RELATION BETWEEN
SARS-COV-2 AND OMICRON VARIANT IN BOTSWANA

By

Tarek Mohamed Abdallah
Mohab Hisham Mahmoud
Baraa Tarek Ewis
Mohamed Ayman Mohamed

Under the SuperVision of
Dr. Ibrahim Mohamed Youssef

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT

Table of Contents

Relation between SARS-Cov-2 and Omicron variant in Botswana

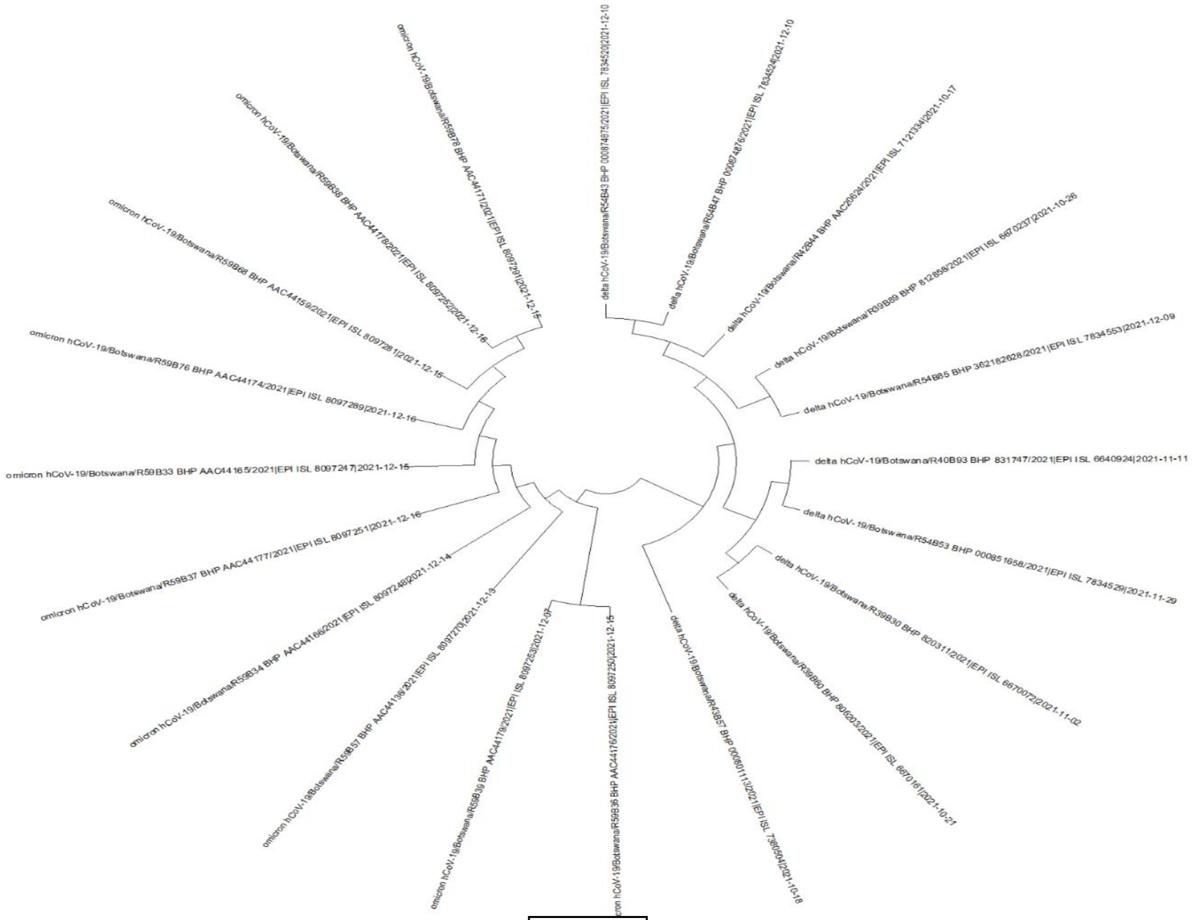
1.0 Introduction

We are trying to analyze SARS-Cov-2 and Omicron variant genes in Botswana, and find the similarities and differences between them, which will help us to understand how the virus is being mutated and what are the new points of strength and weakness of the new variant so we will know how to develop our vaccines to deal with omicron variant also, we will use many methods like phylogenetic trees and chemical constituents to spot the differences and mutations and we will make some conclusions based on our observations.

2.0 Methods

2.1 Phylogenetic Tree

A phylogenetic tree is a diagram that represents evolutionary relationships among organisms or genes from common ancestor, it helps us to know more about biological diversity and structuring classification , In trees, two genes are more related if they have a more recent common ancestor and less related if they have a less recent common ancestor, neighbor-joining is an algorithm used to reconstruct phylogenetic trees from distance data , it’s used to find neighbors based on distance at each stage of clustering, points with least distances are neighbors , we used the phylogenetic tree in our project to analyze the relation between delta and omicron by the help of an external software “ **MEGA** “ **.**



delta hCoV-19/Botswana/R54B43 BHP 000874875/2021|EPI ISL 7834520|2021-12-10
delta hCoV-19/Botswana/R54B47 BHP 000874876/2021|EPI ISL 7834524|2021-12-10
delta hCoV-19/Botswana/R42B44 BHP AAC20624/2021|EPI ISL 7121334|2021-10-17
delta hCoV-19/Botswana/R39B89 BHP 812858/2021|EPI ISL 6670237|2021-10-26
delta hCoV-19/Botswana/R54B85 BHP 362182628/2021|EPI ISL 7834553|2021-12-09
delta hCoV-19/Botswana/R40B93 BHP 831747/2021|EPI ISL 6640924|2021-11-11
delta hCoV-19/Botswana/R54B53 BHP 000851658/2021|EPI ISL 7834529|2021-11-29
delta hCoV-19/Botswana/R39B30 BHP 820311/2021|EPI ISL 6670072|2021-11-02
delta hCoV-19/Botswana/R39B60 BHP 806203/2021|EPI ISL 6670161|2021-10-21
delta hCoV-19/Botswana/R43B57 BHP 000801113/2021|EPI ISL 7380504|2021-10-18
cron hCoV-19/Botswana/R59B36 BHP AAC44176/2021|EPI ISL 8097250|2021-12-15
omicron hCoV-19/Botswana/R59B39 BHP AAC44179/2021|EPI ISL 8097253|2021-12-07
omicron hCoV-19/Botswana/R59B57 BHP AAC44136/2021|EPI ISL 8097270|2021-12-13
omicron hCoV-19/Botswana/R59B34 BHP AAC44166/2021|EPI ISL 8097248|2021-12-14
omicron hCoV-19/Botswana/R59B37 BHP AAC44177/2021|EPI ISL 8097251|2021-12-16
omicron hCoV-19/Botswana/R59B33 BHP AAC44165/2021|EPI ISL 8097247|2021-12-15
omicron hCoV-19/Botswana/R59B76 BHP AAC44174/2021|EPI ISL 8097289|2021-12-16
omicron hCoV-19/Botswana/R59B68 BHP AAC44159/2021|EPI ISL 8097281|2021-12-15
omicron hCoV-19/Botswana/R59B38 BHP AAC44178/2021|EPI ISL 8097252|2021-12-16
omicron hCoV-19/Botswana/R59B78 BHP AAC44171/2021|EPI ISL 8097291|2021-12-15

Figure 1

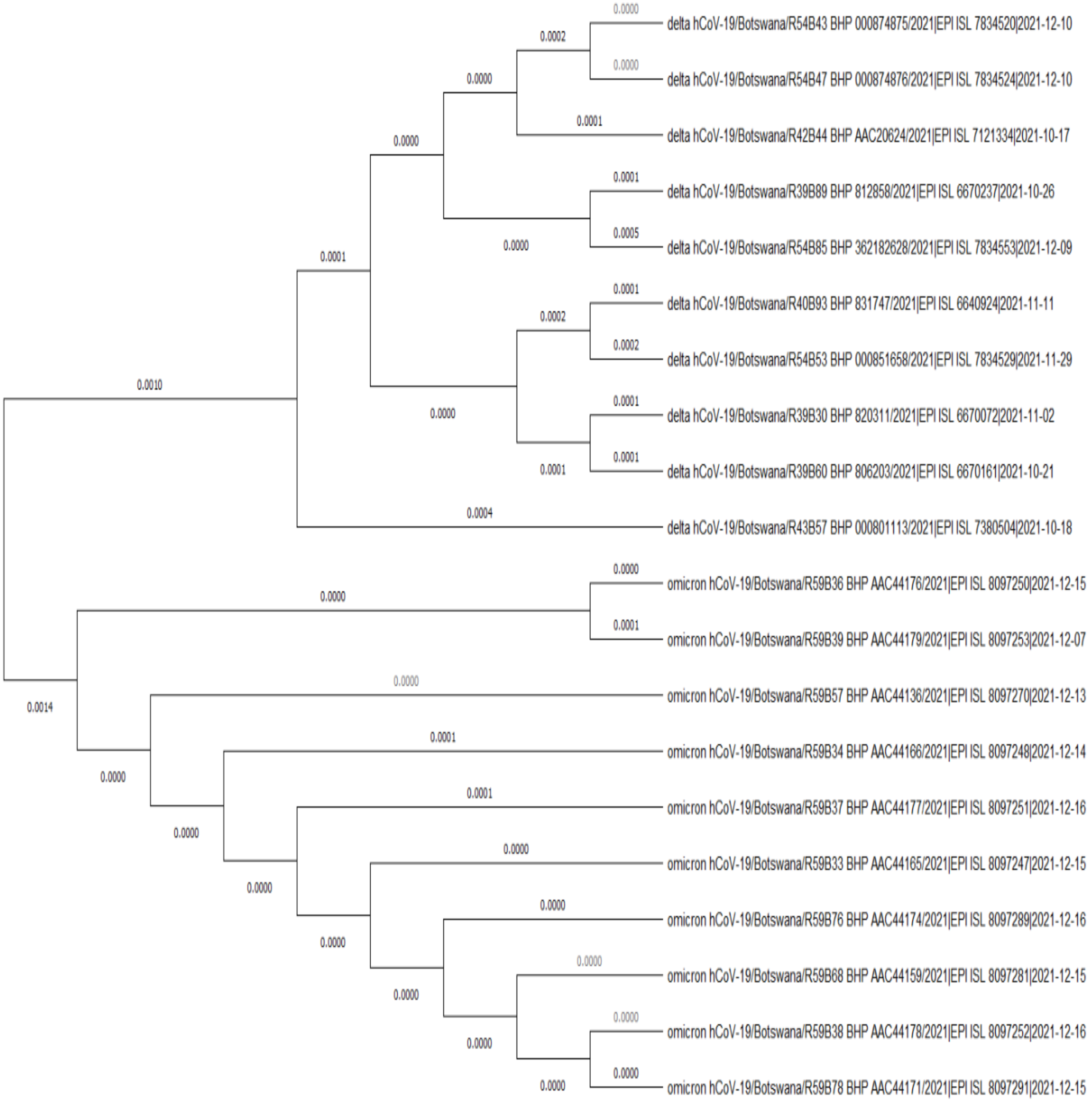

0.0000
0.0002
delta hCoV-19/Botswana/R54B43 BHP 000874875/2021|EPI ISL 7834520|2021-12-10
0.0000
0.0000
delta hCoV-19/Botswana/R54B47 BHP 000874876/2021|EPI ISL 7834524|2021-12-10
0.0001
0.0000
delta hCoV-19/Botswana/R42B44 BHP AAC20624/2021|EPI ISL 7121334|2021-10-17
0.0001
delta hCoV-19/Botswana/R39B89 BHP 812858/2021|EPI ISL 6670237|2021-10-26
0.0005
0.0000
delta hCoV-19/Botswana/R54B85 BHP 362182628/2021|EPI ISL 7834553|2021-12-09
0.0001
0.0001
delta hCoV-19/Botswana/R40B93 BHP 831747/2021|EPI ISL 6640924|2021-11-11
0.0002
0.0002
delta hCoV-19/Botswana/R54B53 BHP 000851658/2021|EPI ISL 7834529|2021-11-29
0.0010
0.0001
0.0000
delta hCoV-19/Botswana/R39B30 BHP 820311/2021|EPI ISL 6670072|2021-11-02
0.0001
0.0001
delta hCoV-19/Botswana/R39B60 BHP 806203/2021|EPI ISL 6670161|2021-10-21
0.0004
delta hCoV-19/Botswana/R43B57 BHP 000801113/2021|EPI ISL 7380504|2021-10-18
0.0000
omicron hCoV-19/Botswana/R59B36 BHP AAC44176/2021|EPI ISL 8097250|2021-12-15
0.0000
0.0001
omicron hCoV-19/Botswana/R59B39 BHP AAC44179/2021|EPI ISL 8097253|2021-12-07
0.0000
omicron hCoV-19/Botswana/R59B57 BHP AAC44136/2021|EPI ISL 8097270|2021-12-13
0.0014
0.0001
omicron hCoV-19/Botswana/R59B34 BHP AAC44166/2021|EPI ISL 8097248|2021-12-14
0.0000
0.0001
omicron hCoV-19/Botswana/R59B37 BHP AAC44177/2021|EPI ISL 8097251|2021-12-16
0.0000
0.0000
omicron hCoV-19/Botswana/R59B33 BHP AAC44165/2021|EPI ISL 8097247|2021-12-15
0.0000
0.0000
omicron hCoV-19/Botswana/R59B76 BHP AAC44174/2021|EPI ISL 8097289|2021-12-16
0.0000
0.0000
omicron hCoV-19/Botswana/R59B68 BHP AAC44159/2021|EPI ISL 8097281|2021-12-15
0.0000
0.0000
0.0000
omicron hCoV-19/Botswana/R59B38 BHP AAC44178/2021|EPI ISL 8097252|2021-12-16
0.0000
0.0000
omicron hCoV-19/Botswana/R59B78 BHP AAC44171/2021|EPI ISL 8097291|2021-12-15

Figure 2

Figure 1 and 2 shows the phylogenetic tree that we constructed between SARS-Cov-2 and Omicron

2.2 Consensus Sequence

A consensus sequence is a sequence of DNA, RNA, or protein that represents aligned, related sequences. The consensus sequence of the related sequences can be defined in different ways, but is normally defined by the most common nucleotide(s) or amino acid residue(s) at each position.

It serves as a simplified representation of the viral population. It represents the results of multiple sequence alignments in which related sequences are compared to each other.

In DNA molecules, consensus sequences are useful for representing short elements that are binding sites for other molecules. Such elements may be recognized by, for example, when considering sequence-dependent enzymes such as RNA polymerase. (**RNA polymerase**, is an enzyme that synthesizes RNA from a DNA template. Using the enzyme helicase, RNAP locally opens the double-stranded DNA so that one strand of the exposed nucleotides can be used as a template for the synthesis of RNA, a process called transcription.)

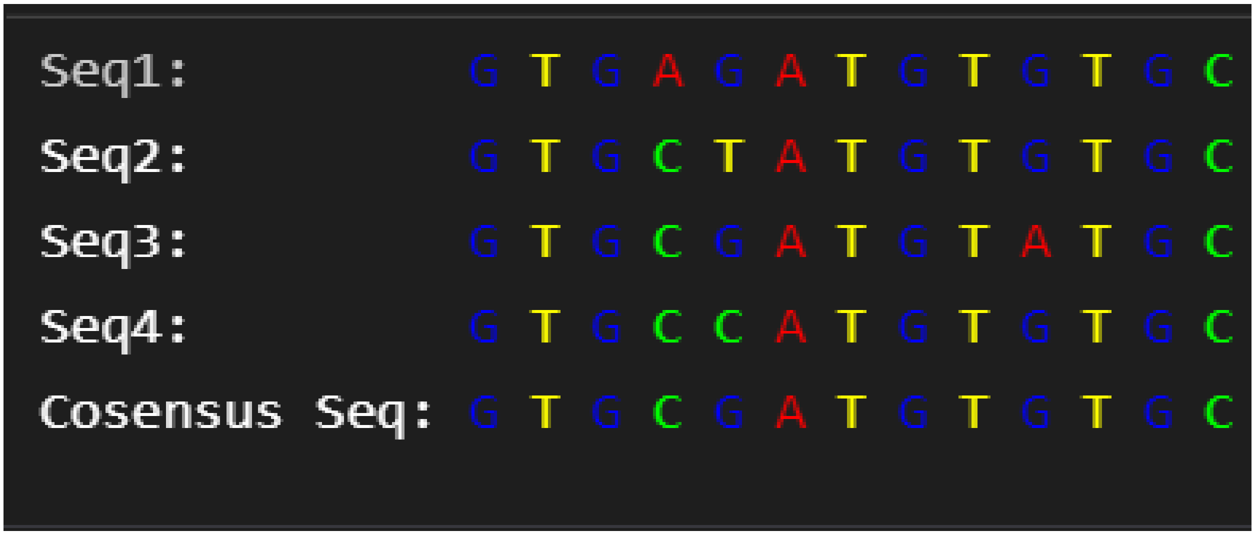
In proteins, consensus sequences may represent entire protein molecules or short fragments of them that correspond to conserved regions of importance for structure and function.

A protein binding site, represented by a consensus sequence, may be a short sequence of nucleotides which is found several times in the genome and is thought to play the same role in its different locations. For example, many transcription factors recognize particular patterns in the promoters of the genes they regulate. In the same way, restriction enzymes usually have palindormic consensus sequences, usually corresponding to the site where they cut the DNA.

Thus consensus sequence is defined as the idealized sequence that represents the predominant base at each position. All the actual examples shouldn't differ from the consensus by more than a few substitutions.

Any mutation allowing a mutated nucleotide in the core promoter sequence to look more like the consensus sequence is known as an **up mutation**. This kind of mutation will generally make the promoter stronger, and thus the RNA polymerase forms a tighter bind to the DNA it wishes to transcribe and transcription is up-regulated. On the contrary, mutations that destroy conserved nucleotides in the consensus sequence are known as **down mutations**. These types of mutations down-regulate transcription since RNA polymerase can no longer bind as tightly to the core promoter sequence.

Samples from Code output:



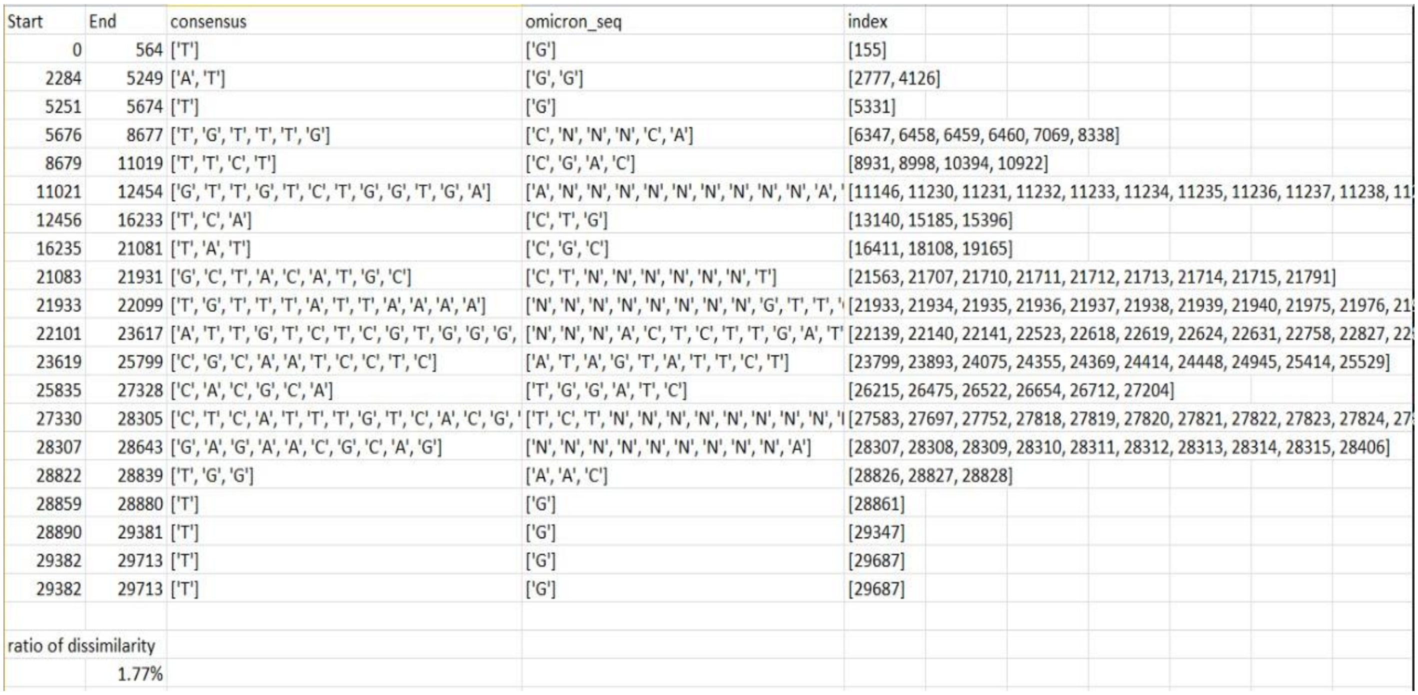
```
Seq1:          G T G A G A T G T G T G C
Seq2:          G T G C T A T G T G T G C
Seq3:          G T G C G A T G T A T G C
Seq4:          G T G C C A T G T G T G C
Cosensus Seq:  G T G C G A T G T G T G C
```

Figure 3

Shows an example for Consensus sequence formation from our code

2.3 Dissimilar Regions

We then used the output of the consensus sequence to compare it with our omicron sequences (as the consensus is a representative of SARS-Cov-2 sequences) and we spotted the differences between them which are our mutations in the original gene, we made a code script to analyze the differences and print them in a CSV file as shown in the figure, we compared the similar regions in the 10 omicron sequences with the consensus sequences at the same position and if any difference (mutation) is spotted, it will be printed as shown , we also calculated the ratio of dissimilarity which represents how many mutations happened in the similar regions as a percentage , [Start , End] represents the indices of similar omicron regions positions, and [index] represents the mutation index we also printed the mutation that has happened.



Start	End	consensus	omicron_seq	index
0	564	['T']	['G']	[155]
2284	5249	['A', 'T']	['G', 'G']	[2777, 4126]
5251	5674	['T']	['G']	[5331]
5676	8677	['T', 'G', 'T', 'T', 'T', 'G']	['C', 'N', 'N', 'N', 'C', 'A']	[6347, 6458, 6459, 6460, 7069, 8338]
8679	11019	['T', 'T', 'C', 'T']	['C', 'G', 'A', 'C']	[8931, 8998, 10394, 10922]
11021	12454	['G', 'T', 'T', 'G', 'T', 'C', 'T', 'G', 'G', 'T', 'G', 'A']	['A', 'N', 'N', 'N', 'N', 'N', 'N', 'N', 'N', 'N', 'A', '	[11146, 11230, 11231, 11232, 11233, 11234, 11235, 11236, 11237, 11238, 11
12456	16233	['T', 'C', 'A']	['C', 'T', 'G']	[13140, 15185, 15396]
16235	21081	['T', 'A', 'T']	['C', 'G', 'C']	[16411, 18108, 19165]
21083	21931	['G', 'C', 'T', 'A', 'C', 'A', 'T', 'G', 'C']	['C', 'T', 'N', 'N', 'N', 'N', 'N', 'N', 'T']	[21563, 21707, 21710, 21711, 21712, 21713, 21714, 21715, 21791]
21933	22099	['T', 'G', 'T', 'T', 'T', 'A', 'T', 'T', 'A', 'A', 'A', 'A']	['N', 'N', 'N', 'N', 'N', 'N', 'N', 'N', 'G', 'T', 'T', '	[21933, 21934, 21935, 21936, 21937, 21938, 21939, 21940, 21975, 21976, 21
22101	23617	['A', 'T', 'T', 'G', 'T', 'C', 'T', 'C', 'G', 'T', 'G', 'G', 'G',	['N', 'N', 'N', 'A', 'C', 'T', 'C', 'T', 'T', 'G', 'A', 'T'	[22139, 22140, 22141, 22523, 22618, 22619, 22624, 22631, 22758, 22827, 22
23619	25799	['C', 'G', 'C', 'A', 'A', 'T', 'C', 'C', 'T', 'C']	['A', 'T', 'A', 'G', 'T', 'A', 'T', 'T', 'C', 'T']	[23799, 23893, 24075, 24355, 24369, 24414, 24448, 24945, 25414, 25529]
25835	27328	['C', 'A', 'C', 'G', 'C', 'A']	['T', 'G', 'G', 'A', 'T', 'C']	[26215, 26475, 26522, 26654, 26712, 27204]
27330	28305	['C', 'T', 'C', 'A', 'T', 'T', 'T', 'G', 'T', 'C', 'A', 'C', 'G', '	['T', 'C', 'T', 'N', 'N', 'N', 'N', 'N', 'N', 'N', 'N', '	[27583, 27697, 27752, 27818, 27819, 27820, 27821, 27822, 27823, 27824, 27
28307	28643	['G', 'A', 'G', 'A', 'A', 'C', 'G', 'C', 'A', 'G']	['N', 'N', 'N', 'N', 'N', 'N', 'N', 'N', 'N', 'A']	[28307, 28308, 28309, 28310, 28311, 28312, 28313, 28314, 28315, 28406]
28822	28839	['T', 'G', 'G']	['A', 'A', 'C']	[28826, 28827, 28828]
28859	28880	['T']	['G']	[28861]
28890	29381	['T']	['G']	[29347]
29382	29713	['T']	['G']	[29687]
29382	29713	['T']	['G']	[29687]
ratio of dissimilarity				
	1.77%			

Figure 4

Shows the dissimilar regions output from our code

2.4 Constituent Percentages

Upon calculating the chemical constituents of variants we used a code that iterates over each sequence of the Omicron and calculates the sum of the calculated constituent in it. Then we calculated its percentage from the whole sequence. After doing so for all of the 10 sequences of the variant we calculated the average percentage of each constituent in Omicron SARS-Cov-2. The Percentages were calculated for delta SARS-Cov-2 using its consensus sequence only not iterating over the 10 sequences. First we used the code that returns the consensus then calculated the percentages in it. The previous steps were done using libraries in python such as bioPython, re and numpy. Then we used some data visualization tools such as pie and bar charts in matplotlib to make the comparison process a little bit easier. The Outputs of this method yielded that the average ratio of:

- C in delta hCoV-19 is 0.1832 while in its omicron variant its 0.1793
- G in delta hCoV-19 is 0.1961 while in its omicron variant its 0.1924
- T in delta hCoV-19 is 0.3216 while in its omicron variant its 0.3154
- A in delta hCoV-19 is 0.2989 while in its omicron variant its 0.2932
- CG in delta hCoV-19 is 0.3794 while in its omicron variant its 0.3718

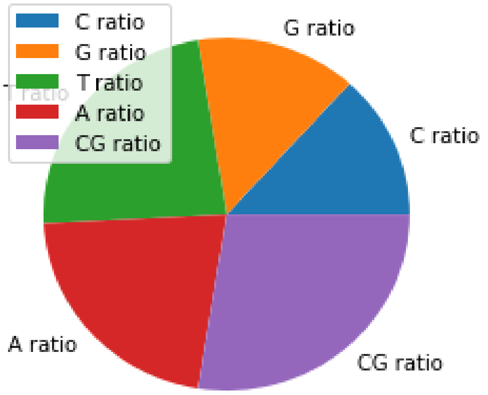


Figure 5

(Delta of hCoV-19)

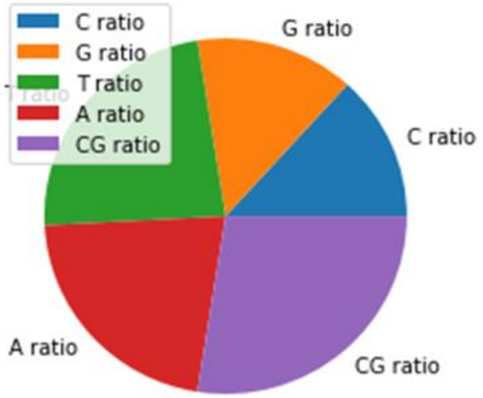


Figure 6

(Omicron of hCoV-19)

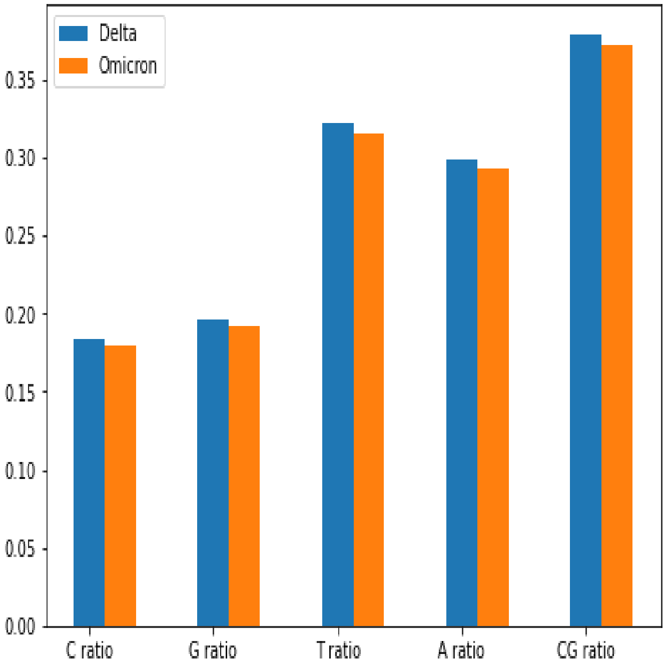


Figure 7

(Bar Chart Comparison)

3.0 Results and Conclusions

As we can see , delta sequence samples are very similar in their constructions and the distances between the neighbors is very small , but in omicron samples , they are approximately typical , as we see many branches are of length of zero which means that the neighbors are typical in their structure

We can also see that the mutations between delta and omicron isn't big and they are close to each other's in structure because the branches length is relatively small, which is a good indication for us as we can understand the structure and behavior of omicron based on our delta studies and also our vaccines and drugs developed for delta may deal with omicron also.

1. Ratio of dissimilarity is around 1.77% which means that the omicron Sequences are very similar to consensus sequence.
2. There are a lot of undefined nucleotides in the omicron sequences so there are some errors in our statistics.
3. The most mutated nucleotide is T and most of the time T nucleotide is mutated to G or undefined nucleotide in the omicron sequences.
4. From index 0 to 6000 and the last 1000 nucleotides the sequences are mostly typical and any mutated nucleotide in consensus is mutated to G.
5. Lot of mutations are to undefined nucleotides in omicrons, So we can't make sure if it's dissimilar or not.
6. There are lots of dissimilarities between each sequence of the 10 sequences of omicron.
7. The CG ratios fell with percentage of 0.76% in the Omicron variant resulting a dna with a slightly lower stability than the delta hCoV-19.
8. Most of the Chemical constituent's percentages in the Omicron variant decreased by a very small amount which may be due to the presence of lots of NaN values in the Omicron sequences
9. Mathematically there's a slight change in the percentages but biologically we can deduce that the 2 variants are very chemically close to each other as the change is insignificant.

4.0 References

1. https://en.wikipedia.org/wiki/Consensus_sequence
2. https://en.wikipedia.org/wiki/RNA_polymerase
3. https://www.sciencedirect.com/topics/medicine-and-dentistry/consensus-sequence
4. https://pubmed.ncbi.nlm.nih.gov/3447015/
5. https://www.nature.com/scitable/topicpage/reading-a-phylogenetic-tree-the-meaning-of-41956/