Comparative Study Related to SARS-Cov-2 and Its Variant Omicron

Nouran Khaled Youssef
Systems and Biomedical Engineering dept.
Cairo University
Cairo, Egypt
nkhaledsoliman@gmail.com

Rania Atef Omar
Systems and Biomedical Engineering dept.
Cairo University
Cairo, Egypt
rannia.attef@gmail.com

Salma Haytham Bahy
Systems and Biomedical Engineering dept.
Cairo University
Cairo, Egypt
salma.h.bahy@gmail.com

***Abstract*—This comparative study aims to put some light on SARS-Cov-2 and its variations. We study the differences between the virus' gene sequence in 2020 and the gene sequence of its variant Omicron found in 2022. Also, we try to analyze some of the mutations found. This study is conducted upon virus samples from England in both 2020 and 2022.**

Keywords—COVID-19, SARS-CoV-2, Omicron variant, evolution, mutations

I. Introduction

Since the reporting of the first cases of coronavirus in China and the publication of the first sequence of SARS-CoV-2 in December 2019, the virus has undergone numerous mutations. In Europe, the spring outbreak (March–April) was followed by a drop in the number of cases and deaths. Evolutionarily important mutations and deletions have emerged in the SARS-CoV-2 genes encoding proteins that interact with the host immune system. In addition, one of the major mutations (in viral polymerase) is logically associated with a higher frequency of mutations throughout the genome. The rate of mutations in proteins involved in the relationship to the immune system continues to increase after the first outbreak. We collected our data from GISAID which included 10 sequences of SARS-Cov-2 and its omicron variant. Our Comparative study is implemented in 3 ways. First way is Phylogenetic tree to build branching diagrams which are made to represent the evolutionary history or relationship between different species, organisms, or characteristics of an organism. Second way is the average percentage of the chemical constituents (C, G, T and A) and the CG content. Third way represents the dissimilar regions which include the point mutation and short indels techniques. Our results are represented in a phylogenetic tree and histogram in order to make obvious differences between mutations along the whole sequence of variants strongly.

II. Methods

We used mainly 10 sequences of SARS-Cov-2 and its omicron which were obtained from England. In order to implement the three comparative ways, we used a mixture of programming and software techniques. We aligned the sequences and constructed the phylogenetic tree by using **MEGA (Molecular Evolutionary Genetics Analysis)** software. MEGA is a software for conducting statistical analysis of molecular evolution and for constructing phylogenetic trees. It includes many sophisticated methods and tools for phylogenomics and phylomedicine and is licensed as proprietary freeware. Using **Biopython** which is an open-source collection of non-commercial python tools for computational biology and bioinformatics, created by an international association of developers, **Pandas** libraries which implement the outputs data in a data frame and **Matplotlib** in order to show the differences in histograms. We implemented a function called "chemical content" to count the chemical constituents then, divide by length of the sequence and calculate the average percent of each chemical constituent. Then we calculated the difference in average chemical content between the reference and the case sequences. The Consensus sequences are constructed through a function also which takes the alignment sequence and calculates the dominant nucleotide at each location. We got the consensus sequence of both SARS-Cov-2 and its omicron variant then, aligned them and we applied three functions to extract dissimilar regions. "get_dissimilar_locations" function which returns dissimilarities between the two consensus sequence locations, "get_dissimilar_regions" which returns a list of these regions and finally, a function to count the point mutations and short indels.

III. Results and Discussion

In this section, we are going to discuss the results of comparing the case sequences (sequences of omicron variant) to the reference sequences (sequences of SARS-Cov-2 from 2020). The phylogenetic tree represents the distances between the 20 sequences. The difference between the chemical constituents' content represents the changes occurring in that content. And finally, we try to analyze the different regions between the two virus' sequences, this helps us infer these mutations' phenotypes.

A. Phylogenetic Tree

Fig. 1. shows the phylogenetic tree generated from MEGA software. The tree shows that 2 pairs of sequences in the 10 sequences of the reference covid-19 genetic material are close to each other, and these 4 sequences with another pair and the rest of the 10 sequences have close distances between them. While in the omicron variant, 9 out of the 10 sequences had a very close distance between them and with the last one they are close to the 10 reference sequences.

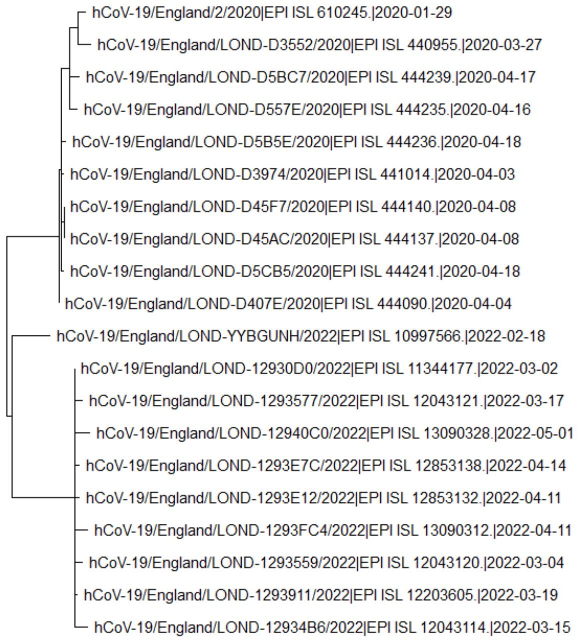
hCoV-19/England/2/2020|EPI ISL 610245.|2020-01-29
hCoV-19/England/LOND-D3552/2020|EPI ISL 440955.|2020-03-27
hCoV-19/England/LOND-D5BC7/2020|EPI ISL 444239.|2020-04-17
hCoV-19/England/LOND-D557E/2020|EPI ISL 444235.|2020-04-16
hCoV-19/England/LOND-D5B5E/2020|EPI ISL 444236.|2020-04-18
hCoV-19/England/LOND-D3974/2020|EPI ISL 441014.|2020-04-03
hCoV-19/England/LOND-D45F7/2020|EPI ISL 444140.|2020-04-08
hCoV-19/England/LOND-D45AC/2020|EPI ISL 444137.|2020-04-08
hCoV-19/England/LOND-D5CB5/2020|EPI ISL 444241.|2020-04-18
hCoV-19/England/LOND-D407E/2020|EPI ISL 444090.|2020-04-04
hCoV-19/England/LOND-YYBGUNH/2022|EPI ISL 10997566.|2022-02-18
hCoV-19/England/LOND-12930D0/2022|EPI ISL 11344177.|2022-03-02
hCoV-19/England/LOND-1293577/2022|EPI ISL 12043121.|2022-03-17
hCoV-19/England/LOND-12940C0/2022|EPI ISL 13090328.|2022-05-01
hCoV-19/England/LOND-1293E7C/2022|EPI ISL 12853138.|2022-04-14
hCoV-19/England/LOND-1293E12/2022|EPI ISL 12853132.|2022-04-11
hCoV-19/England/LOND-1293FC4/2022|EPI ISL 13090312.|2022-04-11
hCoV-19/England/LOND-1293559/2022|EPI ISL 12043120.|2022-03-04
hCoV-19/England/LOND-1293911/2022|EPI ISL 12203605.|2022-03-19
hCoV-19/England/LOND-12934B6/2022|EPI ISL 12043114.|2022-03-15

Fig. 1. The phylogenetic tree between the reference sequences of SARS-Cov-2 in 2020 and it's omicron variant in 2022 in England.

B. Difference in Chemical Constituents

We calculated the average percentage of the chemical constituents (C, G, T, and A) and the CG content in each of the reference sequences and the case sequences. Then we calculated the difference in the average percentage of these constituents in the sequences shown in Fig.2. This shows a change in CG content of around 0.4% and a change in A content of 0.21%.

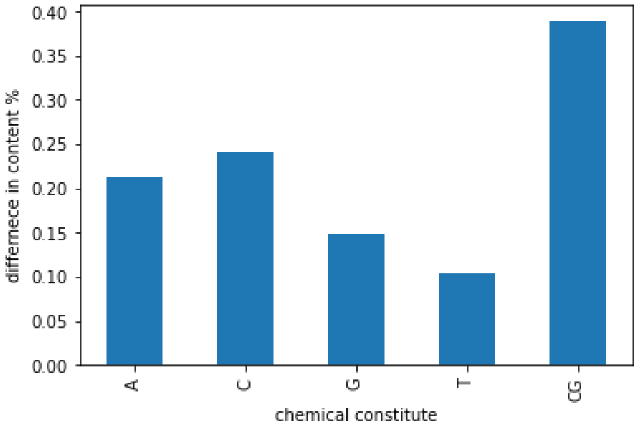

Fig. 2. Difference in average chemical constituents percentage between the reference and the case sequences.

C. Dissimilar Regions

To further compare the two variants, we construct the consensus sequences of them both. Then we apply a global pairwise alignment technique to align the two consensus sequences resulting in an alignment sequence of length 29796 bases. We then extract the dissimilar regions between the two aligned sequences. The dissimilar regions are found to be 1.76% of the sequence, and they are in 73 regions. These regions can be single letter variants, short indels (insertions/deletions) or structural variants as shown in the file attached.

Single letter variants are the most common. In the Omicron variant, there are a total of 60 single letter variants and these variants are shown in Fig. 3. Short indels are short insertions or deletions from the reference sequence. We found that there are 8 short indels of 0.37% of the total sequence shown in Fig. 4. Some other variants between the two sequences were found at 5 regions constituting 1.19% of the total sequence.

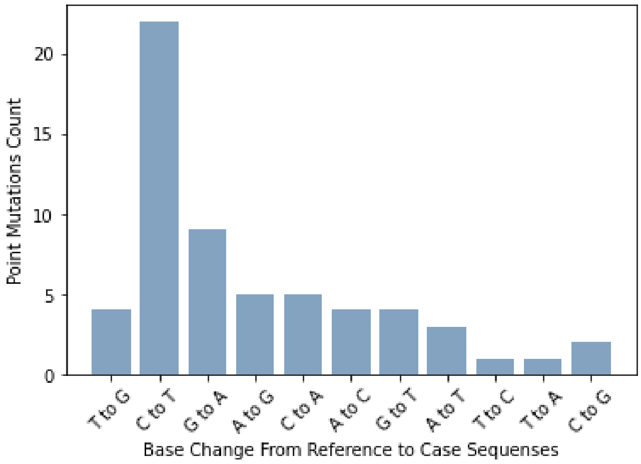

Fig. 3. Single letter variants' count between the two consensus sequences.

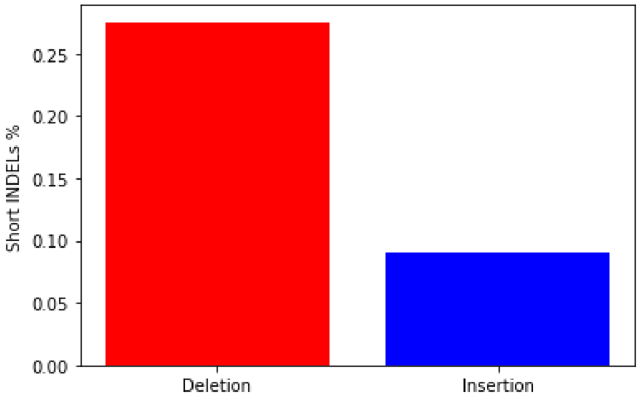

Fig. 4. Percentages of short deletions and insertions in reference sequence.

IV. Conclusion

Scientists are probably exhausted from having to explain that new studies on the coronavirus are far from a sure thing, but, thankfully, they keep setting the record straight anyway. When SARS-Cov-2 first outbreak in 2019, it was deadly. However, various mutations in the genes of the virus have weakened it a lot. Inferring phenotypes of the mutations can be very useful, and Comparing the gene sequences of the Omicron variant with the initial SARS-Cov-2 can help us understand what parts of the gene were actually deadly. Conduct the differences of whole genome sequence to determine the gap and the change alternation of nucleotides and amino acids sequences. We evaluate 10 complete genome sequences of different coronaviruses using MEGA software and Python Programming.