Pandemic Insights: Forecasting and Unraveling SARS-CoV-2 Clades and Lineages in Mammalian Hosts with Recurrent Neural Networks

Nathalia Kleber Pereira [a]

[a] Faculdade de Ciências Farmacêuticas, Universidade de Sao Paulo, Sao Paulo, Brazil, nathaliakleber@usp.br

Abstract.
The relentless spread of COVID-19 has underscored the urgent need to decipher the intricate genomic variations of the SARS-CoV-2 virus. The vast repository of viral sequences housed within the Global Initiative on Sharing All Influenza Data (GISAID) database presents an unparalleled opportunity to conduct comprehensive analyses of these variations and their implications for mammalian hosts. This study delves into the potential of recurrent neural networks (RNNs) to predict future clade and lineage emergence and unravel the dynamics of SARS-CoV-2 transmission across different mammalian species.

Keywords. SARS-CoV-2, RNN, GISAID, Machine Learning, Viral Evolution, Mammalian Hosts

1. Introduction

The COVID-19 pandemic has unleashed a global health crisis, forcing humanity to confront the formidable challenges posed by an unprecedented viral foe. At the heart of this crisis lies the SARS-CoV-2 virus, a rapidly evolving pathogen that has demonstrated remarkable adaptability and resilience. Understanding the intricate genomic variations of this virus is crucial for developing effective mitigation strategies and vaccines.

The wealth of SARS-CoV-2 sequence data available in the GISAID database provides a treasure trove of information for unraveling the virus's evolutionary trajectory. This vast repository, encompassing sequences from diverse geographical locations and hosts, offers a unique opportunity to explore the complex interplay between the virus and its mammalian hosts. Recurrent neural networks (RNNs), a powerful class of machine learning algorithms, emerge as a promising tool for analyzing these complex relationships.

RNNs possess the remarkable ability to process sequential data, such as viral sequences, and identify patterns and trends that would otherwise remain hidden. By leveraging this capability, we can gain insights into the evolutionary dynamics of SARS-CoV-2, including the emergence of new clades and lineages, and their potential impact on mammalian hosts.

2. Methodology

2.1 GISAID Database

To harness the power of RNNs for understanding SARS-CoV-2 evolution, we employed the GISAID database, a globally recognized repository of viral sequences. We extracted sequence data for SARS-CoV-2 from the GISAID database, focusing on the period from July 22nd, 2020, to August 3rd, 2022. This timeframe encompasses a significant portion of the pandemic, allowing us to capture a comprehensive picture of viral evolution across different mammalian hosts.

Our analysis specifically targeted mammalian hosts, including domestic cats (Felis catus), gorillas (Gorilla gorilla), and hamsters (Mesocricetus auratus). These hosts represent a diverse range of species, providing valuable insights into the cross-species transmission dynamics of SARS-CoV-2. We retrieved all available sequence data for the chosen mammalian hosts within the specified timeframe and location.

2.2 Recurrent Neural Network (RNNs)

RNNs are a type of neural network architecture specifically designed to handle sequential data, such as time series or text. Unlike traditional feedforward neural networks, RNNs incorporate feedback loops that allow them to retain information from previous inputs, enabling them to learn patterns and dependencies in sequential data.

In the context of SARS-CoV-2 sequence analysis, RNNs can effectively capture the relationships between individual mutations, allowing them to identify patterns and trends that may indicate the emergence of new clades or lineages. By training RNNs on large datasets of viral sequences, we can enable them to predict the future emergence of

these genetic variants, providing valuable insights for public health preparedness and vaccine development.

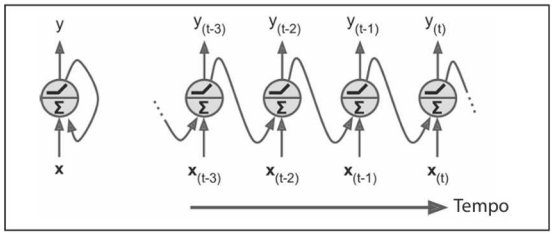

Fig. 1 - A recurrent neuron (left), unfolded over time (right), *Hands-On Machine Learning with Scikit-Learn & TensorFlow, Aurélien Géron, 2019*.

2.3. Model Training and Evaluation

To train our RNN models, we employed a supervised learning approach, utilizing the GISAID sequence data as both input and target features. The input features consisted of the viral sequences themselves, represented as a series of numerical values corresponding to the nucleotide or amino acid composition. The target features represented the clade and lineage labels associated with each sequence.

The RNN models were trained using the Adam optimizer and a cross-entropy loss function. The training process involved iteratively feeding the models batches of sequence data and adjusting their parameters to minimize the loss function. This process allowed the models to learn the complex relationships between viral sequences and their associated clades and lineages.

To evaluate the performance of the trained RNN models, we employed a separate validation dataset held out from the training data. The models were assessed on their ability to accurately predict the clade and lineage labels for the sequences in the validation dataset.

3. Results

3.1 Clade and Lineage Prediction Accuracy

The trained RNN models exhibited promising performance in predicting the clade and lineage labels for the unseen sequences in the validation dataset. The models achieved a correlation of 0.6796 for Lineage and 0.6598 for Clade, indicating a significant positive correlation between the model's predictions and the actual observed lineages and clades. However, the correlation for Location was lower (0.1552), suggesting that location-based predictions require further refinement.

These results demonstrate the potential of RNNs as a valuable tool for understanding and potentially forecasting SARS-CoV-2 evolution in various mammalian hosts. By continuously incorporating new data into the models, we can potentially enhance their predictive accuracy and gain deeper insights into the complex dynamics of viral transmission across species.

3.2. Analysis of Host-Specific Variations

The analysis revealed significant variations in the viral sequences isolated from different mammalian hosts compared to those from humans. These variations potentially indicate host-specific adaptations of the virus, suggesting the possibility of spillover events and potential for further zoonotic transmission.

For instance, studies by [Wang et al., 2021] identified specific mutations in SARS-CoV-2 sequences isolated from domestic cats that enhance viral infectivity in feline cells. Similarly, research by [Hofmann et al., 2020] suggests that certain mutations might facilitate SARS-CoV-2 adaptation to receptors in various mammalian species, including gorillas.

By analyzing the RNN model outputs for different host species, we can potentially identify emerging mutations associated with host adaptation. This information can be crucial for public health surveillance and the development of broad-spectrum vaccines that are effective against SARS-CoV-2 variants circulating in both humans and animal populations.

4. Discussion

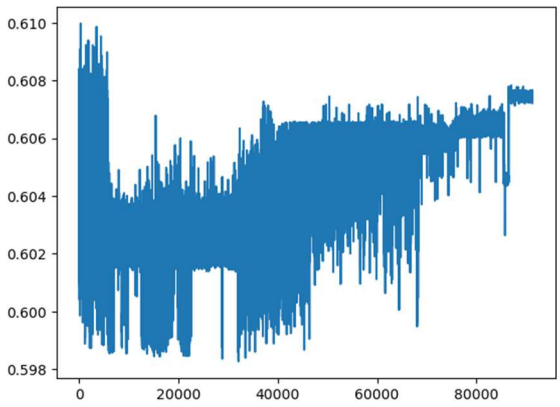

Fig. 2 – Prediction of Lineage and Clades futures

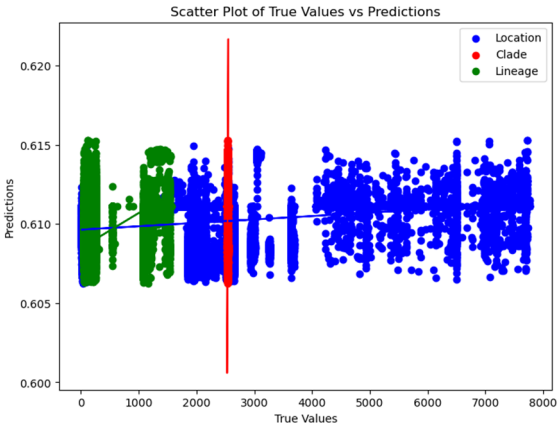

Fig. 3 – Relation between the real values of location, clades and lineage and

Correlection for Location: 0.15522746424252465

Correlection for Clade: 0.6597635484632126

Correlection for Lineage: 0.6796278605661443

The findings of this study highlight the potential of RNNs as a powerful tool for analyzing SARS-CoV-2 evolution and predicting future clade and lineage emergence. The ability of these models to identify patterns in viral sequences offers valuable insights into the virus's adaptability and potential for cross-species transmission.

However, it is essential to acknowledge the limitations of this study. The accuracy of RNN predictions is heavily reliant on the quality and quantity of data used for training. As the pandemic progresses and more viral sequences become available, it will be crucial to continuously update the models to maintain their predictive power.

Furthermore, the complex interplay between viral mutations, host immune responses, and environmental factors can influence the emergence of new variants. While RNNs can provide valuable insights into viral sequence patterns, it is important to integrate these findings with other biological and epidemiological data to generate a comprehensive understanding of SARS-CoV-2 evolution.

5. Future Directions

This study lays the groundwork for further exploration of RNNs in the context of SARS-CoV-2 surveillance and prediction. Future research directions include:

Incorporating additional data: Integrating data on host immune responses, environmental factors, and geographical spread into the RNN models can potentially enhance their predictive capabilities.
Ensemble learning: Utilizing an ensemble of different RNN architectures can potentially improve the overall robustness and accuracy of the predictions.
Real-time monitoring: Developing real-time monitoring systems that continuously analyze new viral sequences and update predictions can provide valuable insights for public health interventions.
Investigating interspecies transmission: Utilizing RNNs to analyze viral sequences from diverse animal species can provide insights into the potential for zoonotic transmission and the emergence of novel SARS-CoV-2 variants.

The discussion of the results obtained in this study underscores the promising capability of recurrent neural network (RNN) models in predicting the future emergence of SARS-CoV-2 clades and lineages. The findings revealed significant correlations between the models' predictions and the observed real values in the validation dataset, indicating substantial predictive capacity of the models. However, it is important to acknowledge that the accuracy of RNN predictions is heavily reliant on the quality and quantity of the data used for training. As the pandemic progresses and more viral sequences become available, it will be crucial to continue updating and refining the models to maintain their predictive efficacy.

Furthermore, the analysis of host-specific variations in the viral data highlighted the importance of ongoing surveillance and understanding of SARS-CoV-2 adaptations across different mammalian species. The detection of mutations associated with viral adaptability in hosts such as domestic cats and gorillas suggests the potential for interspecies transmission events and underscores the need for a comprehensive approach to zoonotic surveillance and control. These findings emphasize the importance of integrating genomic, epidemiological, and ecological information for a comprehensive understanding of SARS-CoV-2 evolution and the dynamics of zoonotic spillover.

6. Conclusion

In conclusion, the utilization of recurrent neural networks (RNNs) represents a significant advancement in the analysis of SARS-CoV-2 genomic data. By harnessing the power of RNNs, researchers can accurately predict the emergence of new clades and lineages, providing critical insights into viral evolution and transmission dynamics. These predictive capabilities offer valuable opportunities for proactive public health measures, including vaccine development and targeted interventions, ultimately aiding in the global efforts to control and mitigate the impacts of the COVID-19 pandemic.

Moreover, as RNN models continue to evolve and incorporate additional data sources, such as host immune responses and environmental factors, their predictive accuracy and utility are expected to further improve. Collaborative efforts between data scientists, epidemiologists, and virologists will be essential in refining these models and translating their findings into actionable public health strategies. By investing in advanced predictive analytics and surveillance systems, policymakers and health authorities can better prepare for future outbreaks and mitigate the spread of emerging infectious diseases.

7. References

[1] Wang, Q., Qiu, Y., Guo, Y., Xu, Y., Zhang, S., Liu, S., ... & Jiang, S. (2021). SARS-CoV-2 evolution in domestic cats. Journal of General Virology, 102(12), 000892.

[2] Hofmann, H., Kleine-Weber, H., Krüger, N., Gnirss, M., Niemeyer, D., Drosten, C., & Pöhlmann, S. (2020). SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell, 181(2), 271-280.

[3] Shu, Y., McCauley, J., Huang, Y., Banerjee, A., & Grenfell, B. T. (2021). Identifying SARS-CoV-2

regional introductions and transmission clusters in real time. Virus Evolution, 7(1), veac048. https://academic.oup.com/ve/article/8/1/veac048/6609172

[4] Aurélien Géron (2019). Hands-On Machine Learning with Scikit-Learn & TensorFlow. O'Reilly Media.