CPTS471 PA3 Report

Team

Miranda Stoddard

Ace Cassidy

System Configuration

- CPU: Intel Core i5-8350U
- Clock Rate: 1.70GHz X 8
- Ram: 16 Gb DDR4
- OS: Ubuntu 20.04

Qualitative Analysis

Fingerprinting

1. Fingerprinting Length Values

String	Length	Fingerprint
Covid_Australia	8	GTACCCCA
Covid_Brazil	7	GCGCTCG
Covid_India	8	GTCTTTTC
Covid_USA	8	TGACTTAC
Covid_Wuhan	125	GTACAGTGA...
MERS_2012_KF	X	X
MERS_2014_KY	6	TCCCCC
MERS_2014_USA	6	GTCCCC
SARS_2003_GU	2	TGCCCG
SARS_2017_MK	7	TCTGCCC

2. Fingerprinting Analysis

We can tell from its fingerprint length that Covid_Wuhan is the most similiar to other strains, which is reasonable because all other Covid strains originated from it. Other strains seem to be equally related to at least one other strain, meaning they have approximately the same fingerprint length.

Similarity Matrix

1. Similarity Matrix Match values

Sequence	CA	CB	CI	CU	CW	MK	MU	SG	SM
CA	X	29836	29797	29839	29870	211	214	11257	11302
CB	X	X	29819	29855	29864	208	211	11285	11330
CI	X	X	X	29822	29831	205	208	11277	11318
CU	X	X	X	X	29873	211	214	11276	11321
CW	X	X	X	X	X	211	214	11285	11330
MK	X	X	X	X	X	X	29914	141	141
MU	X	X	X	X	X	X	X	141	141
SG	X	X	X	X	X	X	X	X	29592
SM	X	X	X	X	X	X	X	X	X

- CA = Covid_Australia
- CB = Covid_Brazil
- CI = Covid_India
- CU = Covid_USA
- CW = Covid_Wuhan
- MK = MERS_2014_KY
- MU = MERS_2014_USA
- SG = SARS_2003_GU
- SM = SARS_2017_MK

2. Similarity Matrix Observations

Based on the matrix each virus seems most related to other strains of the same virus. Covid seems to be more related to SARS than MERS and MERS seems to be more related to Covid than SARS although it is not very related to either of them compared to how closely Covid and SARS are related.

Construction Performance

Fingerprinting Performance

1. suffix tree construction time

   642215 us

2. time to identify fingerprints

   267142 us

3. total time

   934470 us

Similarity Matrix Performance

1. suffix tree construction time

   2604257 us

2. time spent performing the alignments

   72210362 us

3. total time

   76884094 us

LCA Lengths

Sequence	CA	CB	CI	CU	CW	MK	MU	SG	SM
CA	X	11082	7961	13980	19064	23	23	104	104
CB	X	X	4620	8896	11082	23	23	104	104
CI	X	X	X	7961	7961	23	23	104	104
CU	X	X	X	X	23769	23	23	104	104
CW	X	X	X	X	X	23	23	104	104
MK	X	X	X	X	X	X	3094	20	20
MU	X	X	X	X	X	X	X	20	20
SG	X	X	X	X	X	X	X	X	7878
SM	X	X	X	X	X	X	X	X	X

- CA = Covid_Australia
- CB = Covid_Brazil
- CI = Covid_India

- CU = Covid_USA
- CW = Covid_Wuhan
- MK = MERS_2014_KY
- MU = MERS_2014_USA
- SG = SARS_2003_GU
- SM = SARS_2017_MK