Project 3 Report

System configuration: (Abrar Akhyer Abir's Laptop)

Processor: Intel(R) Core(TM) i7-7660U CPU @ 2.50GHz (4 CPUs), ~2.5GHz

Memory: 16384MB RAM

System configuration: (Asif Zaman's Laptop)

Processor: AMD® Ryzen 5 5600u with Radeon graphics × 12

Memory: 13.5 GiB

Cache: 16MB

How to Run the project:

Run this command "***python sf_main.py***"

Quality Report:

Fingerprints Report

Input	Fingerprint	Length
Covid_Wuhan	GTACAGTGAACAATGCTAGGGAGAGCTGCCTATATGGAAGAGCCCTAATGTGTAAAATTAATTTTAGTAGTGCTATCCCCATGTGATTTTAATAGCTTCTTAGGAGAATGACAAAAAAAAAAAAA	125
Covid_USA-CA4	ACTTACCG	8
Covid_Australia	ATACAGTG	8
Covid_India	ACTAAGGA	8
Covid_Brazil	CGCGCTC	7
SARS_2017_MK062179	AACCTCG	7
SARS_2003_GU553363	CAGGAGG	7
MERS_2014_USA_KP223131	GTCCCC	6
MERS_2014_KY581694	CCCCTTG	7
MERS_2012_KF600620	CAAGGGG	7

Here, we can see, Covid_Wuhan has the longest fingerprint available. Since it was the first one strain of Covid-19, we think it has longest fingerprint than other strains.

Length of the Longest Common Substrings Report

We were able to construct the GST for all strains without reducing the size input size.

String	Covid_ Wuhan	Covid_ USA-CA4	Covid_A ustralia	Covid_ India	Covid_ Brazil	SARS_2017_ MK062179	SARS_2003_G U553363	MERS_2014_USA _KP223131	MERS_2014_ KY581694	MERS_2012_ KF600620
Covid_Wuhan	-	23769	19064	7961	11082	104	104	23	23	23
Covid_USA-CA4		-	13980	7961	8896	104	104	23	23	23
Covid_Australia			-	7961	11082	104	104	23	23	23
Covid_India				-	4620	104	104	23	23	23
Covid_Brazil					-	104	104	23	23	23
SARS_2017_MK0 62179						-	7878	20	20	20
SARS_2003_GU5 53363							--	20	20	20
MERS_2014_USA _KP223131								-	3098	3182
MERS_2014_KY5 81694									-	2890
MERS_2012_KF6 00620										-

From the table, we can see similar types of strains has longer common substrings among them. However, if the strains vary the length of longest common substring is also reduced.

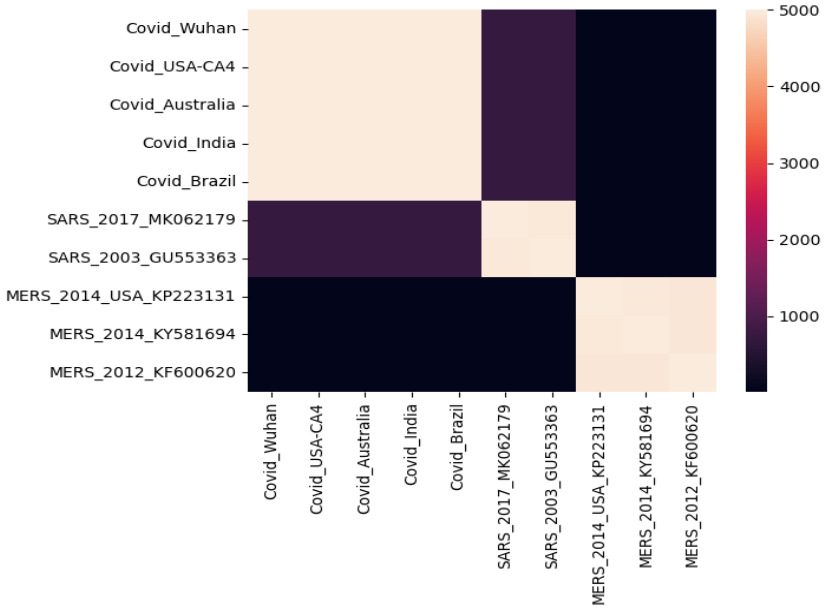
Similarity Matrix Report:

Since our laptop was not able to run the alignment for full input size, we have used 5000 characters for each input and perform the alignment.

Similarity Matrix:

String	Covid_ Wuhan	Covid_ USA-CA4	Covid_A ustralia	Covid_ India	Covid_ Brazil	SARS_2017_ MK062179	SARS_2003_G U553363	MERS_2014_USA _KP223131	MERS_2014_ KY581694	MERS_2012_ KF600620
Covid_Wuhan	5000	4997	4994	4984	4994	733	726	13	13	13
Covid_USA-CA4	4997	5000	4997	4981	4997	730	723	13	13	13
Covid_Australia	4994	4997	5000	4984	4994	733	726	13	13	13
Covid_India	4984	4981	4984	5000	4620	725	723	13	13	13
Covid_Brazil	4994	4997	4994	4984	5000	733	726	13	13	13
SARS_2017_MK0 62179	733	730	733	725	733	5000	4960	21	21	21
SARS_2003_GU5 53363	726	723	726	723	726	4960	5000	21	21	21
MERS_2014_USA _KP223131	13	13	13	13	13	21	21	5000	4961	4930
MERS_2014_KY5 81694	13	13	13	13	13	21	21	4961	5000	4924
MERS_2012_KF6 00620	13	13	13	13	13	21	21	4930	4924	5000

Similarity Matrix Heatmap



From the heatmap and similarity matrix, we can see the covid strains are really similar to each other. Same goes for SARS strain and MERS strain. We can also deduce that SARS has more similarity to Covid Strains than MERS.

Performance:

Task	Time (Seconds)
Generalized suffix tree construction	713.138
Time to identify fingerprints	1.4031
Whole Task 1	716.12

Alignment time for 5000 Characters **(in seconds)**:

String	Covid_ Wuhan	Covid_ USA-CA4	Covid_A ustralia	Covid_ India	Covid_ Brazil	SARS_2017_ MK062179	SARS_2003_G U553363	MERS_2014_USA _KP223131	MERS_2014_ KY581694	MERS_2012_ KF600620
Covid_Wuhan	-	300.42	59.97	303.02	270.01	269.43	264.21	262.04	288.51	249.32
Covid_USA-CA4		-	249.32	80.55	274.54	315.52	306.72	344.63	304.96	282.08
Covid_Australia			-	50.00	260.10	234.13	282.91	332.98	336.46	323.45
Covid_India				-	68.87	292.61	265.37	290.20	293.42	285.85
Covid_Brazil					-	266.67	279.03	288.55	290.78	362.53
SARS_2017_MK062179						-	0.0	270.80	217.27	194.99
SARS_2003_GU553363							-	185.74	196.06	247.39
MERS_2014_USA_KP223131								-	144.97	86.87
MERS_2014_KY581694									-	236.16
MERS_2012_KF600620										-