first month of uploads of that variant. One potential factor that would influence the contagiousness score is when the vaccine came out. Variants that arose after the vaccine came out would end up having lower contagiousness scores since we were more well prepared at that point. One thing we did to reconcile this was to add a small multiplier to the contagiousness scores after the vaccine came out. After some research, in the US around 80 percent of people were able to get the vaccine and the vaccine was able to lower the incoming cases by 40-50% initially. We used these numbers and applied that to variants’ contagiousness scores if their collection date was after the first vaccine came out. This is not a perfect representation of how contagious a virus is but given the available categories of data and the amount of data needed to properly train a model, these scores were sufficient. We then used the pandas library to create a new dataframe containing the following columns: Accession, Date_Collected, Pangolin, a letter for each of the 20 common amino acids, other (encompassing all other amino acids, which are much less frequent in the data), and Contagiousness_Score (our target variable).

When building our model we implemented various concepts that would contribute to the accuracy of the data. Our MLP architecture uses a series of fully connected linear layers stacked upon each other forming a “multilayer” structure with an 80/20 train-test split. This allows our model to extract increasingly complex features from the input data as it progresses through the layers. MLPs are similar to Convolutional Neural Networks (CNN) in that they use a layered approach, but instead of being made with convolutional layers and specializing in image data, MLPs are much better suited for sequential data like amino acids. For our loss function we used Mean Squared Error. Mean Squared error is typically used as a loss function in regression

problems where the target is to predict continuous values and our single target variable was the Contagiousness_Score column. Two strategies that we implemented into the model to increase efficiency and accuracy were Xavier initialization and activation functions. The activation function that we utilized was Rectified Linear Units (ReLu) after each hidden layer. This introduces non-linearity to the layer allowing it to learn more intricate features. The Xavier initialization randomly initializes weights in the linear layers to prevent vanishing gradients during training, leading to increased efficiency. In addition to improving accuracy, we had problems with our model overfitting, leading us to implement the preventative strategies of batch normalization and dropout regularization. Batch normalization helped to stabilize the training process by reducing internal covariate shift and improving convergence. The dropout regularization combated overfitting by dropping out neurons during training, reducing overfitting and promoting generalization

Figure 1:

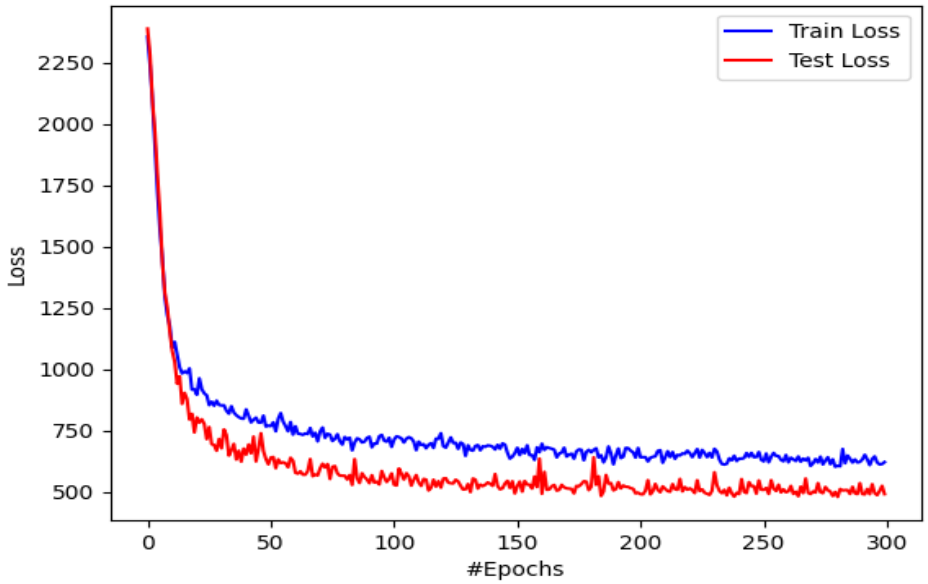
Standard Deviation of Contagiousness_Score: 40.11

	A	C	D	E	F	...	V	W	Y	Other	Contagiousness_Score
count	6000.00	6000.00	6000.00	6000.00	6000.00	...	6000.00	6000.00	6000.00	6000.00	6000.00
mean	372.34	509.19	256.14	265.91	523.43	...	633.02	203.82	427.72	7.02	28.47
std	95.66	127.09	86.70	62.98	62.92	...	87.66	60.71	79.77	26.41	40.11
min	242.00	272.00	112.00	169.00	433.00	...	521.00	91.00	318.00	0.00	1.00
25%	259.00	468.00	133.00	192.00	460.00	...	550.00	191.00	331.00	0.00	3.00
50%	374.00	486.00	289.00	269.00	474.00	...	657.00	207.00	396.00	0.00	9.72
75%	392.00	633.00	311.00	271.00	589.00	...	672.00	259.00	507.00	2.00	27.94
max	535.00	639.00	357.00	383.00	599.00	...	784.00	265.00	516.00	893.00	125.14

[8 rows x 22 columns]

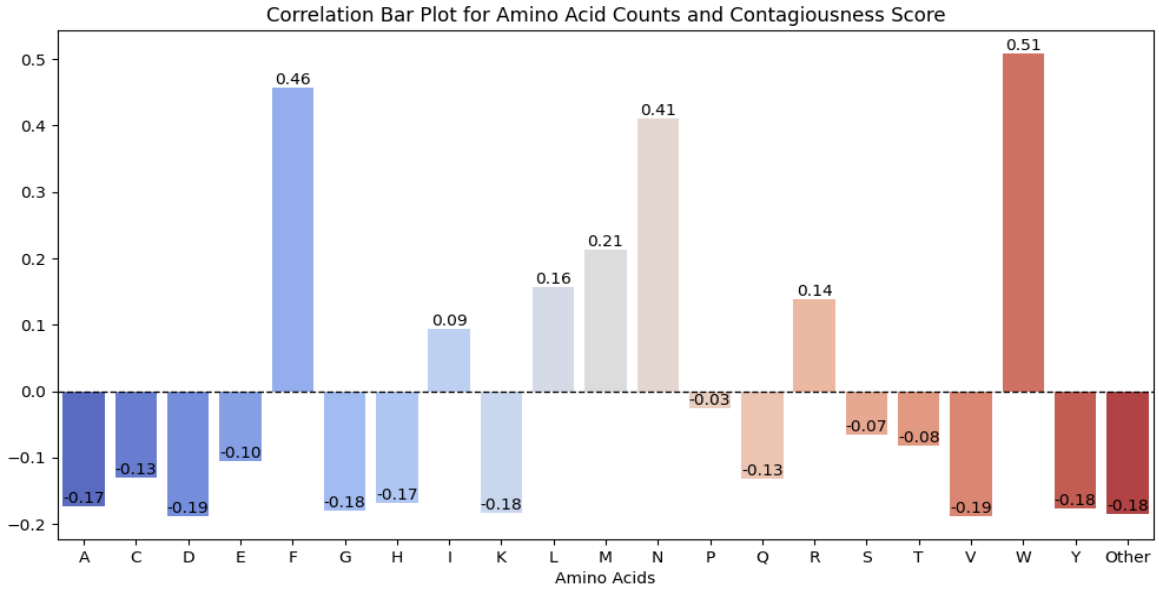
The figure illustrates data from 6,000 random samples, with the first 21 columns as input variables and the last column representing the Contagiousness Score. Variability in raw data stems from imbalances in pangolin sample sizes (many pangolins with small number of samples & contagions scores and a few with very large numbers of samples and thus large contagion scores).

Figure 2:



This figure displays the model's training and testing loss as it processes data. Initial high losses decrease rapidly but plateau around epoch 50, with a mean absolute error of 14.45, significantly lower than the standard deviation of the target variable.

Figure 3:



The above figure shows the correlation between the input counts of each amino acid and the contagiousness score. According to our model, Tryptophan (W), Phenylalanine (F), and Asparagine (N) are the amino acids that have the highest correlation to a high contagiousness score.

Conclusion

The purpose of this study was to determine which amino acids contributed most to the contagiousness of a Covid variant. The results of our efforts turned out only to be somewhat conclusive. Our model ended up predicting that the amount of Tryptophan, Phenylalanine, and Asparagine amino acids were most correlated with how contagious the variant was and the mean absolute error for the best model was 14.45. The model was able to improve with each batch of data, but at some point its performance plateaued. This trend by our model suggests that there are some correlations between the amount of each amino and how contagious the virus is since it was initially able to improve. The error is much lower than the standard deviation of contagiousness values, However, because the model's performance stopped improving at a certain point, we can infer that a sequence has additional factors that are imperative in predicting the contagiousness of a virus.

Methods:

The data we used was a set of 6,000 random samples of complete COVID-SARS-2 nucleotide sequences sourced from humans in the United States, and was downloaded via the NCBI website linked here. Since pangolin data was unavailable in the website's random sample functionality, we downloaded the random samples and then matched their accession numbers to the unique set of ~1.3 million sequences which includes both the pangolin and dates collected for samples. We then calculated the contagiousness score by multiplying the frequency of the pangolin in the first month it appears within our sampled data by either one if the pangolin first appears before the end of the first month of the vaccine's release, and 1.5 * .81 in the "after" case for the reasons outlined within the previous findings section. Codons are then counted and converted into amino acids for each of the sampled sequences. Then, a MLP is trained with an 80/20 train-test-split and oversampling to help prevent majority contagiousness scores dominating prediction results (the issue detailed in Figure 1), a StandardScaler is applied to the data, and the model is saved for analysis.

Acknowledgements:

Data Organization: Romell/Tommy
Model Construction: Tommy
Model Fine Tuning: Romell/Tommy
Report: Romell

References:

1) Melano, I., Kuo, L.-L., Lo, Y.-C., Sung, P.-W., Tien, N., & Su, W.-C. (2021). Effects of Basic Amino Acids and Their Derivatives on SARS-CoV-2 and Influenza-A Virus Infection. *Viruses*, *13*(7), 1301. https://doi.org/10.3390/v13071301
2) Alkady, W., ElBahnasy, K., Leiva, V., & Gad, W. (2022). Classifying COVID-19 based on amino acids encoding with machine learning algorithms. *Chemometrics and Intelligent*

Laboratory Systems, *224*, 104535. https://doi.org/10.1016/j.chemolab.2022.104535
3) Moghadas, S. M., Vilches, T. N., Zhang, K., Wells, C. R., Shoukat, A., Singer, B. H., Meyers, L. A., Neuzil, K. M., Langley, J. M., Fitzpatrick, M. C., & Galvani, A. P. (2021). The impact of vaccination on COVID-19 outbreaks in the united states. *MedRxiv*, *2*(2). https://doi.org/10.1101/2020.11.27.20240051
4) Alfred, R., & Henry Obit, J. (2021). The roles of machine learning methods in limiting the spread of deadly diseases: A systematic review. *Heliyon*, *7*(6), e07371. https://doi.org/10.1016/j.heliyon.2021.e07371

Links:

1) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8310019/
2) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8923015/
3) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7709178/
4) https://pubmed.ncbi.nlm.nih.gov/34179541/

Data Source)
https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Severe%20acute%20respiratory%20syndrome%20coronavirus%202,%20taxid:2697049&SourceDB_s=GenBank&HostLineage_ss=Homo%20sapiens%20(human),%20taxid:9606&BaselineSurveillance_s=include&Completeness_s=complete&Region_s=North%20America

Please note that we received an extension for this project to December 18th, 2023