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

Shania Bu
Radx Radical
University of California, San Diego
La Jolla, United States
Supervised by Dr. Melissa Ledgerwood-Lee
March 2023
tbu@ucsd.edu

***Abstract*—In this research project, different variants of SARS-CoV-2 are classified using artificial intelligence techniques to get primers for qPCR. Dataset 1, including Alpha, Beta, Gamma, and Omicron, and Dataset 2, including BN.1, BQ.1.1, BQ.1, CH.1.1, XBB.1.5, and XBB, were trained. The CNN model was first used to get unique 21-bps sequences. Dataset 1 obtained 14683 sequences and 100% accuracy, and Dataset 2 obtained 25 sequences and 43.12% accuracy. The state-of-the-art algorithms were then used to characterize different features better and get prospective forward primers. Primers unique to the 4 variants in Dataset 1 and primers that could separate BQ.1.1 and BQ.1 from other variants in Dataset 2 were obtained. The reverse primers for the forward primer candidates of Dataset 1 were checked and generated using IDT Primer Quest. The mismatches of Delta forward primer in other variants were checked using Ugene. However, the primers still require validation for implementing qPCR.**

Keywords— CNN, primer design, SARS-CoV-2

I. 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 (Fig. 1.). Thus, a tool that could quickly obtain primers for different variants could be useful for qPCR.

Based on the CNN model and the state-of-the-art algorithms from the research that detected SARS-CoV-2 using deep learning[1], prospective forward primers could be generated. Using Ugene, the sequence of the primer is then compared with other variants in lab to find mutations. The reverse primer is designed using IDT Primer Quest.

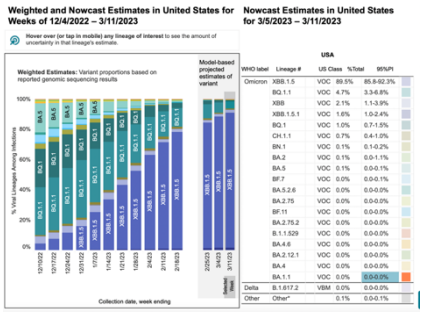

Fig. 1. Nowcast estimates for 3/5/2023-3/11/2023

II. Data Collection

A. Dataset 1

Dataset 1 contains 4 Variants of Concern (VOC) of SARS-CoV-2, including Alpha, Beta, Delta, and Omicron, and the variants were named based on WHO naming convention. There are 100 sequences for each variant, giving a total of 400 sequences in Dataset 1. The data were collected and curated from GISAID.

B. Dataset 2

Dataset 2 contains 6 major variants of Omicron, including BN.1, BQ.1.1, BQ.1, CH.1.1, XBB.1.5, and XBB, and these variants were named based on Pango Lineage naming convention. There are 140 sequences for each variant, giving a total of 840 sequences in Dataset 2. The data were collected and curated from GISAID.

III. Methods

Based on the code of covid_discriminatory_motifs on GitHub[1], experiment 1, using CNN model to get unique 21-bps cDNA sequences, and experiment 2, using state-of-the-art algorithm to characterize different features better, were conducted. After getting sequences that are prospective forward primers in experiment 2, experiment 3 is conducted to get mutations in other variants and design the reverse primers. Experiment 1-3 are implemented using both Dataset 1 and Dataset 2. The experiments are summarized in Fig. 2.

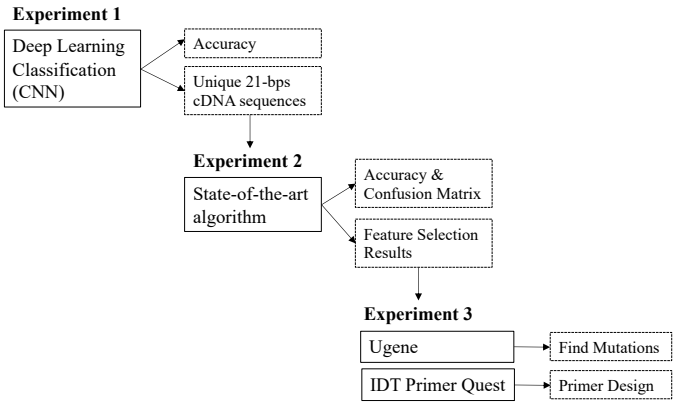

Fig. 2. Summary of the Experiments

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

- Since RT-PCR tests have a length of 18–22 bps normally, 21-bps sequences are designed to get the primers.
- The CNN used for both datasets is composed of one convolutional layer with 12 different filters that have different weights (each with window size 21, and an even padding of 10 steps on each side) with max-pooling (pool size 148 and stride 1), a fully connected layer (196 rectified linear units with dropout probability 0.5), and a final softmax layer with 5 units, to differentiate the different variants of SARS-CoV-2 strains. The batch size and number of iterations are different for two datasets. The values are recorded in Table I. Additionally, the CNN model was performed with 10-fold stratified cross-validation. Data divided into 80% training (8 folds), 10% validation (1 fold), 10% testing (1 fold).
- After obtaining the accuracy of CNN model, at this step, I identify one of the filters as the most promising that could correspond to meaningful cDNA sequences. Given this data, it is now possible to identify the 21-bps sequences that obtained the highest output values in the max pooling layer of the filter, in a section of 148 positions. This results in different numbers of max-pooling features (the value of sequence length divided by 148) in two datasets due to the length difference of sequences. The values of the number of max-pooling features and the length of the sequences in both datasets are also recorded in Table I. Each max-pooling feature identifies the 21-bps sequence that obtained the highest value from the convolutional filter.
- Analyzing the different sequence values appearing in the max pooling feature space, I get unique 21-bps cDNA sequences (different number of the unique sequences for two datasets, the value will be stated in Results), that can potentially be informative for identifying different virus strains.

TABLE I. THE VALUES USED FOR DATASET 1 AND 2

Parameters / Information of Data	Dataset 1	Dataset 2
A. Batch Size	40	70
B. Number of iterations	1000	2000
C. Length of sequences (maximum)	29903	1260
D. Number of max pooling features	203	9

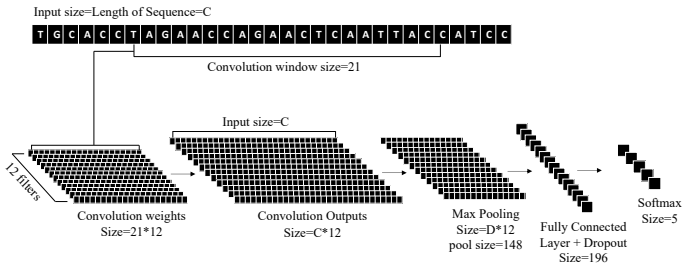

Fig. 3. Structure of CNN model (C, D are values in Table I.)

B. 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, I ran the state-of-the-art feature selection algorithms 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(n_estimators=300), Random Forest Classifier(n_estimators=300), Logistic Regression, Passive Aggressive Classifier, SGD Classifier, SVC(linear), Ridge Classifier, and Bagging Classifier(n_estimators=300), with 10-fold cross-validation. I obtain confusion matrix and accuracy of each algorithm.
- The classifiers were run using different numbers of 21-bps sequences to classify different variants. The values (listed in Table II.) are different for dataset 1 and 2 due to the difference in the length of sequences.

TABLE II. THE NUMBER OF SEQUENCES USED TO CLASSIFY VARIANTS FOR DATASET 1 AND 2

Information of Models	Dataset 1	Dataset 2
Number of runs	39	3
Values of the number of 21-bps squences used to classify different variants	14683, 11746, 9396, 7516, 6012, 4809, 3847, 3077, 2461, 1968, 1574, 1259, 1007, 805, 644, 515, 412, 329, 263, 210, 168, 134, 107, 85, 68, 54, 43, 34, 27, 21, 16, 12, 9, 7, 5, 4, 3, 2, 1	25, 20, 16

C. Experiment 3: Finding Mutations and Designing Primers using Ugene and 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 following 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.*
- I checked if the prospective candidates of forward primers follow the rules. With these forward primers, I located the forward primers, and I got the reverse primer and amplicon (optimal length 150 bps) for the SARS-CoV-2 samples in lab, using IDT Primer Quest.
- 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. The samples in lab include Alpha, Beta, Gamma, Mu, Omicron XBB.1.5, Omicron BA.1, Omicron BA.2.3, Omicron BA.2.12.1, Omicron BA.4.1, and Omicron BA.5.1.

IV. Results

The results for experiment 1, 2, and 3 for Dataset 1 and 2 are stated as follows.

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

The testing, validating, and training accuracy of the CNN model, and the total number of unique 21-bps sequences for two datasets are listed in Table III.

TABLE III. RESULTS OF CNN MODEL FOR DATASET 1 AND 2

Results	Dataset 1	Dataset 2
Testing Accuracy	100.0%	43.21%
Validating Accuracy	100.0%	50.00%
Training Accuracy	100.0%	50.00%
Number of unique 21-bps sequences obtained	14683	25

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

1) *Dataset 1*

a) Accuracy: The model obtained 100% accuracy of classifying the 4 variants, Alpha, Beta, Delta, and Omicron, of SARS-CoV-2.

b) Confusion Matrix: The confusion matrix of classifying 4 varinats, showing 100% accuaracy.

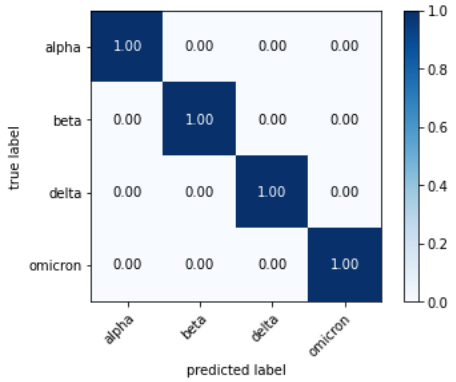

Fig. 4. Confusion matrix of Dataset 1

c) Most meaningful 21-bps sequences to separate 4 variants: This frequency matix (Table IV.) listed the most meaningful 21-bps sequences that the algorithms used to separate different variants. These sequences only appeared in one of the variants and appeared in all of the 100 sequences of the same variant, and they are results of classifying variants using 3 or 4 21-bps sequences.

TABLE IV. FREQUENCY MATRIX OF 21-BPS SEQUENCES OF FOUR VARIANTS (1=ALPHA, 2=BETA, 3=DELTA, 4=OMICRON)

index	21-bps sequences \ Variants	1	2	3	4
1	AATGTCTCTAAATGGACCCCA	1	0	0	0
2	ATAAATATTACAATTTGGTTT	1	0	0	0
3	ATGACACTACTGATGCTGTCC	1	0	0	0
4	ATGTAGAAAACCCTCATCTTA	1	0	0	0
5	ATTATTCAAACAATTGTTGAG	1	0	0	0
6	TGGTAATTTTAACAAAGACTT	1	0	0	0
7	AAAAATTGGAAATACCCACAA	0	1	0	0
8	ACACGAGTAACTCTTCTATCT	0	1	0	0
9	ACGAGTAACTCTTCTATCTTC	0	1	0	0
10	CGAGTAACTCTTCTATCTTCT	0	1	0	0
11	GTAACTCTTCTATCTTCTGCA	0	1	0	0
12	TTGGCAAAGAAATTTGACATC	0	1	0	0
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
18	TATTACTAATTATTATGCGGA	0	0	0	1

2) *Dataset 2*

a) Accuracy: obtained 34.52% accuracy of classifying the 6 variants, BN.1, BQ.1.1, BQ.1, CH.1.1, XBB.1.5, and XBB, of Omicron.

b) Confusion Matrix: The confusion matrix of classifying 6 variants.

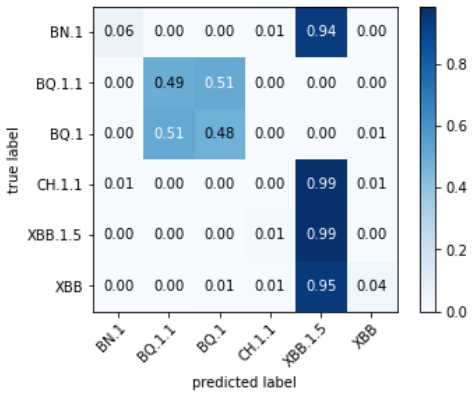

Fig. 5. Confusion matrix of Dataset 2

c) Most meaningful 21-bps sequences to separate 4 variants: Although the algorithm cannot classify the 6 variants with high accuracy, observersing the frequency matrix of the 21-bps sequences appeared in 6 variants, I found the sequences that can separate BQ.1.1 and BQ.1 from other variants (Table V.). The sequences appeared in all of the 140 sequences of the same variant.

TABLE V. FREQUENCY MATRIX OF 21-BPS SEQUENCES OF SIX VARIANTS OF OMICRON
(1=BN.1, 2=BQ.1.1, 3=BQ.1, 4=CH.1.1, 5=XBB.1.5, 6=XBB)

index	Variants / 21-bps sequences	1	2	3	4	5	6
19	AGGGAGCCTTGAATACACCAA	1	0	0	1	1	1
20	AATCAGCGAAATGCACTCCGC	1	0	0	1	1	1
21	AATCAGCGAAATGCACTTCGC	0	1	1	0	0	0

C. *Experiment 3: Finding Mutations and Designing Primers using Ugene and IDT Primer Quest*

1) *Primer Design of Alpha, Beta, Delta in Dataset 1 using IDT Primer Quest*

a) Alpha: Using IDT Primer Quest, I got the melting point, GC content, and start/stop codon of forward primer candidates. Three candidates that succesfully get the assay are sequences index 1, 3, and 4.

TABLE VI. INFORMATION OF FORWARD PRIMER FOR ALPHA

Index (Table IV.)	Start	Stop	Tm(°C)	GC%
1	28255	28276	61	42.9
3	23254	23275	62	47.6
4	15255	15276	57	33.3

Observing the start and stop codon, I found that index 1 and index 3 are the same with only one nucleotide difference. Index 4 is a separate primer. I used index 3 sequence (higher melting point and GC content than index 1) and index 4 sequence as the forward primers for the alpha sample in lab and obtained the reverse primer and amplicon. Details of the assays are listed in Table VII.

TABLE VII. REVERSE PRIMER AND AMPLICON FOR ALPHA (SAMPLE: >CONSENSUS_SEARCH-42976__E0000022__B01__210813_A00953_0374_AHKGLKDSX2__003.TRIMMED.SORTED.PILEUP.CONSENSUS_THRESHOLD_0.5_QUALITY_20)

Reverse primer (index 3)	**CAGGGACTTCTGTGCAGTTA**				
	Start	***Stop***	***Length***	***Tm(℃)***	***GC%***
	23390	23410	20	61	50
Amplicon	GTGATCCACAGACACTTGAGATTCTTGACAT TACACCATGTTCTTTTGGTGGTGTCAGTGTTA TAACACCAGGAACAAATACTTCTAACCAGGT TGCTGTTCTTTATCAGGGTGT (Length=156)				
Reverse primer (index 4)	**CGGTGTGACAAGCTACAA**				
	Start	***Stop***	***Length***	***Tm(℃)***	***GC%***
	15367	15385	18	59	50
Amplicon	TGGGTTGGGATTATCCTAAATGTGATAGAGC CATGCCTAACATGCTTAGAATTATGGCCTCA CTTGTTCTTGCTCGCAAACATACAACGTG (Length=130)				

b) Beta: Using IDT Primer Quest, I got the melting point, GC content, and start/stop codon of forward primer candidates. All six candidates, index 7-12, succesfully get the assay.

TABLE VIII. INFORMATION OF FORWARD PRIMER FOR ALPHA

Index (Table IV.)	Start	Stop	Tm(°C)	GC%
7	5226	5247	57	28.6
8	162	183	58	38.1
9	164	185	57	38.1
10	165	186	56	38.1
11	168	189	57	38.1
12	1041	1062	58	33.3

Observing the start and stop codon, I found that index 8-11 are the same with only a few nucleotides difference. In table, I listed Index 8, which has the highest melting point, and Index 10, which has the most optimal amplicon length. Index 7 and 12 are separate primers. I used index 7, 8, 10, and 12 as the forward primers for the beta sample in lab and obtained the reverse primer and amplicon. Details of the assays are listed in Table IX.

TABLE IX. REVERSE PRIMER AND AMPLICON FOR BETA (SAMPLE: >CONSENSUS_SEARCH-54366__E0001330__A01__211008_A00953_0420_BHV2FMDSX2__004.TRIMMED.SORTED.PILEUP.CONSENSUS_THRESHOLD_0.5_QUALITY_20)

Reverse primer (index 7)	**CAGCCCTTGCTCTGTAATAA**				
	Start	***Stop***	***Length***	***Tm(℃)***	***GC%***
	5360	5380	20	60	45
Amplicon	GTTAATGGTTTAACTTCTATTAAATGGGCAG ATAACAACTGTTATCTTGCCACTGCATTGTT AACACTCCAACAAATAGAGTTGAAGTTTAAT CCACCTGCTCTACAAGATGC (Length=154)				
Reverse primer (index 8)	**GACAAGGCTCTCCATCTTAC**				
	Start	***Stop***	***Length***	***Tm(℃)***	***GC%***
	262	282	20	60	50
Amplicon	TCTGCAGGCTGCTTACGGTCTCGTCCGTGTT GCAGCCGATCATCAGCACATCTAGGTTTTGT CCGGGTGTGACCGAAAG (Length=120)				
Reverse primer (index 10)	**TGAGTTGGACGTGTGTTT**				
	Start	***Stop***	***Length***	***Tm(℃)***	***GC%***
	298	316	18	59	44.4
Amplicon	GCAGGCTGCTTACGGTCTCGTCCGTGTTGCA GCCGATCATCAGCACATCTAGGTTTTGTCCG GGTGTGACCGAAAGGTAAGATGGAGAGCCT TGTCCCTGGTTTCAACGAGA (Length=151)				
Reverse primer (index 12)	**GGTGACGCAACTGGATAG**				
	Start	***Stop***	***Length***	***Tm(℃)***	***GC%***
	1175	1193	18	60	55.6
Amplicon	TTCAATGGGGAATGTCCAAATTTTGTATTTC CCTTAAATTCCATAATCAAGACTATTCAACC AAGGGTTGAAAAGAAAAAGCTTGATGGCTT TATGGGTAGAATTCGATCTGT (Length=152)				

c) Delta: Using IDT Primer Quest, I got the melting point, GC content, and start/stop codon of forward primer candidates. Two candidates that succesfully get the assay are sequences index 13 and 14.

TABLE X. INFORMATION OF FORWARD PRIMER FOR DELTA

Index (Table IV.)	Start	Stop	Tm(℃)	GC%
13	21612	21633	57	33.3
14	21613	21634	57	33.3

Observing the start and stop codon, I found that two sequences are the same with only one nucleotide difference. I used index 14 sequence as the forward primer for the delta sample in lab and obtained the reverse primer and amplicon. Details of the assay are listed in Table XI.

TABLE XI. REVERSE PRIMER AND AMPLICON FOR DELTA (SAMPLE: >CONSENSUS_SEARCH-42977__E0000022__A01__210813_A00953_0374_AHKGLKDSX2__003.TRIMMED.SORTED.PILEUP.CONSENSUS_THRESHOLD_0.5_QUALITY_20)

Reverse primer (index 14)	GTCCCAGAGACATGTATAGC				
	Start	*Stop*	*Length*	*Tm(℃)*	*GC%*
	21762	21782	20	59	50
Amplicon	TACCCCCTGCATACACTAATTCTTTCACACG TGGTGTTTATTACCCTGACAAAGTTTTCAGA TCCTCAGTTTTACATTCAACTCAGGACTTGTT CTTACCTTTCTTTTCCAATGTTACTTGGTTCC AT (Length=169)				

2) Finding mutations of the primer in different variants using Ugene

a) Delta: Using the prospective forward primer, index 14, ATCTTAGAACCAGAACTCAAT, I obtained the mutations of this sequence in other samples.

TABLE XII. MUTATIONS IN SAMPLES IN LAB

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

All samples except for Omicron_XBB.1.5 have mutation at the 7th nucleotide in the sequence (G→C or G→T), and Omicron_XBB.1.5 has more than 4 mutations. When binding to the target sequence, if the primer has 3' mismatch or more than 4 mutations, the primer cannot recognize the sequence and make amplicon. Therefore, I change the 21-bps sequence of delta to start from the mutation "G". GAACCAGAACTCAATTACCCC is the new sequence. I then verified it by checking this new sequence with other samples in lab.

TABLE XIII. MUTATIONS IN SAMPLES IN LAB USING NEW SEQUENCE (*DOES NOT HAVE A SIMILAR SEQEUNCE)

Delta	GAACCAGAACTCAATTACCCC	Mutation count
Alpha	CAACCAGAACTCAATTACCCC	1
Beta	CAACCAGAACTCAATTACCCC	1
Gamma	CAAACAGAACTCAATTACCCT	3
Mu	CAACCAGAACTCAATTACCCC	1
Omicron_XBB.1.5	N/A*	N/A
Omicron_BA.1	CAACCAGAACTCAATTACCCC	1
Omicron_BA.2.3	TAACCAGAACTCAATCATACA	5
Omicron_BA.2.12.1	TAACCAGAACTCAATCATACA	5
Omicron_BA.4.1	TAACCAGAACTCAATCATACA	5
Omicron_BA.5.1	TAACCAGAACTCAATCATACA	5

Since every sequence have mutation at the first nucleotide and some have more than 4 mutations, I obtained the reverse primer and the information of primers using IDT Primer Quest.

TABLE XIV. INFORMATION OF FORWARD AND REVERSE PRIMER AND AMPLICON FOR DELTA

Forward primer	GAACCAGAACTCAATTACCCC				
	Start	*Stop*	*Length*	*Tm(℃)*	*GC%*
	21619	21640	21	60	47.6
Reverse primer	TGGTCCCAGAGACATGTATAG				
	Start	*Stop*	*Length*	*Tm(℃)*	*GC%*
	21763	21784	21	60	47.6
Amplicon	CTGCATACACTAATTCTTTCACACGTGGTGT TTATTACCCTGACAAAGTTTTCAGATCCTCA GTTTTACATTCAACTCAGGACTTGTTCTTACC TTTCTTTTCCAATGTTACTTGGTTCCATG (Length=165)				

V. DISCUSSION

A. CNN Classification for Dataset 1 and Dataset 2

Since 100% testing accuracy was obtained after training the CNN model using Dataset 1, it can be concluded that this model can successfully classify the four variants of SARS-CoV-2, including Alpha, Beta, Delta, and Omicron. Therefore, the 14683 unique 21-bps sequences that the CNN model obtained can be used for experiment 2 to reduce the sequences needed to identify the different virus strains.

However, when using Dataset 2 to train the CNN model, only 43.21% accuracy was obtained. It can be concluded that it is difficult to separate the different variants of Omicron, BN.1, BQ.1.1, BQ.1, CH.1.1, XBB.1.5, and XBB, since the virus strains of these variants might only have slight difference.

B. Feature Selection using state-of-the-art algorithms for Dataset 1 and Dataset 2

According to the frequency matrix for Dataset 1 in Table IV., there are 6 21-bps sequences (index 1-6) that are unique to alpha, 6 for beta (index 7-12), 5 for delta (index 13-17), and

1 for omicron (index 18). Therefore, these 21-bps sequences can be forward primer candidates.

Although the 3 21-bps sequences in the frequency matrix for Dataset 2 in Table V. are not unique for all 6 variants, the 3 sequences can classify BN.1, CH.1.1, XBB.1.5, XBB from BQ.1.1 and BQ.1, since the first two sequences (index 19-20) only appeared in BN.1, CH.1.1, XBB.1.5, XBB and the last sequence only appeared in BQ.1.1 and BQ.1 (index 21). This result was possibly due to the fact that BQ.1.1 and BQ.1 have similar strains and can be separated from other variants.

C. *Designing Reverse Primers for Dataset 1 and Checking Mutations for Delta*

When designing reverse primers in IDT Primer Quest, the forward and reverse primers must follow the rules of primer. Some forward primer candidates in Table IV. cannot obtain a melting point that is high enough or cannot get a reverse primer that has similar melting point with an amplicon length smaller than 300. Thus, the forward primers that successfully got the assay listed in Table VI., VIII., and X. for alpha, beta, and omicron are not as many as the forward primer candidates in Table IV. Afterwards, when I observed the start and stop codon of the forward primers, it is obvious that some sequences are the same only with 1 or 2 shifts. For example, the first sequence (index 13) for delta starts on 21612 and the second sequence (index 14) starts on 21613: the two sequences are the same only with 1 shift. Therefore, the sequences with higher melting point, higher GC content, or better amplicon length (closest to 150) will be used to design the reverse primers.

However, if the primer has less than 4 mismatches, it is still possible to bind to the virus strain. Therefore, the location and number of mutations in variants other than delta are also checked by comparing the delta forward primer sequence (index 14) with other samples in lab. Although most of the samples do not have more than 4 mismatches, while observing the location of the mutations, I found that all samples have the mutation on the 7th nucleotide of the delta forward primer. Since primers cannot bind to the target sequence with 3' mismatch, I then shifted the delta forward primer for 6 nucleotides to get a new sequence. With the new sequence, the mutations in other samples in lab were also found and the reverse primer and amplicon for the new sequence were also obtained from IDT Primer Quest. However, whether the primers would work or not still need to be validated by future experiments in lab.

D. *Furture Directions*

1) *Dataset used to train the AI models*: There are some possible adjustments for the dataset used to train the CNN model and state-of-the-art algorithms to get different results or improve them. For example, changing Omicron in Dataset 1 to be labeled first to get more primer candidates for Omicron, and adding more variants such as XBB.1.5 to Dataset 1 to classify more variants with the 21-bps sequences.
2) *Validation of the Primers and Application*: The forward and reverse primers obtained in this research still need to be verified by using the primers in qPCR. If validated, the primers can be used for the mini-PCR machine project in lab.

VI. Conclusions

The CNN model and state-of-the-art algorithms are promising techniques that can classify different variants of SARS-CoV-2, since 21-bps primers unique to the four variants, Alpha, Beta, Delta, and Omicron are obtained. However, due to the rules and restrictions of designing primers, further checking and validation of the primers are still required in order to use the primers in qPCR.

Acknowledgment

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'd also like to recognize Numaan for providing me with WHO 1200 and WHO 1008 datasets that contains data collected and curated from GISAID. I am also grateful to UC San Diego Jupyterhub (Data Science) Platform for giving me GPU resources to train the machine learning model. Additionally, I had the pleasure of working and collaborating with Dr. Melissa Ledgerwood-Lee and other members in the team.

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.