System Configuration

CPU Used: Intel i5-6300HQ

Clock Rate: 2.30GHz

RAM: 8GB

Cache Sizes –

L1: 256KB

L2: 1.0MB

L3: 6.0MB

Quality

a. Fingerprint Lengths

String ID	Fingerprint Length
1	125
2	8
3	8
4	8
5	7
6	7
7	7
8	7
9	7
10	6

b. Similarity Matrix D

String IDs	1	2	3	4	5	6	7	8	9	10
1	29903	29879	29890	29849	29872	23708	23782	2515	2515	2519
2	29879	29882	29866	29846	29869	23705	23779	2515	2515	2516
3	29890	29866	29893	29836	29861	23698	23772	2515	2515	2516
4	29849	29846	29836	29854	29845	23705	23772	2513	2513	2514
5	29872	29869	29861	29845	29876	23708	23782	2514	2514	2515
6	23708	23705	23698	23705	23708	29644	29629	529	529	529
7	23782	23779	23772	23772	23782	29629	29727	529	529	529
8	2515	2515	2515	2513	2514	529	529	30055	30005	30003
9	2515	2515	2515	2513	2514	529	529	30005	30123	30053
10	2519	2516	2516	2514	2515	529	529	30003	30053	30123

c.

The fingerprint lengths let us tell how unique each string is. The shortest unique substring for each DNA sequence tell us how unique a sequence is by essentially saying that any substring shorter than the length given is no longer unique. Therefore the longer the shortest unique substring is, the less unique that substring is to the other ones given because a larger portion of it is contained in the other DNA sequences. From this we can determine that the Wuhan variant of COVID is the most similar to the other variants and to SARS and MERS since a fairly large portion of it is found in the other sequences. If we try to find a unique region of less than 125 base pairs we would not be able to. We can also tell from this table that the other COVID variants along with SARS and MERS are rather unique since we find rather short unique substrings.

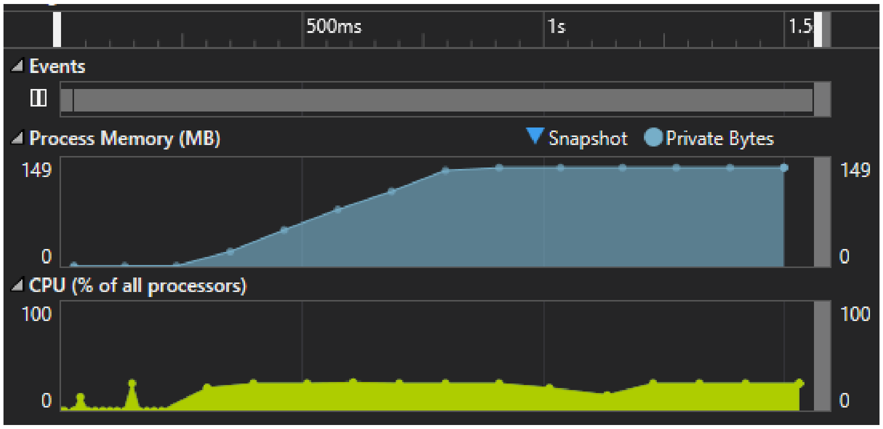
The similarity matrix shows us which DNA sequences are most similar to each other. Since they are all approximately the same length we do not need a denominator to parse the results. The similarity matrix shows that the viruses fall into 3 distinct categories as shown in the highlighter above. The yellow region are the COVID strains, green is SARS, and blue is MERS. This similarity matrix shows that the viruses that we have classified in the same category are indeed genetically similar. In columns 6 and 7 until the green portion is reached we can see that COVID and SARS is fairly similar which suggests that COVID may have mutated from some SARS variant.

Performance

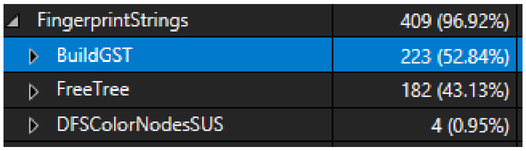
Task 1 --

Time to Build Suffix Tree: 225,534 microseconds

Time to Identify Fingerprints: 120,218 microseconds



CPU Profile and Memory Usage during fingerprinting



FingerprintStrings	409 (96.92%)
BuildGST	223 (52.84%)
FreeTree	182 (43.13%)
DFSColorNodesSUS	4 (0.95%)

While fingerprinting the sequences Building the GST accounted for 52.84% of CPU usage and the node coloring that actually determines the fingerprints accounts for 0.95% of the CPU usage

Task 2 –

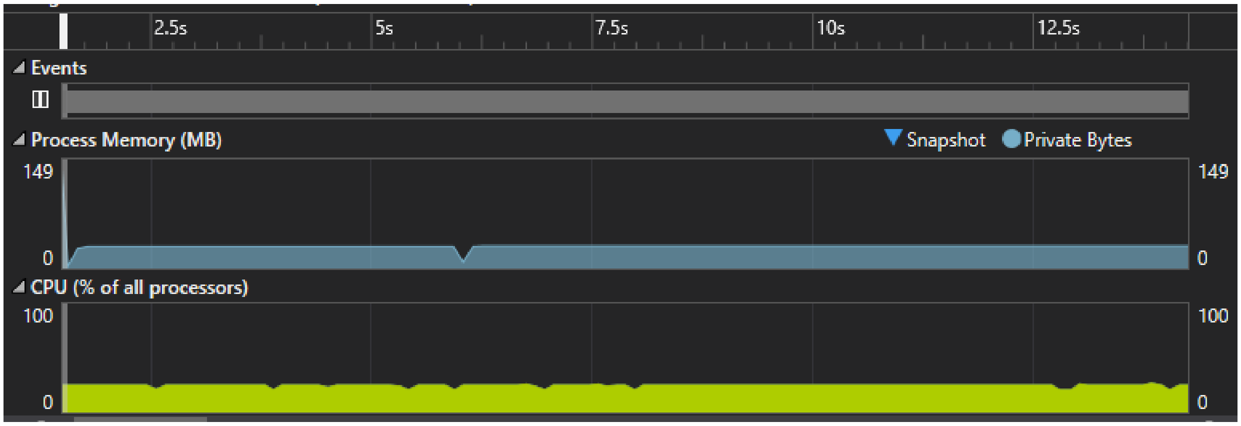
Time Building Suffix Trees: 1,811,803 microseconds

Time Performing Alignments: 1,214,856,668 microseconds

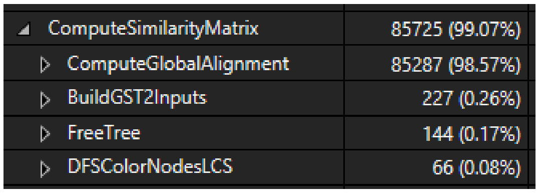
Total Time to Compute Similarity Matrix: 1,219,284,791 microseconds

LCS Lengths --

String IDs	1	2	3	4	5	6	7	8	9	10
1		23769	19064	7961	11082	104	104	23	23	23
2			13980	7961	8896	104	104	23	23	23
3				7961	11082	104	104	23	23	23
4					4620	104	104	23	23	23
5						104	104	23	23	23
6							7878	20	20	20
7								20	20	20
8									2890	3182
9										3094
10										



CPU Profile and Memory Usage during Computation of Similarity Matrix



ComputeSimilarityMatrix	85725 (99.07%)
ComputeGlobalAlignment	85287 (98.57%)
BuildGST2Inputs	227 (0.26%)
FreeTree	144 (0.17%)
DFSColorNodesLCS	66 (0.08%)

While computing the similarity matrix, building the suffix trees accounts for 0.26% of the CPU usage, coloring the nodes accounts for 0.08% of the CPU usage, and computing the global alignment values accounts for 98.57% of CPU usage