INF503_Assignment_5_mp2525 - Solutions

Problem #1 (of 1): Prefix trie

A. Generate 5K, 50K, and 100K random 36-mers from the SARS-CoV2 genome sequence (Appendix A) and store them in the prefix trie. Hint: generate a random starting position somewhere in the genome and read 36 characters starting from that position.

- What are the sizes of the trie (# of nodes)? Does it make sense to you? Explain why.
- Iterate through all possible 36-mers in the SARS-CoV2 genome, using each to search / traverse the prefix trie with up to 1 mismatch. How many of your 36-mers had a match? Does it make sense? Explain why

Ans:

```
[mp2525@wind ~/large-scale-data-structures/homework5 ]$ cat /scratch/mp2525/trie_q1a.txt
g++ -c homework.cpp -O3 -std=c++11
g++ -c prefixTrie.cpp -O3 -std=c++11
g++ -o homework homework.o prefixTrie.o -std=c++11
The number of arguments passed: 3
The first argument is: /home/mp2525/large-scale-data-structures/homework5/./homework
The second argument is: problem1A
The third argument is: ./genome.txt
==========
Initialized 29903 character sequences.
==========
Generating 5000 Random 36-Mers to the Prefix-Trie
PrefixTrie size (including root node) for 5000 random 36-mers: 140553
==========
0 Mismatch: 4616
1 Mismatch: 0
1+ Mismatch: 25252
**********
Total Matches: 4616
==========
==========
Generating 50000 Random 36-Mers to the Prefix-Trie
PrefixTrie size (including root node) for 50000 random 36-mers: 729205
==========
0 Mismatch: 24998
1 Mismatch: 0
1+ Mismatch: 4870
**********
Total Matches: 24998
==========
==========
Generating 100000 Random 36-Mers to the Prefix-Trie
PrefixTrie size (including root node) for 100000 random 36-mers: 861921
==========
0 Mismatch: 29680
1 Mismatch: 0
1+ Mismatch: 188
**********
Total Matches: 29680
==========
rm -rf *.o homework
```

Sizes of the trie (# of nodes) for 5K, 50K, and 100K random 36-mers.

Random 36-mers	Sizes of the trie (# of nodes)
5,000	140553
50,000	729205
100,000	861921

It does make sense to me because as we insert more random 36-mers into the prefix trie, the size of the prefix trie grows.

The number of matches for 5K, 50K, and 100K random 36-mers.

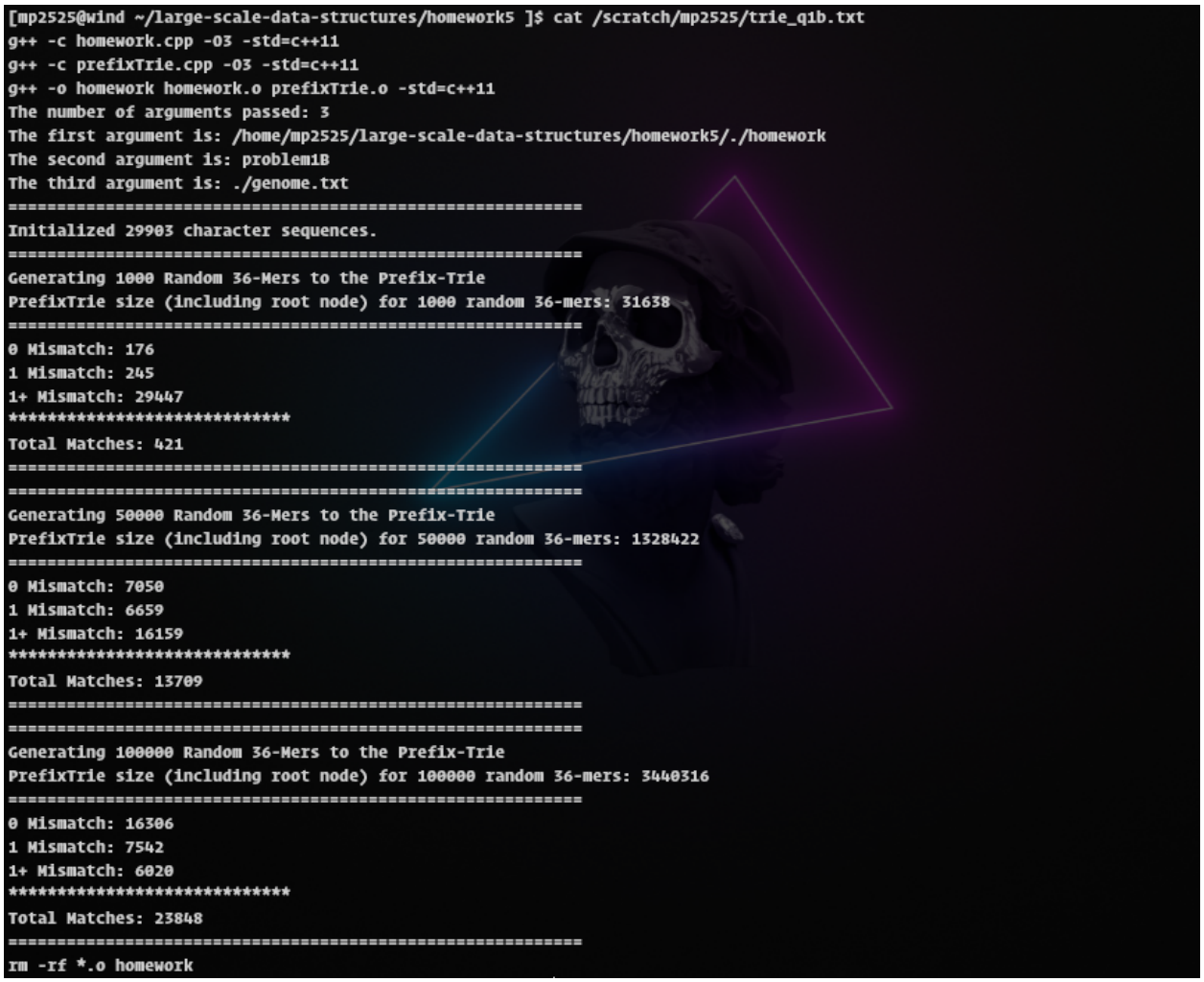
Random 36-mers	Total matches (up to 1 mismatch)
5,000	4616
50,000	24998
100,000	29680

It does make sense to me, as most of the sequence matches are found in the prefix trie. The remaining mismatch is due to the randomized insertion of the sequences from the subject into the prefix trie. It means that some of the query sequences are not inserted in a tree during the randomized insertion though they are generated from the same subject.

B. Generate 1K, 50K, and 100K random 36-mers from the SARS-CoV2 genome sequence with a 5% per-base error rate and store them in the prefix trie. Hint: repeat the process from part A, except each base of 36-mer has a 5% chance of mutation/error.

- What are the sizes of the trie (# of nodes)? Does it make sense to you? Explain why.
- Iterate through all possible 36-mers in the SARS-CoV2 genome, using each to search / traverse the prefix trie with up to 1 mismatch. How many of your 36-mers had a match? Does it make sense? Explain why.

Ans:



```
[mp2525@wind ~/large-scale-data-structures/homework5 ]$ cat /scratch/mp2525/trie_q1b.txt
g++ -c homework.cpp -O3 -std=c++11
g++ -c prefixTrie.cpp -O3 -std=c++11
g++ -o homework homework.o prefixTrie.o -std=c++11
The number of arguments passed: 3
The first argument is: /home/mp2525/large-scale-data-structures/homework5/./homework
The second argument is: problem1B
The third argument is: ./genome.txt
==========================================================
Initialized 29903 character sequences.
==========================================================
Generating 1000 Random 36-Mers to the Prefix-Trie
PrefixTrie size (including root node) for 1000 random 36-mers: 31638
==========================================================
0 Mismatch: 176
1 Mismatch: 245
1+ Mismatch: 29447
****************************
Total Matches: 421
==========================================================
==========================================================
Generating 50000 Random 36-Mers to the Prefix-Trie
PrefixTrie size (including root node) for 50000 random 36-mers: 1328422
==========================================================
0 Mismatch: 7050
1 Mismatch: 6659
1+ Mismatch: 16159
****************************
Total Matches: 13709
==========================================================
==========================================================
Generating 100000 Random 36-Mers to the Prefix-Trie
PrefixTrie size (including root node) for 100000 random 36-mers: 3440316
==========================================================
0 Mismatch: 16306
1 Mismatch: 7542
1+ Mismatch: 6020
****************************
Total Matches: 23848
==========================================================
rm -rf *.o homework
```

Sizes of the trie (# of nodes) for 1K, 50K, and 100K random 36-mers.

Random 36-mers	Sizes of the trie (# of nodes)
1,000	31638
50,000	1328422
100,000	3440316

It does make sense to me because as we have introduced error to the random sequence generation, the trie tree is built based on more random sequences which increased the size of the overall prefix trie. The size is more than the trie we built in problem 1A because of the error.

The number of matches for 5K, 50K, and 100K random 36-mers.

Random 36-mers	Total matches (up to 1 mismatch)
1,000	421
50,000	13709
100,000	23848

It does make sense to me because as we have more random sequences, fewer matches were found in the prefix trie tolerating up to 1 mismatch.