

**Faculty of Engineering   
Cairo University**



**Cairo University**

**Project Final**

**Submitted to:**

**Dr: Ibrahim Youssef**

**Prepared by**

|  |  |  |
| --- | --- | --- |
| Name | Section | BN |
| Abbas Mohammed Abbas | 1 | 46 |
| Ahmed Hatem Farouk | 1 | 5 |
| Abd Allah Mohammed Shazly | 1 | 52 |
| AbdUlkhalek Mohammed Alfakih | 1 | 47 |

**Introduction**

Ever since the corona outbreak, 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*. 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 that took place for the original mutant.

The 10 sequences for *SARS-Cov-2* for *Italy*:

|  |  |
| --- | --- |
| Original 1 | hCoV-19/Italy/PIE-SLL-MS45/2020|EPI\_ISL\_569866|2020-02-25 |
| Original 2 | hCoV-19/Italy/PIE-SLL-MS46/2020|EPI\_ISL\_569867|2020-02-22 |
| Original 3 | hCoV-19/Italy/TUS-C44/2020|EPI\_ISL\_738147|2020-02-25 |
| Original 4 | hCoV-19/Italy/TUS-C198/2020|EPI\_ISL\_738194|2020-02-24 |
| Original 5 | hCoV-19/Italy/LOM-UniMI05/2020|EPI\_ISL\_779704|2020-02-24 |
| Original 6 | hCoV-19/Italy/LOM-UniMI07/2020|EPI\_ISL\_779709|2020-02-24 |
| Original 7 | hCoV-19/Italy/LOM-UniMI08/2020|EPI\_ISL\_779712|2020-02-24 |
| Original 8 | hCoV-19/Italy/LOM-UniMI09/2020|EPI\_ISL\_779713|2020-02-24 |
| Original 9 | hCoV-19/Italy/LOM-UniSR2/2020|EPI\_ISL\_1499504|2020-02-28 |
| Original 10 | hCoV-19/Italy/EMR-UA-02\_00108/2020|EPI\_ISL\_5687737|2020-02-29 |

The 10 sequences for SARS-Cov-2 Omicron variant for *Italy*:

|  |  |
| --- | --- |
| variant 1 | hCoV-19/Italy/UMB-IZSGC-318786.1.36/2021|EPI\_ISL\_7952669|2021-12-14 |
| variant 2 | hCoV-19/Italy/UMB-IZSGC-318786.1.33/2021|EPI\_ISL\_7952666|2021-12-14 |
| variant 3 | hCoV-19/Italy/UMB-IZSGC-318786.1.37/2021|EPI\_ISL\_7952670|2021-12-14 |
| variant 4 | hCoV-19/Italy/UMB-IZSGC-318786.1.32/2021|EPI\_ISL\_7952665|2021-12-14 |
| variant 5 | hCoV-19/Italy/UMB-IZSGC-318786.1.35/2021|EPI\_ISL\_7952668|2021-12-14 |
| variant 6 | hCoV-19/Italy/UMB-IZSGC-318786.1.34/2021|EPI\_ISL\_7952667|2021-12-14 |
| variant 7 | hCoV-19/Italy/UMB-IZSGC-318786.1.68/2021|EPI\_ISL\_7952701|2021-12-14 |
| variant 8 | hCoV-19/Italy/UMB-IZSGC-318786.1.38/2021|EPI\_ISL\_7952671|2021-12-14 |
| variant 9 | hCoV-19/Italy/UMB-IZSGC-318786.1.69/2021|EPI\_ISL\_7952702|2021-12-14 |
| variant 10 | hCoV-19/Italy/CAL-AOCatanzaro-12171195\_CZ/2021|EPI\_ISL\_7952114|2021-12-17 |

**Information About Data Used**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Gender | Patient Age | Patient status |
| Original 1 | Female | 83 | Deceased |
| Original 2 | Male | 41 | Live |
| Original 3 | Male | 49 | Deceased |
| Original 4 | Male | 63 | Released |
| Original 5 | Male | 43 | Hospitalized |
| Original 6 | Female | 72 | Hospitalized |
| Original 7 | Male | 60 | Hospitalized |
| Original 8 | Female | 61 | Hospitalized |
| Original 9 | Male | 58 | Not Hospitalized |
| Original 10 | Male | 66 | Deceased |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Gender | Patient Age | Patient status |
| variant 1 | Female | 22 | Live |
| variant 2 | Male | 40 | Live |
| variant 3 | Male | 24 | Live |
| variant 4 | Male | 34 | Live |
| variant 5 | Male | 56 | Live |
| variant 6 | Female | 23 | Live |
| variant 7 | Male | 25 | Live |
| variant 8 | Female | 57 | Live |
| variant 9 | Female | 49 | Live |
| variant 10 | Male | 37 | Live |

**methodology and findings**

We Construct a consensus sequence from the reference sequences.by using UGENE software we 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.

The 10 SARS-Cov-2 sequences:



We have a lot of Consensus types:

* *Strict* — specifies that a set of species must appear in all input trees to be included in the strict consensus tree.
* *Majority Rule (extended)* — specifies that any set of species that appears in more than 50% of the trees is included. The program then considers the other sets of species in order of the frequency with which they have appeared, adding to the consensus tree any which are compatible with it until the tree is fully resolved. This is the default setting.
* *M1* — includes in the consensus tree any sets of species that occur among the input trees more than a specified fraction of the time (see the *Fraction* parameter below). The *Strict* consensus and the *Majority Rule* consensus are extreme cases of the Ml consensus, being for fractions of 1 and 0.5 respectively.
* *Majority Rule* — specifies that a set of species is included in the consensus tree if it is present in more than half of the input trees.

In our example we choose the Strict type and we specify the threshold by 39% but before that we will do the multiple sequence alignment.

Now the file contains the consensus sequence.



The 10 sequences for the SARS-Cov-2 Omicron variant:



WE choose to Apply the align with MAFFT:

In bioinformatics, MAFFT is a program used to create multiple sequence alignments of amino acid or nucleotide sequences. MAFFT use an algorithm based on progressive alignment, in which the sequences were clustered with the help of the Fast Fourier Transform.

A progressive alignment method is described that utilizes the Needleman and Wunsch pairwise alignment algorithm iteratively to achieve the multiple alignment.

It has parameters in which you should choose to begin the alignment

1. Gap opening penalty we used 3
2. Offset (works like gap extension penalty) we used 1
3. Maximum number of iterative refinements we used the default which is 2

file after apply MAFFT:

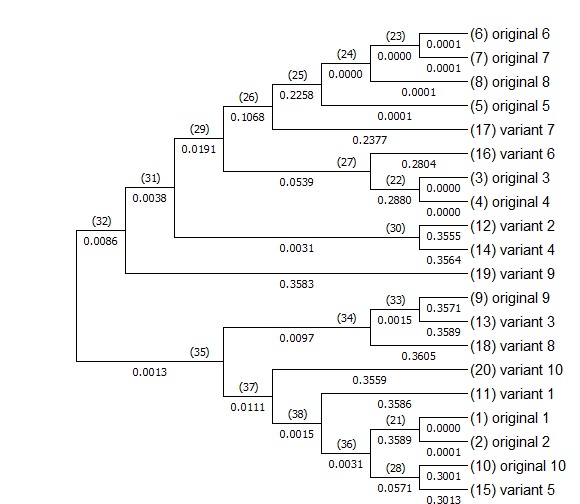


We Construct a phylogenetic tree between all the above 20 sequences:

We constructed the phylogenetic tree using the statistical method neighbor-joining.

The model used is the p-distance.

This distance is the proportion (p) of nucleotide sites at which two sequences being compared are different. It is obtained by dividing the number of nucleotide differences by the total number of nucleotides compared



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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AncId | Desc1 | Desc2 | Branch Length 1 | Branch Length 2 |
| 21 | original 1 | original 2 | 4.18E-05 | 5.88E-05 |
| 22 | original 3 | original 4 | 3.27E-05 | 3.47E-05 |
| 23 | original 6 | original 7 | 7.03E-05 | 6.40E-05 |
| 24 | 23 | original 8 | 7.42E-06 | 9.33E-05 |
| 25 | 24 | original 5 | 1.58E-05 | 0.000119 |
| 26 | 25 | variant 7 | 0.22584 | 0.237745 |
| 27 | variant 6 | 22 | 0.280379 | 0.287968 |
| 28 | original 10 | variant 5 | 0.300108 | 0.301322 |
| 29 | 26 | 27 | 0.106773 | 0.053892 |
| 30 | variant 2 | variant 4 | 0.355503 | 0.356402 |
| 31 | 29 | 30 | 0.019106 | 0.00313 |
| 32 | 31 | variant 9 | 0.003826 | 0.358349 |
| 33 | original 9 | variant 3 | 0.357081 | 0.358899 |
| 34 | 33 | variant 8 | 0.001494 | 0.360514 |
| 35 | 34 | 37 | 0.00971 | 0.011143 |
| 36 | 21 | 28 | 0.35889 | 0.057082 |
| 37 | variant 10 | 38 | 0.355892 | 0.001529 |
| 38 | variant 1 | 36 | 0.358615 | 0.003127 |
| 39 | 32 | 35 | 0.008608 | 0.001311 |

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



now we Extract the dissimilar regions/columns between the alignment of the case sequences and the consensus sequence:



and we Extract the similar regions/columns between the alignment of the case sequences and the consensus sequence:



the numbers of similar for each Nucleotide: the numbers of dissimilar for each Nucleotide:

|  |  |
| --- | --- |
| Nucleotide | Number of dissimilar |
| A | 6228 |
| C | 4504 |
| G | 4816 |
| T | 6608 |

|  |  |
| --- | --- |
| Nucleotide | Number of similar |
| A | 2693 |
| C | 987 |
| G | 1044 |
| T | 3004 |

Most converted to nucleotides in dissimilar case:

|  |  |
| --- | --- |
| Nucleotide | Number of Change |
| A | 6170 |
| C | 4449 |
| G | 4772 |
| T | 6525 |

Table showing the amount of similarity between each variant and case sequence:

**Similar table**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variants  Case sequence | Variant 1 | Variant 2 | Variant 3 | Variant 4 | Variant 5 | Variant 6 | Variant 7 | Variant 8 | Variant 9 | Variant 10 |
| # Of similar | 7771 | 7749 | 7699 | 7769 | 7759 | 7735 | 7761 | 7742 | 7762 | 7762 |
| Percentage% | 26% | 26% | 26% | 26% | 26% | 26% | 26% | 26% | 26% | 26% |

Table showing the amount of dissimilarity between each variant and case sequence:

**Dissimilar table**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variants  Case sequence | Variant 1 | Variant  2 | Variant 3 | Variant 4 | Variant 5 | Variant 6 | Variant 7 | Variant 8 | Variant 9 | Variant 10 |
| #Of dissimilar | 22121 | 22143 | 22193 | 22123 | 22133 | 22157 | 22131 | 22150 | 22130 | 22130 |
| Percentage% | 74% | 74% | 74% | 74% | 74% | 74% | 74% | 74% | 74% | 74% |



***Conclusion:***

* In conclusion, we observed that CG contented, which represents the stability of a sequences, was pretty low in the corona sequences which led to the Inevitable mutations resulting in the omicron variant.
* As the excel sheet shows, the differences that occur between SARS-Cov-2 Omicron variant and the Consensus sequence are large, but we note that the percentage of amino acid change was not that large, and this shows that most of the differences that occur are nothing but silent mutation

Ex:

GAA GLU

GAG GLU

Both ***GAG*** and ***GAG*** refers to the same Amino acid (***GLU***)

* With the same idea, we note that there are slight changes in some amino acid present in SARS-Cov-2 Omicron variant compared to the Consensus sequence, an example of this (***Tyr***)

We note that the ***Tyr*** has decreased in the variant 1 compared to the Consensus sequence, it has decreased by 0.65 % and this Missense *mutation*.

* It can be seen that the omicron variant is quite aggressive, so one of the best ways if not the best is to take the corona vaccine seeing that omicron originated from corona. It might not be the optimal solution, but it is sure to help reduce the adverse effects that omicron bears.