**Name**: Ali Eastman Oku

**Z-ID**: Z1893417

**Final Project**

**Programming Language Used**: Python (3.9.6). I also used the version on the server to view the results (version 2.7.16).

**Time Spent:** I spent about 25 hours completing this assignment. I used about 12 hours to write the code and spent the remaining time, creating the PowerPoint presentation, writing documentation, and writing the final report for project.

**Results:**

**Sample 119 \* 119 Distance Matrix (using entire genome)**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **…** | **118** |
| **0** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | **…** | 0.408 |
| **1** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | **…** | 0.408 |
| **2** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | **…** | 0.408 |
| **3** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | **…** | 0.408 |
| **4** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | **…** | 0.408 |
| **6** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | **…** | 0.408 |
| **7** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | **…** | 0.408 |
| **…** | … | … | … | … | … | … | … | **…** | **…** |
| **119** | 0.408 | 0.408 | 0.408 | 0.408 | 0.408 | 0.408 | 0.408 | … | 0.000 |

**For phylogenetic tree information, see supplementary documents**

**Supplementary Documents**

1. for phylogenetic tree using entire genome, see “global\_align\_distance.pdf”
2. for phylogenetic tree using orf1a gene, see “orf1a\_align.pdf”
3. for phylogenetic tree using orf1b, see “orf1b\_align\_align.pdf”
4. for phylogenetic tree using S gene, see “s\_align.pdf”
5. for phylogenetic tree using M gene, see “m\_align.pdf”
6. for phylogenetic tree using E gene, see “e\_align.pdf”
7. for phylogenetic tree using N gene, see “n\_align.pdf”

**Summary**

Using, “WUHCOR1.NC\_045512V2” as reference genome

|  |  |  |
| --- | --- | --- |
| **Gene** | **Closest Genome** | **Most Distant Genome** |
| Whole genome | WIV04 | BELUGA WHALE COV SW1 |
| Orf1a | WIVO4 | MAGPIE-ROBIN COV HKU18 |
| Orf1b | WIV04 | WIGEON COV HKU20 |
| M | WIV04 | WIGEON COV HKU20 |
| S | WIV04 | WIGEON COV HKU20 |
| E | WIV04 | WIGEON COV HKU20 |
| N | WIVO4 | WIGEON COV HKU20 |

**Discussion:**

The input file contained the genome of SARS\_CoV\_2 with other related corona viruses. After analysis, it was observed that the input file also contained different SARS-CoV-2 viruses isolated from different patients. However, “WUHCOR1.NC\_045512V2” was used as the reference genome during analysis and was compared to all other genomes in the input file. Using the whole genomes to generate a distance matrix showed that about the first 50 genomes in the file were very similar. This was due to the low substitution rates observed across these 50 genomes. This was expected after discovering that the first few genomes in the input file were mostly SARS\_CoV\_2 viruses isolated from different patients. Using this distance matrix to generate a phylogenetic tree showed that the closest genome to our reference SARS\_CoV\_2 genome (“WUHCOR1.NC\_045512V2”) was “WIV04”. The phylogenetic tree also captured the close level of similarity between the first 50 genomes since the branch length between these genomes were quite minimal. The phylogenetic tree also showed that the closest genome to SARS\_CoV\_2 was “BAT COV TG13” which corresponded with previous reported research [1]. The close similarity with SARS\_CoV\_2 might indicate that the virus was previously in bats before crossing over to humans. The most distantly related species according to the phylogenetic tree belonged to the BELUGA WHALE COV.

Using the M, S, E, N, Orf1a, Orf1b genes produced very similar phylogenetic trees with the same closest and farthest related genomes. However, using only the specified genes for analysis produced sequences that had no matching bases with the sequence of comparison. This is indicated the there was a high level of mutation in those genes. This had a high occurrence in the M, S, E and N genes with the N genes having the highest amount of dissimilarity.

References

1. Zhou, Peng, et al. “A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin.” *Nature*, vol. 579, no. 7798, 2020, pp. 270–273., https://doi.org/10.1038/s41586-020-2012-7.