Cpt S 471/571 Assignment Cover Sheet

(To be turned in along with each homework and program project submission)

Assignment # Programming Project 2

For individual assignments:

Student name (Last, First): Senguttuvan, Nithyashree

For team projects:

List of all students (Last, First):

List of collaborative personnel (excluding team participants):

I certify that I have listed above all the sources that I consulted regarding this assignment, and that I have not received or given any assistance that is contrary to the letter or the spirit of the collaboration guidelines for this assignment. (print name here if using a word processor).

Assignment Project Participant(s): Senguttuvan, Nithyashree

Today's Date: 30th March 2023

School of EECS, Washington State University

Washington State University

Computational Genomics

Programming Project 2

Submitted by:
Nithyashree Senguttuvan
011808333

March 31, 2023

Contents

1 System Configuration

CPU used: Intel Core i7-1195G7
Clock rate: 2.92 GHz
RAM: 16 GB
Cache size: L1: 320 KB, L2: 5.0 MB, L3: 12.0 MB

2 Construction Performance

Table 1: Performance Table

Input	**Running Time**
s1	0.007324 (ms)
s2	0.012939 (ms)
Opsin gene in human	0.886230 (ms)
Opsin gene in mouse	0.862061 (ms)
Human BRCA2 gene	4.066895 (ms)
Tomato's Chloroplast Genome	94.932129 (ms)
Yeast's Chromosome 12	815.978027 (ms)
Covid-19 Wuhan China	12.679932 (ms)
Covid-19 USA	7.339111 (ms)
Covid-19 Australia	14.485107 (ms)
Covid-19 India	11.772949 (ms)
Covid-19 Brazil	10.760986 (ms)

A detailed performance statistics for each input is provided in the "Statistics Output.txt" file.

3 Justification

The performance insights reported above are in line with what I anticipated. As the length of the input sequence affects how long it takes to create a ST, I anticipated that longer input sequences would require more time to produce and is proportional. As the tested inputs can be handled by the system configuration utilized for this investigation, I do not anticipate any appreciable performance gains from upgrading to a more powerful system.

4 BWT Index

The BWT (Burrows-Wheeler Transform) index for each input is provided in a separate file, named after the input sequence name of the file.

5 Implementation Constant

Table 2: Implementation Constant (bytes consumed by code for every input byte) Table

Input	Running Time
s1	950
s2	512
Opsin gene in human	24
Opsin gene in mouse	24
Human BRCA2 gene	23
Tomato's Chloroplast Genome	22
Yeast's Chromosome 12	23
Covid-19 Wuhan China	22
Covid-19 USA	23
Covid-19 Australia	22
Covid-19 India	23
Covid-19 Brazil	22

6 Exact Matching Repeat

For each input sequence, the longest exact matching repeat and its start position have been identified. The results are as follows:

Table 3: Exact Matching Repeat Table

Input	Length	Start Coordinates
s1	3	4, 2,
s2	4	5, 2,
Opsin gene in human	18	1818, 1811,
Opsin gene in mouse	14	2840, 2839,
Human BRCA2 gene	14	6165, 6405,
Tomato's Chloroplast Genome	48	88151, 88131,
Yeast's Chromosome 12	8375	451419, 460556,
Covid-19 Wuhan China	32	29872, 29871,
Covid-19 USA	17	3255, 14836,
Covid-19 Australia	32	29862, 29861,
Covid-19 India	17	3242, 14823,
Covid-19 Brazil	17	3255, 14836,

The method searches a given sequence for the longest exact matching repeat. The deepest internal node in the suffix tree is located first by the algorithm, which then prints out its path label, which denotes the lengthiest exact matching repeat along with its length. The method then locates the leaf nodes that are descendants of the deepest internal node, printing out the places of the repetition.

A pointer to the first child of the deepest internal node in the suffix tree is initialized at the start of the method. The characters in the sequence that match the path label of the

deepest internal node are then iterated through and printed. The loop runs from the ID of the deepest internal node's first child minus one to the sum of the ID of the deepest internal node and its depth minus one. By doing this, the loop is guaranteed to iterate through the appropriate set of characters in the sequence.

The function first displays the path label before publishing the depth of the deepest internal node followed by the length of the longest exact matching repetition.

By iterating through the children of the deepest internal node and reporting them, the method then writes out the places of the repetition. The places in the sequence where the repeats occur are indicated .