FINAL PROJECT COVID-19

Bioinformatics

DR. IBRAHIM YOUSEF

Team 4

FEBRUARY 6, 2023

Name	Section	B.N
osamah Faisal	1	11
Shuaib Abdulsalam	1	48
Gufran Mohammed	2	8
Yasmin Yasser Ali	2	52

Introduction

Severe acute respiratory syndrome coronavirus type 2 (SARS-CoV-2) is a virus that caused the COVID-19 disease outbreak in late 2019 in Wuhan China. In early 2020, the disease had rapidly spread around the whole world and was announced as a global pandemic. Out in late 2020 the delta variant was discovered in India, and due to the out-break many activities were limited which affected a wide variety of businesses and economies. And just as the corona outbreak began to subside, the new variant going by the name omicron came to exist which is called variant since it's actually a mutated version of corona. Consequently, trying to know the origins of these viruses and how they are related became a top priority if the world is going to go back to how it was. For that, we did this simplified research by selecting a specific number of sequences for SARS-Cov-2 delta variant. In addition, we chose 10 sequences for the SARS-Cov-2 Omicron variant in order to make a simple comparison between them and deduce the differences and mutations.

We have 10 sequences for hCoV-19 Delta variant in Belgium:

Delta1	hCoV-19/Belgium/rega-42191/2021|EPI_ISL_14673035|2021-12-14
Delta2	hCoV-19/Belgium/rega-42244/2021|EPI_ISL_14673060|2021-08-26
Delta3	hCoV-19/Belgium/rega-42262/2021|EPI_ISL_14673078|2021-08-18
Delta4	hCoV-19/Belgium/rega-42270/2021|EPI_ISL_14673086|2021-07-14
Delta5	hCoV-19/Belgium/rega-42271/2021|EPI_ISL_14673087|2021-08-20
Delta6	hCoV-19/Belgium/rega-42279/2021|EPI_ISL_14673095|2021-08-26
Delta7	hCoV-19/Belgium/rega-42288/2021|EPI_ISL_14673104|2021-08-26
Delta8	hCoV-19/Belgium/rega-42293/2022|EPI_ISL_14673109|2022-01-04
Delta9	hCoV-19/Belgium/rega-44886/2022|EPI_ISL_14732879|2022-08-01
Delta10	hCoV-19/Belgium/rega-44889/2022|EPI_ISL_14732881|2022-08-04

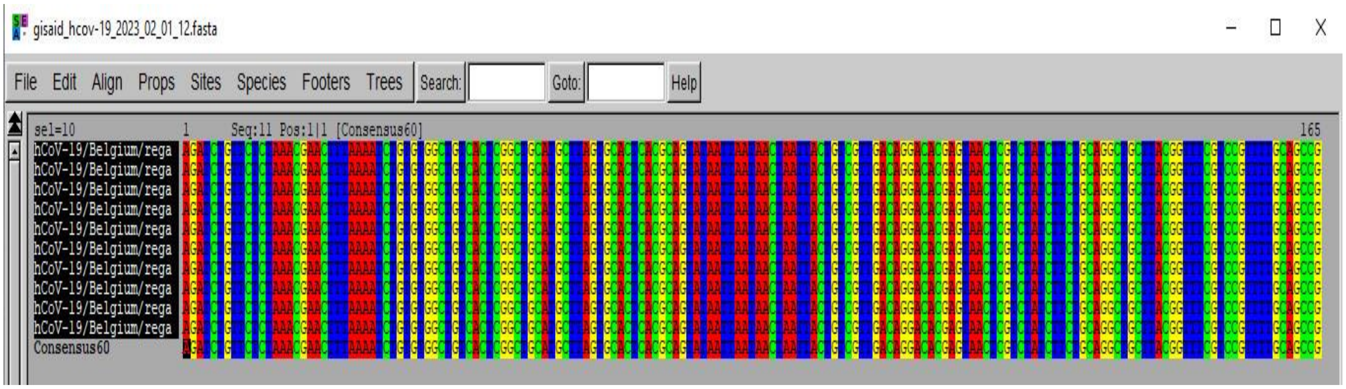
We have 10 sequences for hCoV-19 Omicron variant in Belgium:

Omicron1	hCoV-19/Belgium/UZA-UA-MI23022376/2023|EPI_ISL_16740361|2023-01-14
Omicron2	hCoV-19/Belgium/UZA-UA-MI23022449/2023|EPI_ISL_16740362|2023-01-14
Omicron3	hCoV-19/Belgium/UZA-UA-MI23030429/2023|EPI_ISL_16740363|2023-01-16
Omicron4	hCoV-19/Belgium/UZA-UA-MI23030472/2023|EPI_ISL_16740364|2023-01-17
Omicron5	hCoV-19/Belgium/UZA-UA-MI23031234/2023|EPI_ISL_16740365|2023-01-18
Omicron6	hCoV-19/Belgium/UZA-UA-MI23031323/2023|EPI_ISL_16740366|2023-01-18
Omicron7	hCoV-19/Belgium/UZA-UA-MI23032050/2023|EPI_ISL_16740367|2023-01-20

Omicron8	hCoV-19/Belgium/UZA-UA-MI23032253/2023|EPI_ISL_16740368|2023-01-20
Omicron9	hCoV-19/Belgium/UZA-UA-MI23040190/2023|EPI_ISL_16740369|2023-01-23
Omicron10	hCoV-19/Belgium/UZA-UA-MI23040195/2023|EPI_ISL_16740370|2023-01-23

Project Steps

- At First, we considered that hCoV-19 Delta variant sequences are the reference sequences. To Construct a consensus sequence from the reference sequences, we used the **Seaview** software (a multiplatform, graphical user interface for multiple sequence alignment and molecular phylogeny). The way of doing that is to get at each sequence location, the nucleotide/amino acid of the consensus sequence will be the most dominant one across all the sequences at that location.



So, the reference sequence is specified in a certain file.

Delta_consensus_sequence_reference.fa

- Then, we applied a multiple sequence alignment on hCoV-19 Omicron variant sequences (the case sequences).

We used the clustal omega technique to apply the multiple sequence alignment:

- Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences.

The link of the clustal omega alignment:

http://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?tool=clustalo&jobId=clustalo-E20230205-133925-0717-49179804-p1m

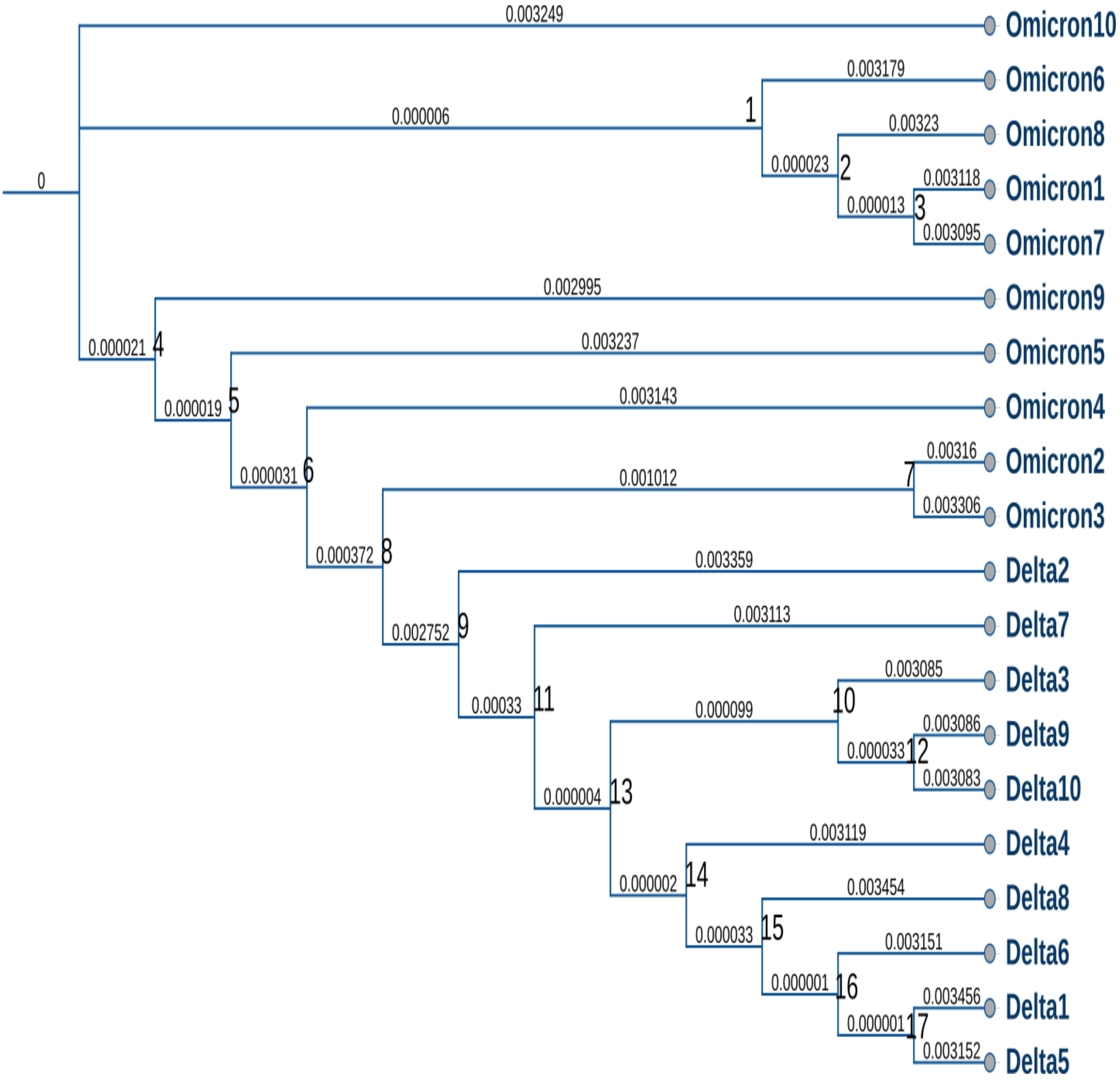
Phylogenetic Tree

- The second step was constructing a phylogenetic tree between all the above 20 sequences.

- In a phylogenetic tree, the species or groups of interest are found at the tips of lines referred to as the tree's **branches**.

- The pattern of connection between these branches helps us in understanding how the species in the tree evolved from a series of common ancestors.

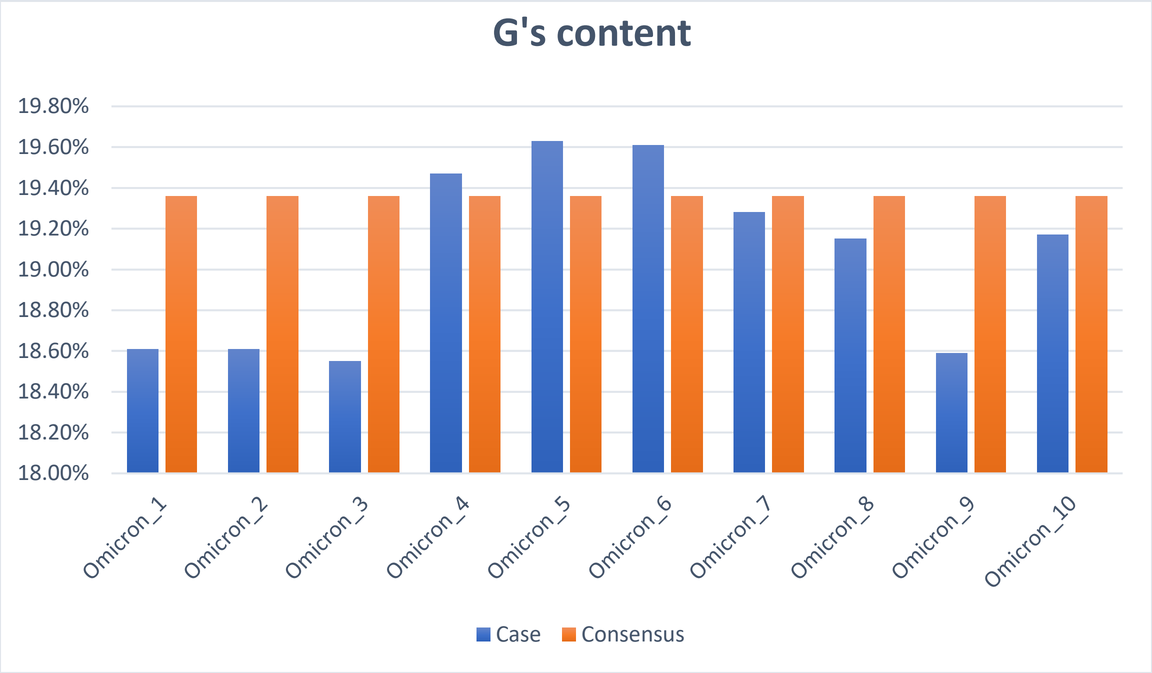
- Each internal node or point represents a separation event of a group into two groups.

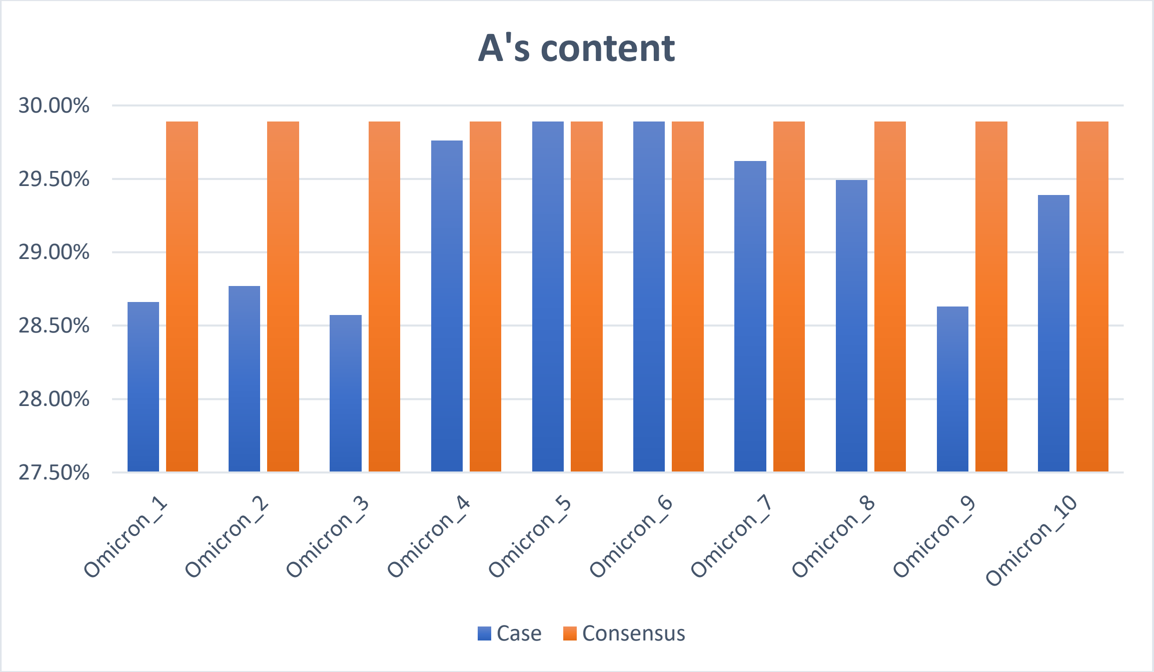

0.003249
Omicron10
0.003179
Omicron6
0.000006
1
0.00323
Omicron8
0
0.000023
2
0.003118
Omicron1
0.000013
3
0.003095
Omicron7
0.002995
Omicron9
0.000021
4
0.003237
Omicron5
0.000019
5
0.003143
Omicron4
0.00316
Omicron2
0.000031
6
7
0.001012
0.003306
Omicron3
0.000372
8
0.003359
Delta2
0.002752
9
0.003113
Delta7
0.003085
Delta3
0.00033
11
10
0.000099
0.003086
Delta9
0.000033
12
0.003083
Delta10
0.000004
13
0.003119
Delta4
0.000002
14
0.003454
Delta8
0.000033
15
0.003151
Delta6
0.000001
16
0.003456
Delta1
0.000001
17
0.003152
Delta5

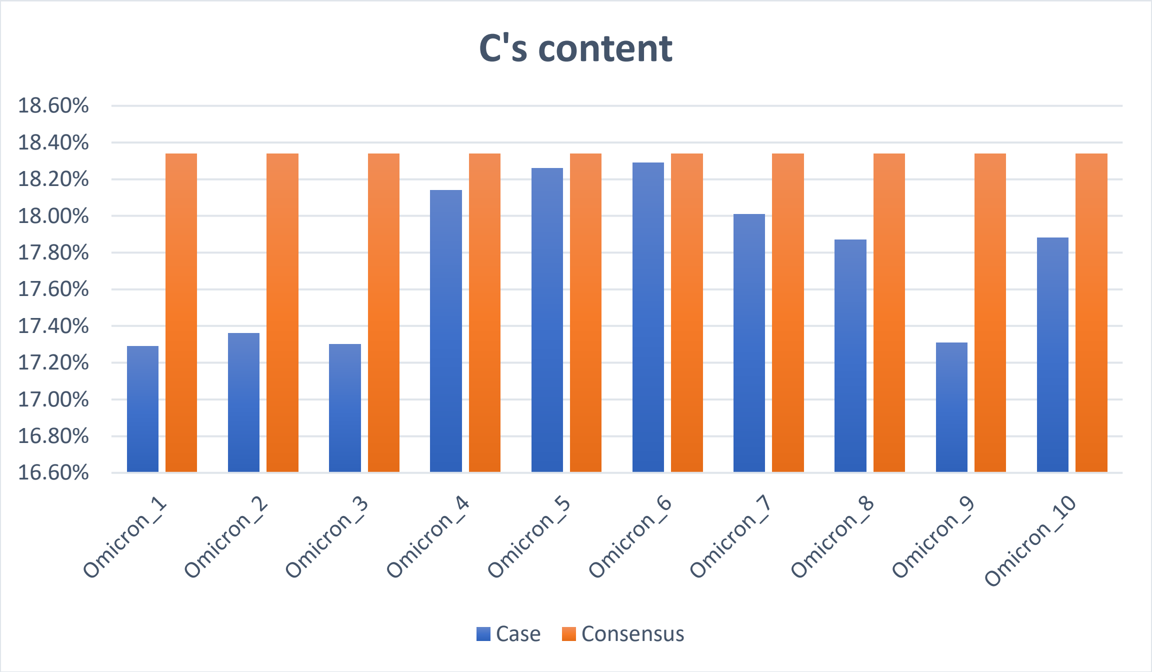
This table shows the correlation between each node and the other and the length of each branch:

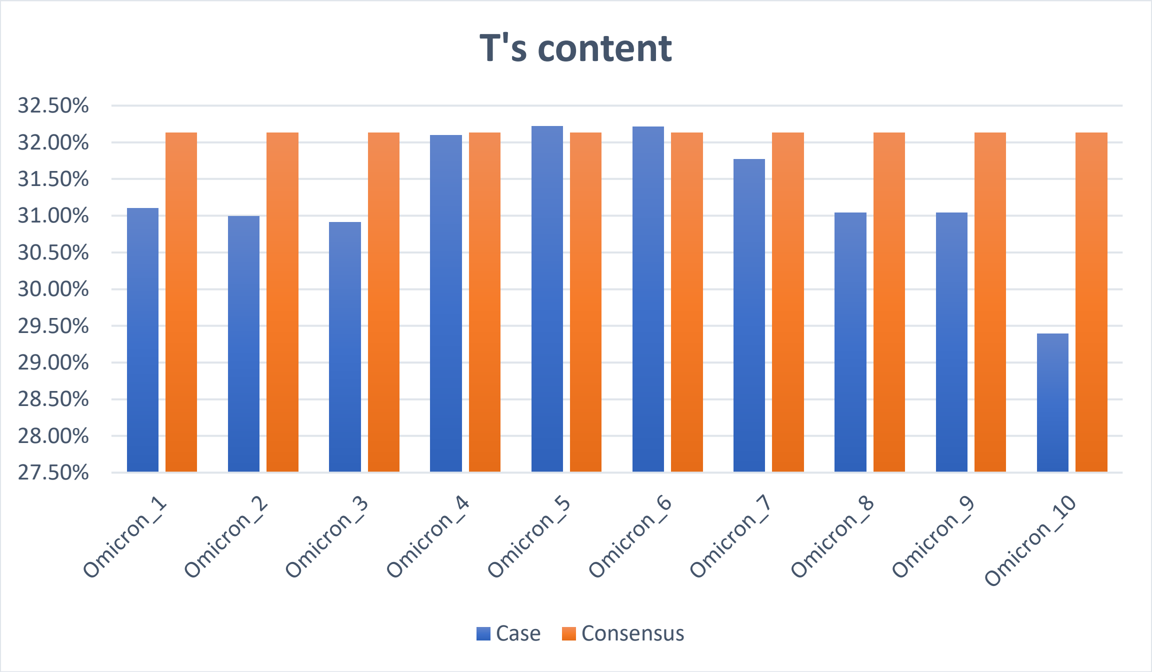
Node	Branch 1	Branch 2	Distance 1	Distance 2
1.	Omicron 6	2	0.003179	0.000023
2.	Omicron 8	3	0.00323	0.000013
3.	Omicron 1	Omicron 7	0.003118	0.003095
4.	Omicron 9	5	0.002995	0.000019
5.	Omicron 5	6	0.003237	0.000031
6.	Omicron 4	8	0.003143	0.000372
7.	Omicron 2	Omicron 3	0.00316	0.003306
8.	7	9	0.001012	0.002752
9.	Delta 2	11	0.003359	0.00033
10.	Delta 3	12	0.003085	0.000033
11.	Delta 7	13	0.003113	0.000004
12.	Delta 9	Delta 10	0.003086	0.003083
13.	10	14	0.000099	0.000002
14.	Delta 4	15	0.003119	0.000033
15.	Delta 8	16	0.003454	0.000001
16.	Delta 6	17	0.003151	0.000001
17.	Delta 1	Delta 5	0.003456	0.003152

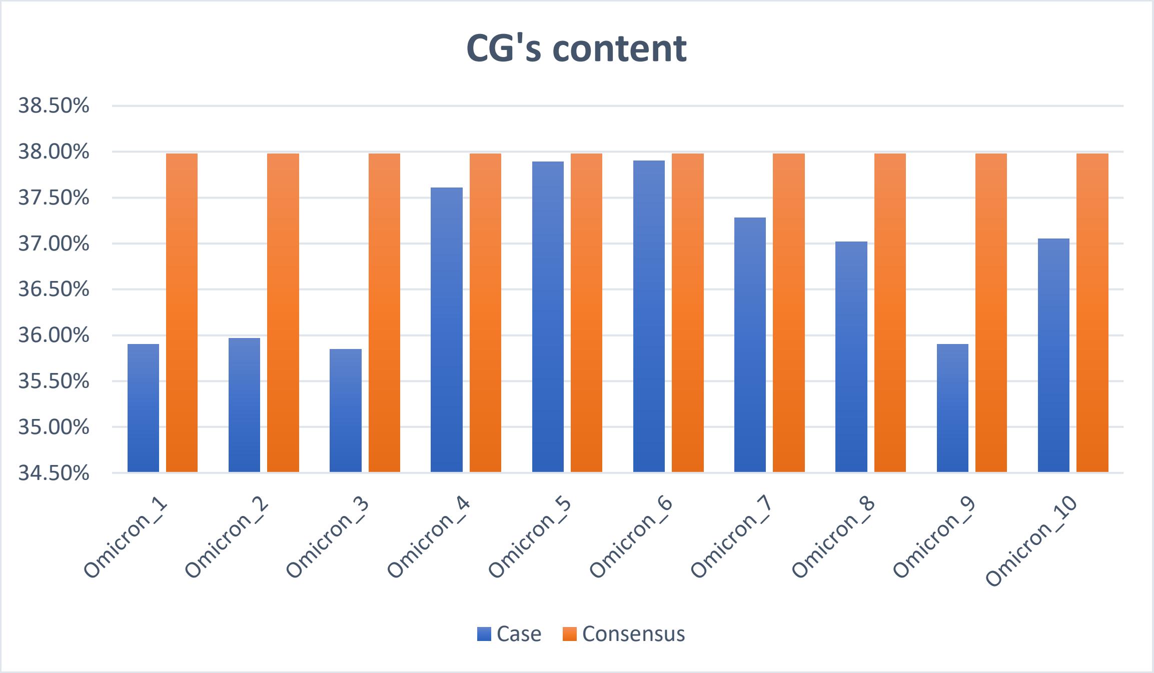
- Then, we found out The average percentage of the chemical constituents (C, G, T, and A) and the CG content between the reference sequences and the case sequences.






C's content
18.60%
18.40%
18.20%
18.00%
17.80%
17.60%
17.40%
17.20%
17.00%
16.80%
16.60%
Omicron_1
Omicron_2
Omicron_3
Omicron_4
Omicron_5
Omicron_6
Omicron_7
Omicron_8
Omicron_9
Omicron_10
🟦 Case
🟧 Consensus


T's content
32.50%
32.00%
31.50%
31.00%
30.50%
30.00%
29.50%
29.00%
28.50%
28.00%
27.50%
Omicron_1
Omicron_2
Omicron_3
Omicron_4
Omicron_5
Omicron_6
Omicron_7
Omicron_8
Omicron_9
Omicron_10
🟦 Case
🟧 Consensus



Code:

```
fastaSequences = SeqIO.parse(open("data/Omicron
.fasta"),'fasta')
for fasta in fastaSequences:
    name, sequence = fasta.id, str(fasta.seq)
    contan_of_A = sequence.count("A")
    contan_of_T = sequence.count("T")
    contan_of_C = sequence.count("C")
    contan_of_G = sequence.count("G")


    print(f"{name},length:{len(sequence)}\ncontan_of_A:{contan_
of_A}, contan_of_T:{contan_of_T}, contan_of_C:{contan_of_C},
contan_of_G:{contan_of_G}")
```

- After applying the multiple sequences alignment: we need to extract the dissimilar regions/columns between the alignment of the case sequences and the consensus sequence (the representative reference).

- The overall similarity regions between the alignment of the case sequences and the consensus sequence are: **27612** regions

- The overall dissimilarity regions between the alignment of the case sequences and the consensus sequence are: **2139** regions

The amount of similarity between each Omicron variant and case sequence variant is shown in the table below:

	# Of similar	Percentage%
Omicron1	28268	95.17%
Omicron2	28283	95.2%
Omicron3	28161	94.8%
Omicron4	29403	98.99%
Omicron5	29555	99.5%
Omicron6	29558	99.51%
Omicron7	29205	98.33%
Omicron8	29046	97.79%
Omicron9	28249	95.11%
Omicron10	29030	97.7%

The amount of dissimilarity between each Omicron variant and case sequence variant is shown in the table below:

	# Of dissimilar	Percentage%
Omicron1	1483	4.99%
Omicron2	1468	4.94%
Omicron3	1590	5.35%
Omicron4	348	1.17%
Omicron5	196	0.65%
Omicron6	193	0.64%
Omicron7	546	1.83%
Omicron8	705	2.37%
Omicron9	1502	5.05%
Omicron10	721	2.42%

The amount of similarity of each nucleotide between Omicron sequences and case sequence is shown in the table below:

	# Of Similar
A	8301
C	4970
G	5367
T	8974

The amount of similarity of each nucleotide between Omicron sequences and case sequence is shown in the table below:

	# Of Dissimilar
A	568
C	477
G	462
T	462

Code:

```
from Bio import  SeqIO
fasta_sequnece=SeqIO.parse(open("data/data_dissmilarty.fasta"),'fasta')
seqs=[]
similarities=0
dissimilarities=0
A_similarty=0 C_similarty=0 G_similarty=0 T_similarty=0 A_disimilarty=0
C_disimilarty=0 G_disimilarty=0 T_disimilarty=0

for fasta in fasta_sequnece:
    seqs.append(str(fasta.seq))

maximamlength=max(len(seqs[0]) , len(seqs[1]),len(seqs[2]),
len(seqs[3]),len(seqs[4]), len(seqs[5]),len(seqs[6]), len(seqs[7]),
len(seqs[8]),len(seqs[9]),len(seqs[10]))
minlength=min(len(seqs[0]) , len(seqs[1]),len(seqs[2]),
len(seqs[3]),len(seqs[4]), len(seqs[5]),len(seqs[6]), len(seqs[7]),
len(seqs[8]),len(seqs[9]),len(seqs[10]))

for i in range(maximamlength):
    try:
        if i<minlength :
            if seqs[0][i]==seqs[1][i]==seqs[2][i]==seqs[3][i]==
seqs[4][i]==seqs[5][i]==seqs[6][i]==seqs[7][i]==seqs[8][i]==seqs[9][i]==seqs[10][
i]:
                similarities+=1
# similarty nucl
                if seqs[10][i]=='A':
                    A_similarty+=1
                elif seqs[10][i]=='C':
                    C_similarty+=1
                elif seqs[10][i]=='G':
                    G_similarty+=1
                else:
                    T_similarty+=1
# disimilarty nucl
            else :
                dissimilarities+=1
                if seqs[10][i]=='A':
                    A_disimilarty+=1
                elif seqs[10][i]=='C':
                    C_disimilarty+=1
                elif seqs[10][i]=='G':
                    G_disimilarty+=1
                else:
                    T_disimilarty+=1
    except:
        pass
```

Conclusion:

- The similarity between Omicron's sequences is very close also the similarity between Delta's sequences is very close as we can see that in the phylogenetic tree.

- There is a great similarity between the delta virus and the Omicron mutant, as the similarity rate is approximately 93%.

- CG represent the stability of DNA, was low in the corona sequences that led to many variants of virus.

- The similarity between the consensus sequence and the individual Omicron is very close so we can say that Delta and Omicron are from the same family.

➢ Reference

- Clustal omega
  https://www.ebi.ac.uk/Tools/msa/clustalo/
- Itol
  https://itol.embl.de/
- Seaview
  https://doua.prabi.fr/software/seaview