SARS-Cov-2 Comparative Study

Name: Lamis Kamal Omara
Department: *Faculty of Engineering (System and biomedical department)*
Institution: *Cairo University*
Sec:2 BN:7

Name: Nada Nasr Ali
Department: *Faculty of Engineering (System and biomedical department)*
Institution: *Cairo University*
Sec:2 BN:36

Name: Shaimaa Mamdouh Ahmed
Department: *Faculty of Engineering (System and biomedical department)*
Institution: *Cairo University*
Sec:1 BN:42

Name: Suhaila Ahmed Bekhet
Department: *Faculty of Engineering (System and biomedical department)*
Institution: *Cairo University*
Sec:1 BN:4

Abstract--- The current and future situation globally according the ongoing persistent pandemic relies on the study of the behavior of the SARS-COV2 strains. In this study we have focused on the behavior of a two of the most resent and the most popular corona-virus strains (Omicron and Delta), this study is aiming at investigating the structure and the similarities between these two variants in order to get a sense of how these variants might behave relative to one another, also, to what to expect from the usage of the same drugs on both of them and how a new variant is more stable or unstable than the previous one.

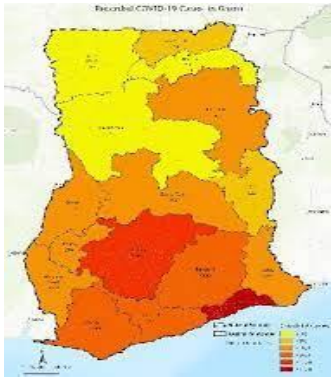
All the samples used where taken from Ghana.

I. INTRODUCTION

Viruses such as SARS-CoV-2 evolve throughout a time when changes in the genetic mutations arise during genome replication. A lineage is a group of virus variants that have a common ancestor and are genetically related. A variant of the SARS-CoV-2 virus has one or more mutations that distinguish it from other SARS-CoV-2 virus variants. The Delta variant of SARS-CoV-2, the virus that causes COVID-19, is a variant of SARS-CoV-2. It was discovered for the first time in India in late 2020. On May 31, 2021, and had spread to over 179 countries by 22 November 2021. The World Health Organization (WHO) indicated in June 2021 that the Delta variant was becoming the dominant strain globally. Another new variant called Omicron is the most recent variation as of December 2021. It was initially reported to the World Health Organization (WHO) on November 24, 2021, by South Africa. the WHO designated it as a variant of concern and named it "Omicron". Omicron is believed to be far more contagious (spreading much faster) than previous variants, spreading around 70 times faster in the bronchi (lung airways), but it is less able to penetrate deep lung tissue, which may explain why there is a significant reduction in the risk of severe disease requiring hospitalization. In this research, we made a comparative study between the two variants in Ghana by some methods of comparison and phylogenetic tree.

II. METHODS

- Preparations:
  From Ghana, the data of SARS-COV-2(COVID-19) and its variant Omicron were collected.
  And the data was available in GISAID which is a public database of SARS-Cov-2 sequences and its variant.



A. Choosing and manipulating the sequences

We chose 10 sequences from Delta and 10 from Omicron from Ghana by reading the Fasta files using Bio python library in Jupiter notebook and saving the new files containing 10 sequences each.
-We considered Delta the reference one, omicron the case one.

B. Building the Consensus Sequence

We made a multiple sequence alignment for the reference ones in order to construct the consensus sequence which we're going to use as a single representative for the reference sequences.

Alignment using muscle:
Alignment using → Mega Software for alignment (muscle), align by codon.

Constructing the sequence:
Then we construct the consensus sequence using python, we extracted the seqs from the Fasta file (type: string) then we added them into a list.

Then we used a library called collections, importing Counter () → a function returns the most frequent nucleotide at each index of the 10 sequences; thus, we get the consensus sequence.

C. MSL for Case Sequences (Omicron)

Then we applied a multiple seq alignment with the same technique for the case one (omicron sequence).

D. The Phylogenetic tree

- Generating a fasta file: we merged all the above 20 sequences together in a single fasta file using Bio python.
- MSL for all sequences: Then we applied a multiple sequence alignment on them using mega software, then we built the phylogenetic tree on them by neighbor joining using the same software

E. Average Percentage of Chemical Constituents

-We calculated the average percentage of the chemical constituents using:

Bio python→ reading the fasta file using → SeqIO.parse() and re. findall () → returns the length of each chemical constituent in the whole sequence

-For every sequence:
We used the re function to get the average:

$$\frac{Length\ of\ each\ nucleotide}{Length\ of\ sequence} \times 100$$

we got the avg percent value for each one, then we got the CG content:

$$\frac{Length(C) + Length(G)}{Length\ of\ sequence} \times 100$$

We've done this for the consensus / representative sequence for the reference ones and for all of the case sequences.

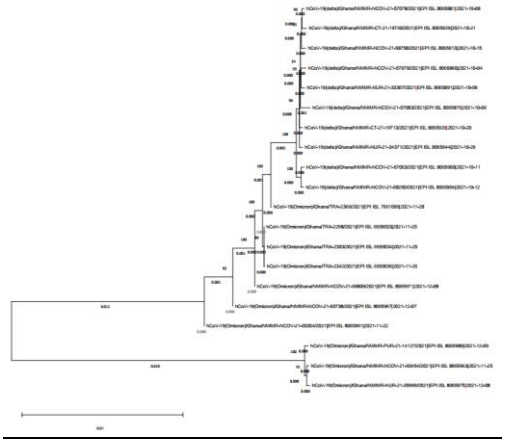
F. Extracting the dissimilar regions

- Apply MSL technique: We applied the multiple sequence alignment on the 10 aligned omicron sequences with the representative one, then we separated the omicron sequences from the representative one.

- Extracting the regions:
- With the help of Jupiter notebook, we iterated on the case sequences comparing the sequence at each index with each other and with the representative sequence at the same index.
- The dissimilar regions/columns state that at each index the case ones have the same or almost the same nucleotide but different from the representative one.
- We implemented a function which returns a dictionary containing the index at which the most dissimilarities occur. The function takes a third a parameter which is the percentage of the alignment the user can choose → A threshold percentage for the dissimilarities.
- We generated a csv excel file containing the index and the dissimilar region.

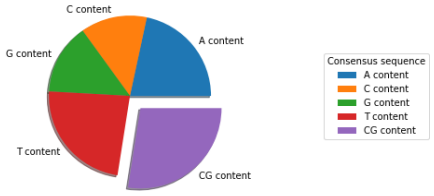
III. RESULTS AND DISCUSSION

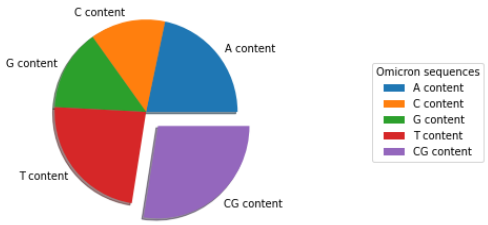
- From the Phylogenetic tree:



we figured out that the distance between sequences of omicron and delta in Ghana is very small, thus they are almost the same.
The figure shows the distance and the confidence percentage of the distance.

- From the Chemical Constituents:






```
the average A content in delta:  29.83000586186683
the average C content in delta:  18.29247267335609
the average G content in delta:  19.547601806834248
the average T content in delta:  32.28509361746147
the average CG content in delta:  37.84007448019034
```

```
the average A content in omicron:  29.51261297235775
the average C content in omicron:  18.074339322571362
the average G content in omicron:  19.37441362672501
the average T content in omicron:  31.779675412785707
the average CG content in omicron:  37.44875294929637
```

We found out here that a insignificant mutation occurs in Omicron sequences resulting in different average percent for each nucleotide.

- From The dissimilar regions:

	Index	Dissimilar Regions
0	155	(G, T)
1	1075	(G, N)
2	1076	(A, N)
3	1140	(T, N)
4	1141	(G, N)
...	...	...
690	29869	(T, -)
691	29872	(A, -)
692	29873	(G, -)
693	29874	(T, -)
694	29875	(G, -)

695 rows × 2 columns

When we extracted the regions, we noticed that it contains only 694 dissimilar regions/columns of nearly 30,000 nucleotides.

IV. Conclusion

It's clear from the constructed phylogenic tree between the sequences of the two variants that:

- The variant delta has gone a lot of mutations itself and it could manage to stay under the umbrella of delta, before it could finally mutate into omicron
- generally, the delta variants are closely related to each other as a group of sequences than to the family of Omicron
- Time plays a significant role in letting more and more mutations to occur and new variants to emerge

Also, from the average percentages of the chemicals constructing the two variants and their CG content, we can conclude that:

- The two variants have a very close percentages of each chemical constituent, maybe with a very slight difference in the T content

By analyzing the dissimilar regions it's clear that generally the two sequences are very similar, however, some parts are dissimilar that might potentially be a crucial part or might not be

V. Future Work

Throughout this paper a number of sequences that belongs to both delta and omicron where thoroughly analyzed, however the confidence in these results would be boosted if a larger set of data and more diverse were used in the future. Also, this paper hasn't deviated to a more thorough investigation of what the extracted dissimilar regions stand for or represent.

VI. Contribution Table

Name	Contribution
Suhaila Ahmed Bekhet	Equally contributed
Shaimaa Mamdouh Ahmed	Equally contributed
Lamis Kamal Omara	Equally contributed
Nada Nasr Ali	Equally contributed