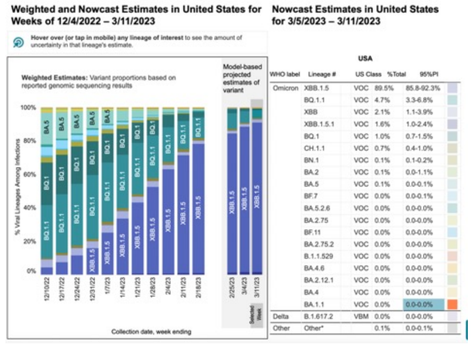
Classification and Primer Design for Different Variants of SARS-CoV-2 using AI

Shania Bu, Supervised by Dr. Melissa Ledgerwood-Lee
Radx Radical, University of California, San Diego

INTRODUCTION

According to the nowcast estimates provided by CDC, there are different and multiple Variants of Concern (VOC) of SARS-CoV-2 every week. For example, the nowcast estimates in United States for 3/5/2023-3/11/2023 stated that the top five VOCs are XBB.1.5, BQ.1.1, XBB, XBB.1.5.1, and BQ.1, which are all Omicron variants. Thus, a tool that could quickly obtain primers for different variants could be useful for qPCR.



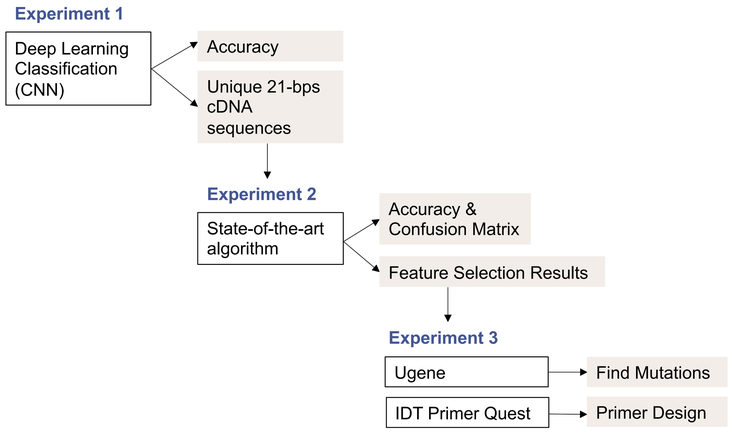
METHODS

I. Datasets data collected and curated from GISAID

	Dataset 1	Dataset 2	Dataset 3
Number of VOC	4	6	8
Variants of Concern (VOC)	Alpha, Beta, Delta*, Omicron	XBB.1.5*, Alpha, Beta, Delta, Gamma, Mu	XBB.1.5, 229E*, OC43*, Alpha, Beta, Delta, Gamma, Mu
Naming convention	WHO	WHO, Pango Lineage	WHO, Pango Lineage
Number of sequences	100 each variant	140 for XBB.1.5; 100 for others	140 for XBB.1.5; 100 for others
Total number	400	640	840

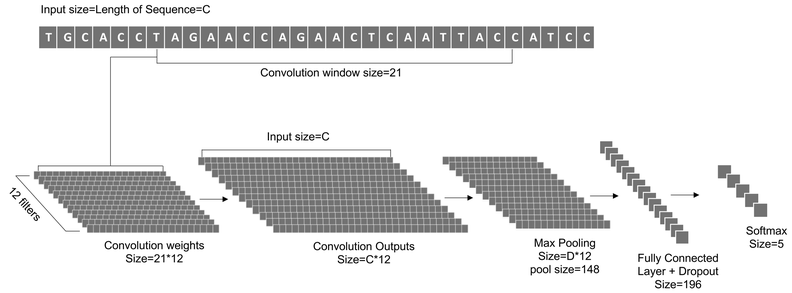
*most concerned variants

II. Experiment Logistics



III. Experiment 1 Using Deep Learning Classification CNN to get unique 21-bps sequences of different variants

RT-PCR tests use 21-bp sequences as primers. A CNN model with one convolutional layer, 12 filters, and max-pooling is used for three datasets. The highest output values in the max pooling layer identify 21-bp cDNA sequences. Unique sequences are analyzed, providing potential information for identifying different virus strains.



IV. Experiment 2 Running state-of-the-art algorithms to get most meaningful 21-bps sequences

After obtaining the unique 21-bps cDNA sequences from the CNN model, the state-of-the-art feature selection algorithms wan run to reduce the sequences needed to identify the virus strain of different variants to the bare minimum. The state-of-the-art feature selection algorithms include 8 simpler and more traditional classifiers: Gradient Boosting Classifier, Random Forest Classifier, Logistic Regression, Passive Aggressive Classifier, SGD Classifier, SVC(linear), Ridge Classifier, and Bagging Classifier, with 10-fold cross-validation. I obtain confusion matrix and accuracy of each algorithm.

V. Experiment 3 Finding Mutations using Ugene and Designing Primers using IDT Primer Quest

The most meaningful 21-bps sequences obtained from Experiment 2 are prospective candidates of forward primers. Primers for qPCR must follow the rules.

a) GC content in the 30–70% range.
b) The optimal primer length is 20 bases
c) Avoid runs of identical nucleotides. If repeats are present, there must be fewer than four consecutive G residues.
d) Make sure the last five nucleotides at the 3' end contain no more than two G and/or C bases.

The prospective candidates of forward primers of delta are also compared with other samples in lab. Mutations in the sequence are pointed out and annotated, using Ugene.

RESULTS

I. Dataset 1 Alpha, Beta, Delta, Omicron

- 100% Testing Accuracy, Validating Accuracy, Training Accuracy
- 21-bps unique sequences of Delta

index	Sequence	Alpha	Beta	Delta	Omicron
13	AATCTTAGAACCAGAACTCAA	0	0	1	0
14	ATCTTAGAACCAGAACTCAAT	0	0	1	0
15	CTTAGAACCAGAACTCAATTA	0	0	1	0
16	GTTAATCTTAGAACCAGAACT	0	0	1	0
17	TTAGAACCAGAACTCAATTAC	0	0	1	0

- The melting point, GC content, start/stop codon of forward primer candidates of index 13 and 14, and mutations in other variants.

Index	Start	Stop	Tm(°C)	GC%
13	21612	21633	57	33.3
14	21613	21634	57	33.3

Delta	ATCTTAGAACCAGAACTCAAT	Mutation count
Alpha	ATCTTACAACCAGAACTCAAT	1
Beta	ATTTTACAACCAGAACTCAAT	2
Gamma	ATTTTACAAACAGAACTCAAT	3
Mu	ATCTTACAACCAGAACTCAAT	1
Omicron_XBB.1.5	ATTTTGGGGACCAGGAACTAAT	N/A
Omicron_BA.1	ATCTTACAACCAGAACTCAAT	1
Omicron_BA.2.3	ATCTTATAACCAGAACTCAAT	1
Omicron_BA.2.12.1	ATCTTATAACCAGAACTCAAT	1
Omicron_BA.4.1	ATCTTATAACCAGAACTCAAT	1
Omicron_BA.5.1	ATCTTATAACCAGAACTCAAT	1

II. Dataset 2 XBB.1.5, Alpha, Beta, Delta, Gamma, Mu

XBB.1.5	ACCATGTCATACCAACATCAC
Alpha, Beta, Delta, Gamma, Mu	ACCATGTCATATCAACATCAC
XBB.1.5	CCCAGACCCATCAAGAATCCT
Alpha, Beta, Delta, Gamma, Mu	CCCAGATCCATCAAGAATCCT
XBB.1.5	CAAGAAACCAATTGAAACGAT
Alpha, Beta, Delta, Gamma, Mu	CAAGAAACCAACTGAAACGAT
XBB.1.5	GATATTACTAATTATTATGCG
Alpha, Beta, Delta, Gamma, Mu	GATATTACTAATTATTATGAG
XBB.1.5	ATTGCAACAATCCATGAGCCG
Alpha, Beta, Delta, Gamma, Mu	ATTGCAACAATCCATGAGCAG

III. Dataset 3 XBB.1.5, 229E, OC43, Alpha, Beta, Delta, Gamma, Mu

- Unique 21-bps sequences of 229E: TCATAACCTACCTGAATACAT, TGTAGGTACATTTGAAAGTGC, ATATTACACTGGCCAATTTTA, TGGTGATTTCGCACAAGGACC, CAAATGCAATGCTTAAGTGTG, TTAAAACGTAGTCTTAAATAT
- Unique 21-bps sequences of OC43: CTGTTACATATTCAAGATTAA, TTTGGACATGTTTATGATTTT, GGCTTACAGGTTGTAAGTATG, AAAGGGTATATTGCTAATATA, AAATATCGTATTCTAATTTTT

REFERENCES

[1] Integrated DNA Technologies (IDT), PrimerQuest™ Tool, https://rb.gy/gojnaz
[2] Lopez-Rincon, A., Tonda, A., Mendoza-Maldonado, L. et al. Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning. Sci Rep 11, 947 (2021).
[3] Lopez-Rincon, A.; Mendoza-Maldonado, L.; Martinez-Archundia, M.; Schönhuth, A.; Kraneveld, A.D.; Garssen, J.; Tonda, A. Machine Learning-Based Ensemble Recursive Feature Selection of Circulating miRNAs for Cancer Tumor Classification. Cancers 2020, 12, 1785. H
[4] Simsek M, Adnan H. Effect of single mismatches at 3'-end of primers on polymerase chain reaction. J Sci Res Med Sci. 2000;2(1):11-14.
[5] Thermofisher Scientific, "PowerTrack SYBR Green Master Mix USER GUIDE, Master mix with a two-dye tracking system for real-time PCR," applied biosystems, pp. 28-29, January 2020.

ACKONWLEDGEMENT

I have many thanks to Dr. Melissa Ledgerwood-Lee for providing me with this research idea, important knowledge and information about qPCR and making primers, and instructions about using IDT Primer Quest and Ugene. I am also grateful to UC San Diego Jupyterhub (Data Science) Platform for giving me GPU resources to train the machine learning model.