A comparison between COVID-19 variants (Delta and Omicron)

Safwan Mahmoud, Sec: 1, BN: 43

Laila Hamdy, Sec: 2, BN: 8

Nouran FakhrEldin, Sec: 2, BN: 40

Youssef Shawki, Sec: 2, BN: 51

Under Supervision of:

Dr. Ibrahim Mohamed Youssef, PhD

Abstract

In this paper, we are examining the similarities and differences between Delta and Omicron variants of SARS-COV-2 virus. In order to do that, we have constructed a consensus sequence for 10 sequences of SARS-COV-2 Delta variant, called the reference sequence, and applied multiple sequence alignment for our 10 sequences of SARS-COV-2 Omicron variant being the case sequences, all the sequences were obtained from Ghana. We have executed several methods for comparison: Constructed a phylogenetic tree, calculated average percentage of the chemical constituents (C, G, T, and A) and the CG content and extracted the dissimilar regions between the aligned omicron sequences and the consensus sequence. The results of the comparison show that there are indeed similarities between the two variants.

Keywords: SARS-COV-2, Delta, Omicron

A comparison between COVID-19 variants (Delta and Omicron(

Introduction

Since the start of COVID-19 pandemic era, the mentioned virus has undergone multiple mutations as it spreads across countries worldwide, which led to emergence of variants with new characteristics other than the first variant. However, the most concerning variants are considered to be Delta and Omicron, with the latter being the last variant to be discovered. The common factor between these two variants which puts them on top of the danger scale is the spreading rate, which is being rising to its spike recently, thanks to the Omicron variant that was discovered first in Botswana and South Africa.

It is necessary to utilize techniques of bioinformatics such as sequence alignment and phylogenetic trees to differentiate between Delta and Omicron in terms of genetics, and hence, capture an idea on the current situation scientifically and how to deal with it.

Methods

In order to acquire a general overview on the difference between Delta and Omicron through construction of a phylogenetic tree as well as inspecting the chemical constituents of the gathered set of nucleotide sequences, an optimized mixture of coding as well as software techniques were involved in such a process.

MEGA

MEGA (Molecular Evolutionary Genetics Analysis), is a computer software released initially in 1993 by Pennsylvania State University. MEGA is used for conducting statistical analysis of molecular evolution as well as construction of phylogenetic trees.

Biopython

Biopython is a set of open-source tools written in Python by international team of developers used in biological computation. It is considered as the most popular python library in terms of serving the bioinformatics field.

Procedure

The data used in the analysis is 10 sequences of SARS-COV-2 Delta variant and 10 sequences of SARS-COV-2 Omicron variant, both of which are obtained from Ghana.

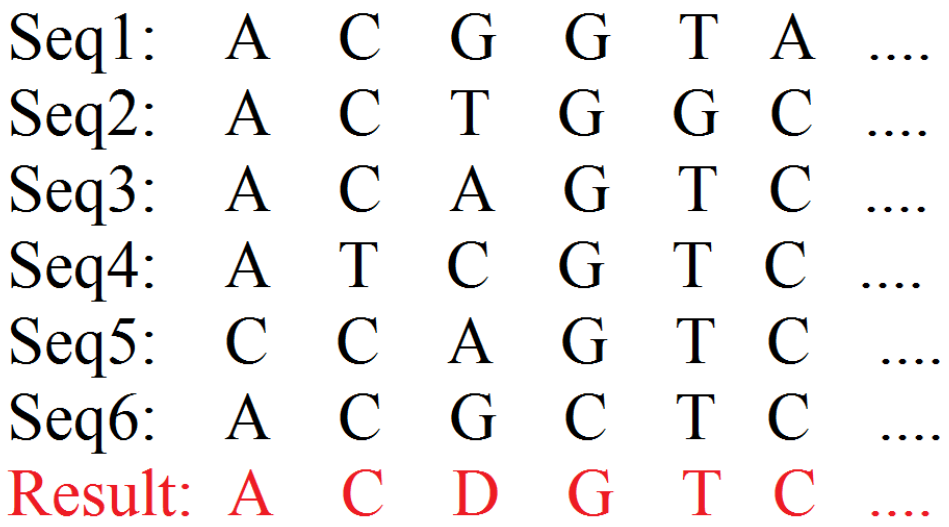
Initially, MEGA was used to apply MSA (Multiple Sequence Alignment) to the Delta variant sequences (Reference Sequences), as well as Omicron variant sequences (Case Sequences), which leads to the next step of constructing a representative sequence from the alignment result of the reference sequences which is called “Consensus Sequence”.

The consensus sequence is acquired through a python code which takes the output of the MSA of reference sequences and computes the most dominant nucleotide across the sequences at a specific location, and concatenating them together eventually.

Next, MEGA was used again to construct a phylogenetic tree between all the 20 sequences, which will be later used in discussion of the results of the analysis.

For the sake of having a deeper view for the data as well as obtaining more evidence on the genetic difference between Delta and Omicron, chemical constituents (C, G, A and T) of the 20 sequences were computed in average, exhibited and analyzed with the help of the Python library Pandas and Microsoft Excel.

Lastly, the case sequences were aligned with the consensus sequence using MEGA, then the output was passed to a python code which computed the dissimilar regions between them. In order to enable us to compare between the 10 case sequences and the consensus sequence, we had to construct a sequence which should represent the case sequences to ease the comparison. This representative sequence was constructed through checking for a minimum threshold of 70% of similarity between nucleotides of the case sequences in each column in order to be taken into consideration during comparison with the consensus sequence. If this threshold is met, the dominant nucleotide is placed in the resulting representative at the same location, if the threshold is not met, 'D' is placed, which stands for "don't care", which wouldn't be taken in consideration in the process of extracting the dissimilar regions. This assumption might be considered a compromise between the accuracy of the output and the ease of coding. An example for how the function works is shown in the figure below, where the sequences are random for the sake of illustration, and the result is considered their representative sequence.



Example for the representative sequence at 70% threshold.

After getting the representative of the case sequences, the resultant sequence was passed to another function that extracts the dissimilar regions/columns, where it returns the dissimilar region in both the case representative and the reference sequence and the indices of the start and end of the dissimilar region. The dissimilar regions were then exported in an excel sheet with the format shown in the figure below. 230 dissimilar regions/columns were found.

	A	B	C	D	E	F
		Staring Index	Ending Index	Region in Omicron	Region in Consensus	Length
0		155	155	G	T	1
1		1140	1209	TGAATGCAACCAAATGTGCCTTTCAACTCT	----------	70

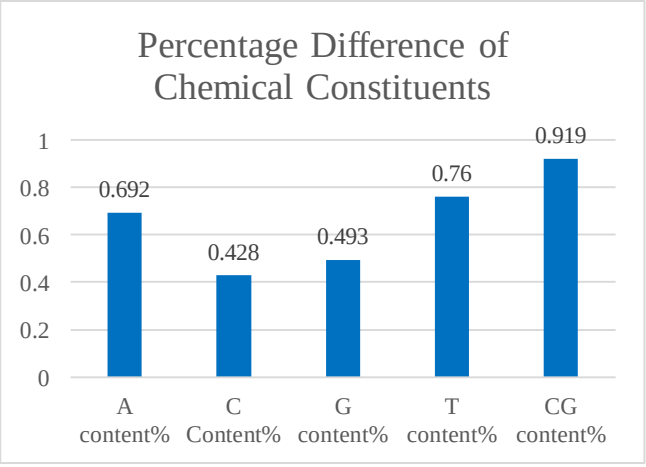
Results and Discussion

In terms of comparison, the output data from the techniques mentioned in the previous section was collected and analyzed, and to be discussed below.

Chemical Constituents

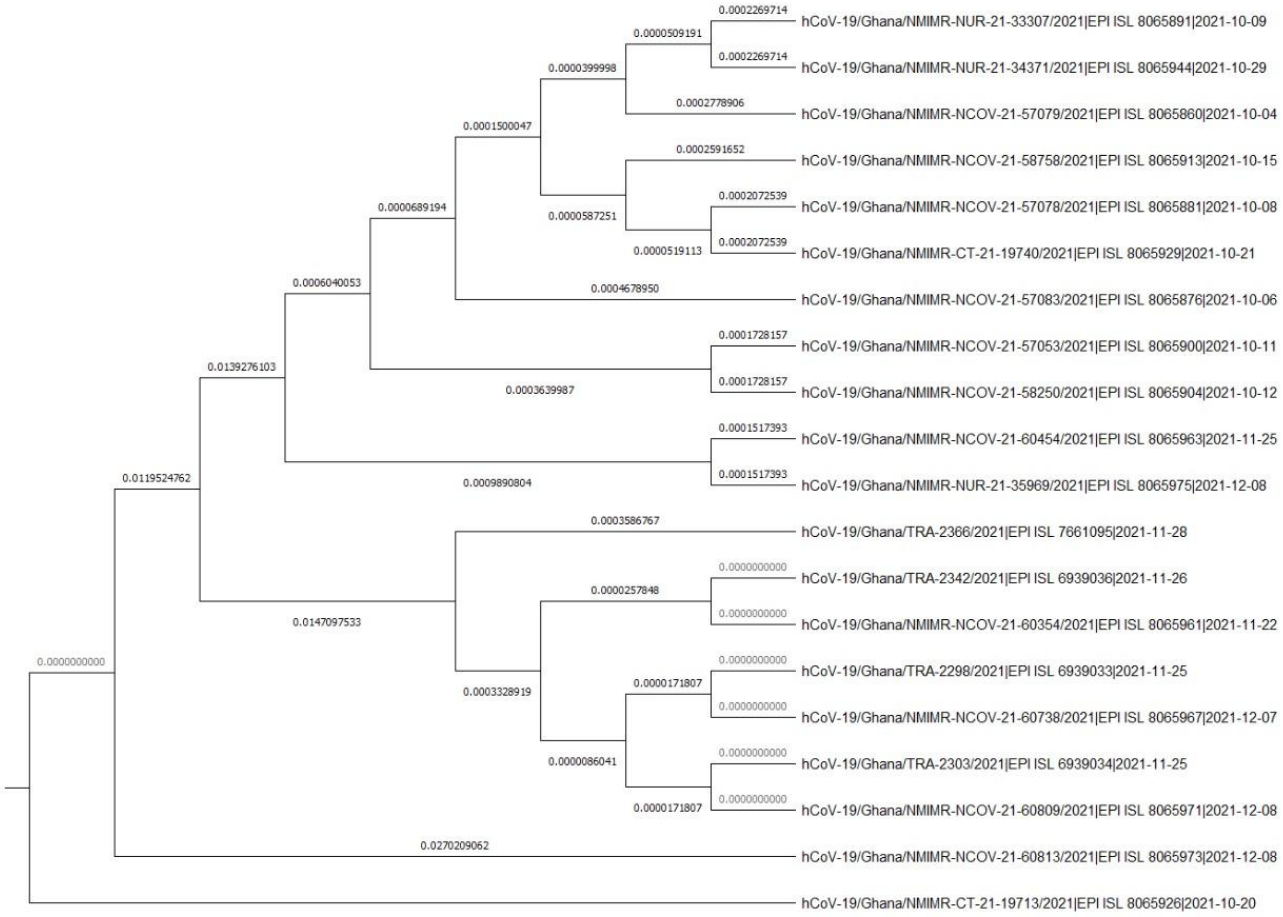
	A content%	C Content%	G content%	T content%	CG content%
Consensus Sequence	29.84	18.29	19.58	32.29	37.87
Omicron Sequence (Average)	29.148	17.862	19.087	31.53	36.951
Difference	0.692	0.428	0.493	0.76	0.919

The previous table exhibits a comparison between the amount of chemical constituents as a percentage of the sequence length for each the consensus sequence as well as the average of the same parameter computed for the multiple alignment of the case sequence, the table also shows the difference between them.



The data analysis above shows a difference of approximately 1% between the CG Content values in favor of Reference sequences. CG content is known generally to encode for higher DNA stability. The reason behind this is the fact that there is a triple hydrogen bond between G and C, which makes it harder to denature than the double hydrogen bond between A and T. Nevertheless, the 1% difference is insignificant to conclude any biological difference.

Phylogenetic tree

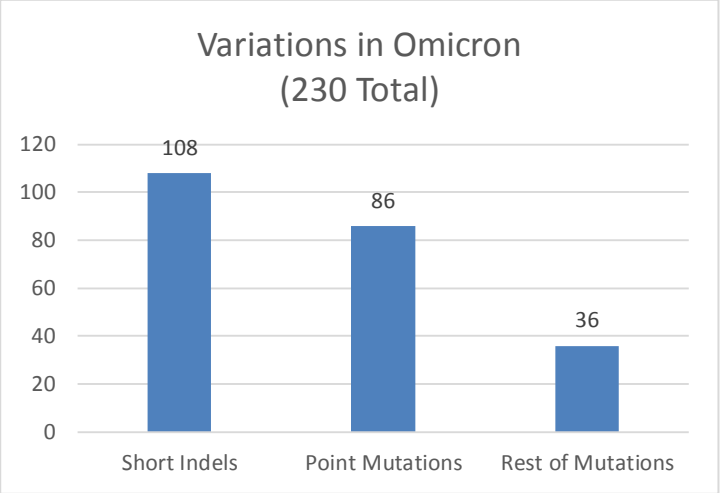


After observing the phylogenetic tree constructed by MEGA between the 20 sequences, we can tell that the distances between the sequences are very small. Thus, it can be deduced that the omicron variant is very similar to delta variant. However, one important limiting factor

needed to be taken into consideration is that the sample collected and used for the data analysis process is very small in size since time is needed to obtain more data from different countries.

Dissimilar Regions Extraction

After applying the technique mentioned above and analyzing the exported excel sheet data, we have found that out of 230 variations: 108 of them were short indels of varying length, 86 were point mutations and 36 were different regions of short length in both sequences.



The total length of all the variations was calculated through excel, it was found to be 2091 bases, which is relatively small in comparison with the length of the aligned sequences (above 30K bases).

Total length of variations:	2091

Conclusion

After this analysis that followed multiple steps, from constructing a phylogenetic tree to analyzing the content of the cases and the reference, ending with extracting the dissimilar regions, we can safely say that _quantitively_ the Omicron variant is not so different from the Delta variant.

Analyzing where the differences lie, can show us how the effect of each of them on the patient differs. On the other hand, seeing how similar they are, sheds the light on the possibility that the same procedures that are applied on finding the vaccine for Delta may also be efficient on Omicron. In the end, recognizing where the two variants differ could indeed be the key to finding the correct procedure to dealing with and treating the Omicron variant patients.

Members Contribution

All team members have contributed equally and efficiently into producing the output from the project with all its requirements, from coding to reporting and presentation.