**Programming Project 3 Report**

1. **System configuration:**

CPU used - Intel Core i9-9900k

Clock rate - 4.7 GHz

RAM - 32.0 GB @ 3200Mhz

Cache size - L1:512KB, L2:2MB, L3:16MB

1. **Quality:**

**Input String Fingerprint Lengths**

|  |  |  |
| --- | --- | --- |
| **String ID** | **Strain Name** | **Fingerprint Length** |
| 1 | COVID-19 China | 125 |
| 2 | COVID-19 USA | 8 |
| 3 | COVID-19 Australia | 8 |
| 4 | COVID-19 India | 8 |
| 5 | COVID-19 Brazil | 7 |
| 6 | SARS\_2013 | 7 |
| 7 | SARS\_2017 | 7 |
| 8 | MERS\_2012\_Saudi | 7 |
| 9 | MERS\_2014\_Saudi | 7 |
| 10 | MERS\_2014\_USA | 6 |

**Similarity Matrix D**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 |  |  |  |  |  |  |  |  |  |  |
| 2 | 28965 |  |  |  |  |  |  |  |  |  |
| 3 | 28976 | 29865 |  |  |  |  |  |  |  |  |
| 4 | 28947 | 29845 | 29835 |  |  |  |  |  |  |  |
| 5 | 28958 | 29868 | 29861 | 29844 |  |  |  |  |  |  |
| 6 | 20746 | 23687 | 23679 | 23688 | 23690 |  |  |  |  |  |
| 7 | 20789 | 23761 | 23753 | 23744 | 23764 | 29629 |  |  |  |  |
| 8 | 2506 | 2506 | 2506 | 2504 | 2505 | 529 | 529 |  |  |  |
| 9 | 2506 | 2506 | 2506 | 2504 | 2505 | 529 | 529 | 30005 |  |  |
| 10 | 2507 | 2507 | 2507 | 2505 | 2506 | 529 | 529 | 30003 | 30052 |  |

**What does the fingerprint lengths tell you about each strain and its uniqueness?**

The fingerprint length shows us to what extent a strain is unique. Covid-19 China was expected to have the longest fingerprint (Least Unique), due to all the other strains coming directly from it. Since other COVID’s mutated from it, it would have the most in common with other strings; therefore, its fingerprint would have to be the longest since we are looking for a shortest common string where a single new char would identify it from the rest.

It is weird to see that the MERS have such short Fingerprints, we were expecting larger fingerprints due to them being the same strain technically.

**What does the similarity matrix D tell you about the relationship between these strains? Do you see any logical groupings visible through this matrix?**

The similarity matrix D is very helpful to us in showing the relationship between strains, the higher the score between two strains, the more similar they are. This can actually be seen quite well on our Similarity matrix above. As expected, the COVID strains have a very high similarity between each other, and they can actually be grouped, same goes for SARS and MERS. All of these have scores of 28 thousand and higher between each other. It is interesting to see that even though it’s not 28 thousand when COVID and SARS are compared, their scores are higher than 20 thousand. SARs, on the other hand, though don’t have much in common with the other two. MERS\_2012\_Saudi is in a category of its own and barely has any connection with any of the other strains.

1. **Performance:**

**Task 1 Fingerprinting performance**

|  |  |  |
| --- | --- | --- |
| **Task Name** | **Time Taken for Tasks in Seconds** | **Total Time for Tasks in Seconds** |
| Construction of Suffix Tree | 0.0 | 0.146818 |
| Fingerprint Identification | 0.072108 |

**Task 2 Similarity matrix performance**

|  |  |  |
| --- | --- | --- |
| **Task Name** | **Time Taken for Tasks in Seconds** | **Total Time for Matrix D Computation in Seconds** |
| Building Suffix Trees | 0.492241 | 125.258331 |
| Time spent performing alignments | 107.295857 |

**Longest Common Substring Length for Each Pair of Strains**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 |  |  |  |  |  |  |  |  |  |  |
| 2 | 23769 |  |  |  |  |  |  |  |  |  |
| 3 | 18150 | 13980 |  |  |  |  |  |  |  |  |
| 4 | 7961 | 7961 | 7961 |  |  |  |  |  |  |  |
| 5 | 10168 | 8896 | 11082 | 4620 |  |  |  |  |  |  |
| 6 | 104 | 104 | 104 | 104 | 104 |  |  |  |  |  |
| 7 | 104 | 104 | 104 | 104 | 104 | 7878 |  |  |  |  |
| 8 | 23 | 23 | 23 | 23 | 23 | 20 | 20 |  |  |  |
| 9 | 23 | 23 | 23 | 23 | 23 | 20 | 20 | 2890 |  |  |
| 10 | 23 | 23 | 23 | 23 | 23 | 20 | 20 | 3182 | 3094 |  |

1. **Other comments:**    If you have any other observations or comments, you can add them here in this section.

It’s interesting to see how the LCS sizes compare. Out of all the COVID LCS’s there is a clear correlation between the size of the LCS and when the two strains were first found. Since China and USA had their cases very early in comparison to others, we can see how alike they are, with their LCS being over 2/3 of their lengths. When we compare COVID-19 China to other COVIDS, the LCS size goes down with time.