Machine learning-based Forward Primer design for the detection of SARS-CoV-2 emerging variants

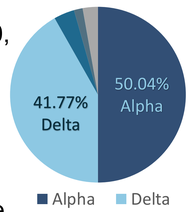
Hanyu Wang[a], Emmanuel Kagning Tsinda[b], Anthony Dunn[a], Alain B. Zemkoho[a]

[a] The University of Southampton, School of Mathematical Sciences, Southampton SO17 1BJ, United Kingdom
[b] Graduate School of Medicine, University of Tohoku, Sendai 980-8575, Japan

Introduction

Since the COVID-19 pandemic started in December 2019, over 220 million cases of SARS-CoV-2 have been reported cumulatively, and eleven different variants of the virus have also evolved.

The Alpha variant which predominated until March 2020, infected more than 50% of patients. Since April this year, the ultra-transmissible Delta variant has emerged, cumulatively infecting more than 40% of patients worldwide.



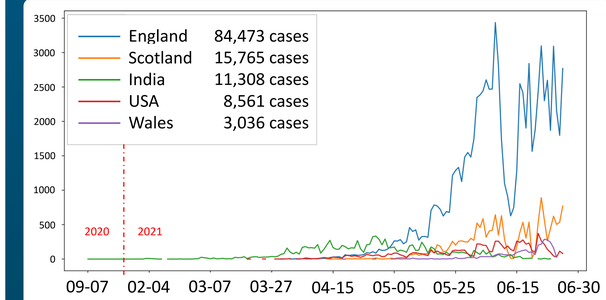
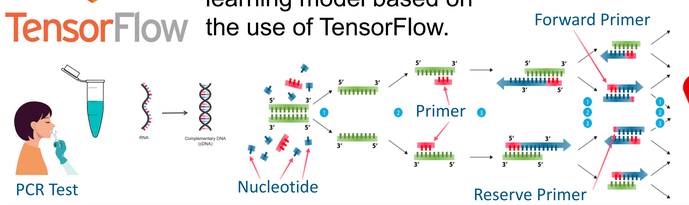


Figure 1. Top 5 regions with the fastest-growing Delta variants worldwide

Aim and methods

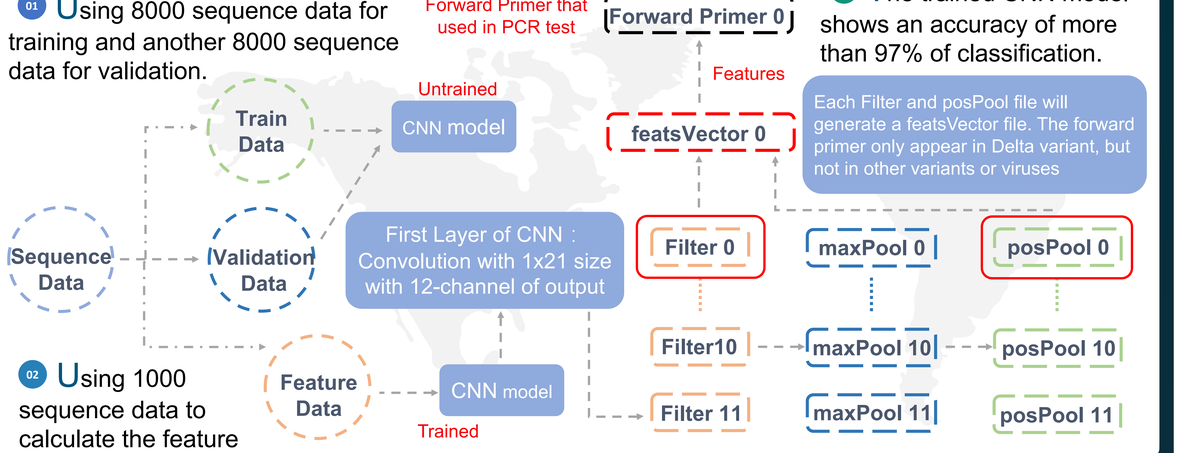
We aimed to use artificial intelligence techniques for the discovery of representative genomic sequences in SARS-CoV-2 Delta variant.

Fast and accurate **Forward Primer** design using CNN model, which is a deep learning model based on the use of TensorFlow.



01 Using 8000 sequence data for training and another 8000 sequence data for validation.

02 Using 1000 sequence data to calculate the feature



Experiments

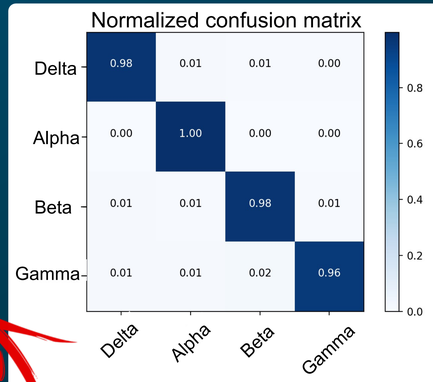


Figure 2. The visualization of the confusion matrix

03 The trained CNN model shows an accuracy of more than 97% of classification.

Results

The Figure 2. shows the trained CNN model works well. And the model with high accuracy of classification could efficiently identify the differences between variants and therefore extract features efficiently, contributing to a higher quality forward primer.

	feature_alpha	feature_beta	feature_gamma	feature_delta
CCGGTGGAATTGCTACCGCAA	0.0	0.001	0.0	0.998
TTGCTACCGCAATGGCTTGTC	0.0	0.001	0.0	0.997
CACCGGTGGAATTGCTACCGC	0.0	0.001	0.0	0.997
AATTGCTACCGCAATGGCTTG	0.0	0.001	0.0	0.997
GAATTGCTACCGCAATGGCTT	0.0	0.001	0.0	0.997
ACCGCAATGGCTTGTCTTGTA	0.0	0.001	0.0	0.997
ATTGCTACCGCAATGGCTTGT	0.0	0.0	0.0	0.996
GTGGAATTGCTACCGCAATGG	0.0	0.0	0.0	0.996
TGGAATTGCTACCGCAATGGC	0.0	0.0	0.0	0.996
TACCGCAATGGCTTGTCTTGT	0.001	0.0	0.0	0.996

Figure 3. The frequency of appearance of each Forward Primer in different variant virus (part)

I got a total of 32845 features. Then, I calculated the appearance frequency of each feature by using 5000 genetic sequences of each variant with different requests for appearance:

Delta >= 0.99 / Other <= 0.01

Then extend the range to other taxa (e.g., MERS-CoV and so on), and beta virus on others host (e.g., cat, dog, bat and so on)

Finally, from 32,845 features get several forward primer which satisfied with requests will be validated with the existing reserve primer by In-Silico PCR in FastPCR which works well. (Figure 4, in the bottom left)

In-Silico PCR

Genome: SARS-CoV-2 – Delta Variant

- Forward: 5'- GAAGGCCTTAAATTCCCTCGA -3'
  Position: 28412->28432 Tm = 55.9℃
- Reverse: 5'- TGTAGCACGATTGCAGCATTG -3'
  3'- CAATGCTGCAATCGTGCTACA -5'
  Position: 28687<-28707 Tm = 57.0℃
- The size of the amplified sequence: 296 bps
  Position: 28412-28707 Ta = 60.0℃

Figure 4. The result of the In-Silico PCR in the Fast PCR Software

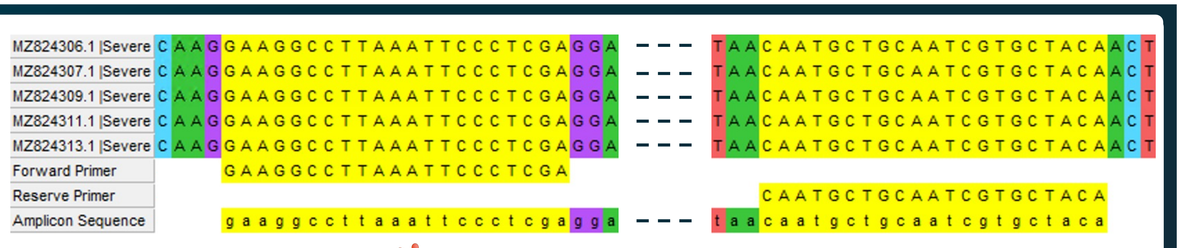


Figure 5. Using MEGA X software for visualizing the SARS-CoV-2 gene sequence and showing the forward primer, reserve primer and amplicon sequence

Acknowledgements

This work was supported by the CORMSIS External Summer Project of the University of Southampton, sponsored by Tohoku University. Also, many thanks to Adithya Madhusoodanan for their help during this work.

Contact:

Hanyu Wang : harrywang1997@outlook.com
Emmanuel Kagning Tsinda : kagningemmanuel2@med.tohoku.ac.jp
Anthony Dunn : ajd1g15@soton.ac.uk
Alain B. Zemkoho : a.b.zemkoho@soton.ac.uk

[1] Lopez-Rincon, A., Tonda, A., Mendoza-Maldonado, D.G., Molenkamp, R., Perez-Romero, Garssen, J. and Kraneveld, A.D., 2021. Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning. Scientific reports, 11(1), pp.1-11.
[2] GISAID (https://www.gisaid.org/)
[3] Johns Hopkins COVID-19 Map (https://coronavirus.jhu.edu/map.html)
[4] Kalendar, R., Khassenov, B., Ramankulov, Y., Samuilova, O. and Ivanov, K.I., 2017. FastPCR: An in silico tool for fast primer and probe design and advanced sequence analysis. Genomics, 109(3-4), pp.312-319.