INF-503

Table of Contents

Homework – 5 .... *2*

Commands..7

Video Presentation Link..8

Homework – 5

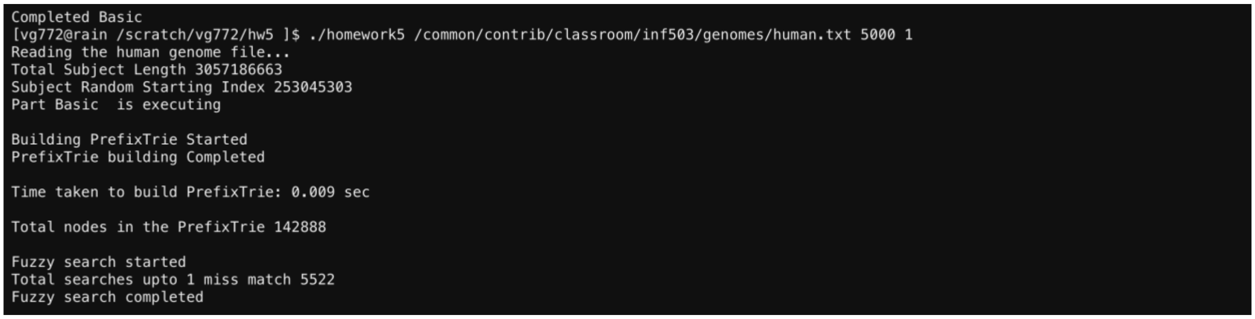
Create a class called Prefix_Trie. The purpose of the class will be to contain a dataset of genomic sequences (queries) and all the functions needed to operate on this set. Use the prefix trie data-structure to store the genomic fragments of a given size. Here you will be performing fuzzy matching, tolerating up to 1 mismatch

Part - A

(20 pts) Basic prefix trie: Pick a random 50K long segment from the human genome assembly. Generate 5K, 50K, 100K, and 1M random 36-mers this segment and store them in the prefix trie. Hint: generate a random starting position somewhere in the segment and read 36 characters
starting from that position.

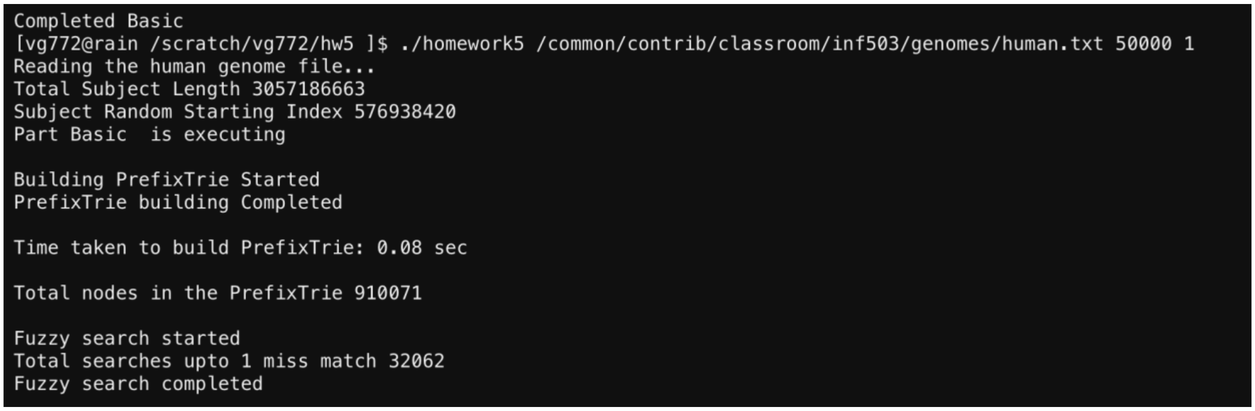
● For each of the 36-mer datasets, what are the sizes of the trie (# of nodes)? Explain the pattern that you observed.

- 5k



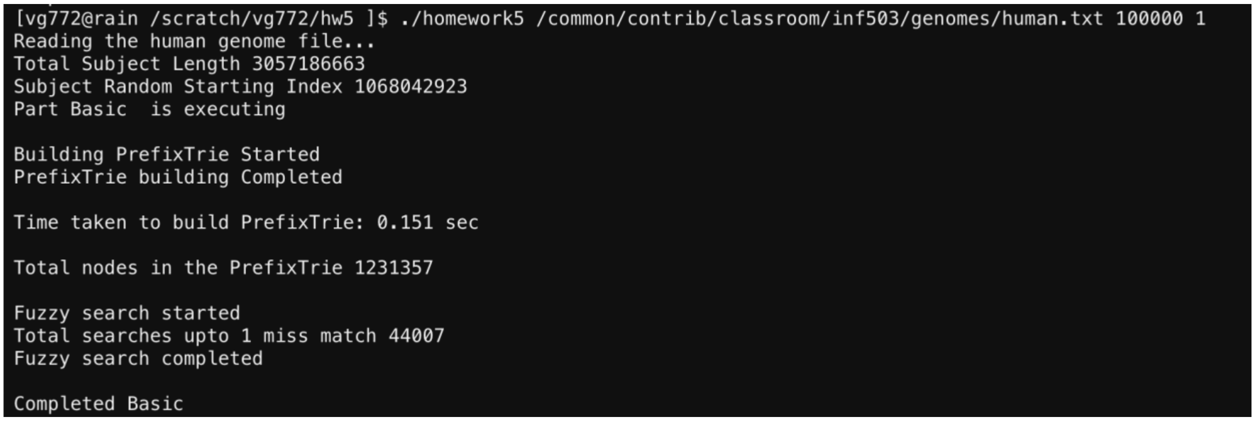
```
Completed Basic
[vg772@rain /scratch/vg772/hw5 ]$ ./homework5 /common/contrib/classroom/inf503/genomes/human.txt 5000 1
Reading the human genome file...
Total Subject Length 3057186663
Subject Random Starting Index 253045303
Part Basic  is executing

Building PrefixTrie Started
PrefixTrie building Completed

Time taken to build PrefixTrie: 0.009 sec

Total nodes in the PrefixTrie 142888

Fuzzy search started
Total searches upto 1 miss match 5522
Fuzzy search completed
```

- 50k



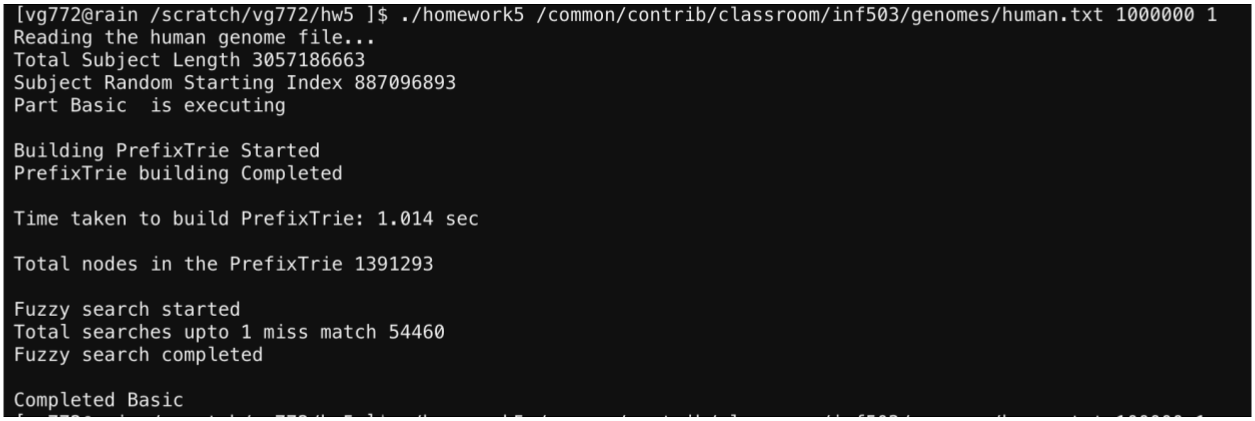
```
Completed Basic
[vg772@rain /scratch/vg772/hw5 ]$ ./homework5 /common/contrib/classroom/inf503/genomes/human.txt 50000 1
Reading the human genome file...
Total Subject Length 3057186663
Subject Random Starting Index 576938420
Part Basic  is executing

Building PrefixTrie Started
PrefixTrie building Completed

Time taken to build PrefixTrie: 0.08 sec

Total nodes in the PrefixTrie 910071

Fuzzy search started
Total searches upto 1 miss match 32062
Fuzzy search completed
```

- 100k



```
[vg772@rain /scratch/vg772/hw5 ]$ ./homework5 /common/contrib/classroom/inf503/genomes/human.txt 100000 1
Reading the human genome file...
Total Subject Length 3057186663
Subject Random Starting Index 1068042923
Part Basic  is executing

Building PrefixTrie Started
PrefixTrie building Completed

Time taken to build PrefixTrie: 0.151 sec

Total nodes in the PrefixTrie 1231357

Fuzzy search started
Total searches upto 1 miss match 44007
Fuzzy search completed

Completed Basic
```

- 1M



```
[vg772@rain /scratch/vg772/hw5 ]$ ./homework5 /common/contrib/classroom/inf503/genomes/human.txt 1000000 1
Reading the human genome file...
Total Subject Length 3057186663
Subject Random Starting Index 887096893
Part Basic  is executing

Building PrefixTrie Started
PrefixTrie building Completed

Time taken to build PrefixTrie: 1.014 sec

Total nodes in the PrefixTrie 1391293

Fuzzy search started
Total searches upto 1 miss match 54460
Fuzzy search completed

Completed Basic
```

● Iterate through all possible 36-mers in the segment, 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.

- 5k - 5522
- 50k - 32062
- 100k - 44007
- 1M – 54460

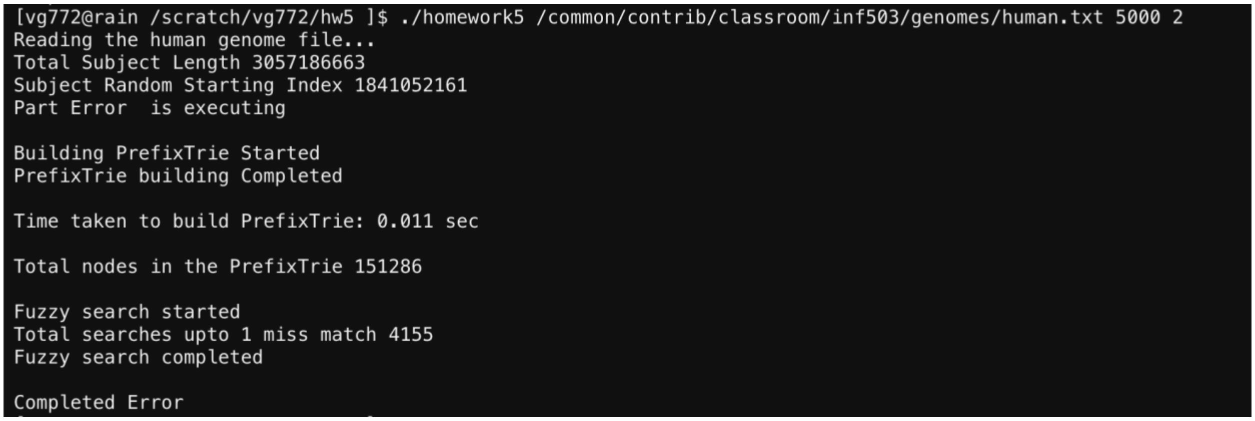
The match count incrased as we incres the total numbers of query size. It does make scence. As we are increasing the query size. For 5k it’s 5522 and 1M 54450

Part - B

(20 pts) Impact of error rate on trie structure: Use the same random 50K long segment from the human genome assembly that you used in part A. Generate 5K, 50K, 100K, and 1M random 36-mers from this segment with 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.

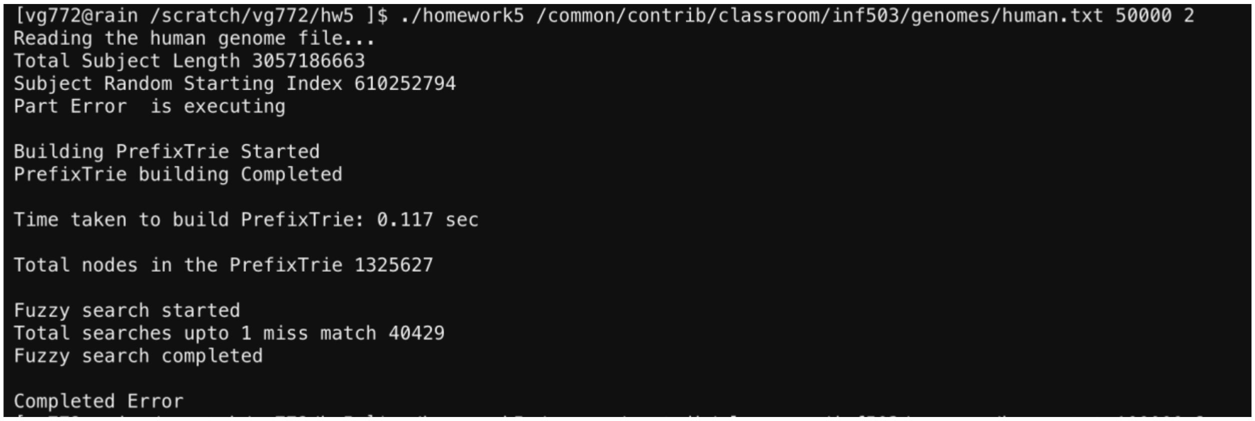
a) For each of the 36-mer datasets, what are the sizes of the trie (# of nodes)? Explain differences (if any) between the trie sizes in partA and part B.

- 5k



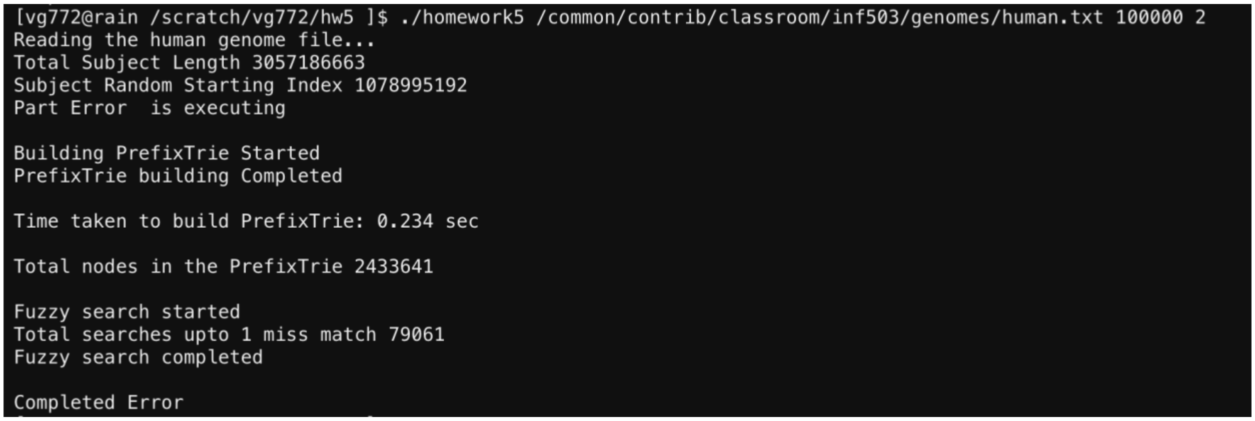
```
[vg772@rain /scratch/vg772/hw5 ]$ ./homework5 /common/contrib/classroom/inf503/genomes/human.txt 5000 2
Reading the human genome file...
Total Subject Length 3057186663
Subject Random Starting Index 1841052161
Part Error  is executing

Building PrefixTrie Started
PrefixTrie building Completed

Time taken to build PrefixTrie: 0.011 sec

Total nodes in the PrefixTrie 151286

Fuzzy search started
Total searches upto 1 miss match 4155
Fuzzy search completed

Completed Error
```

- 50k



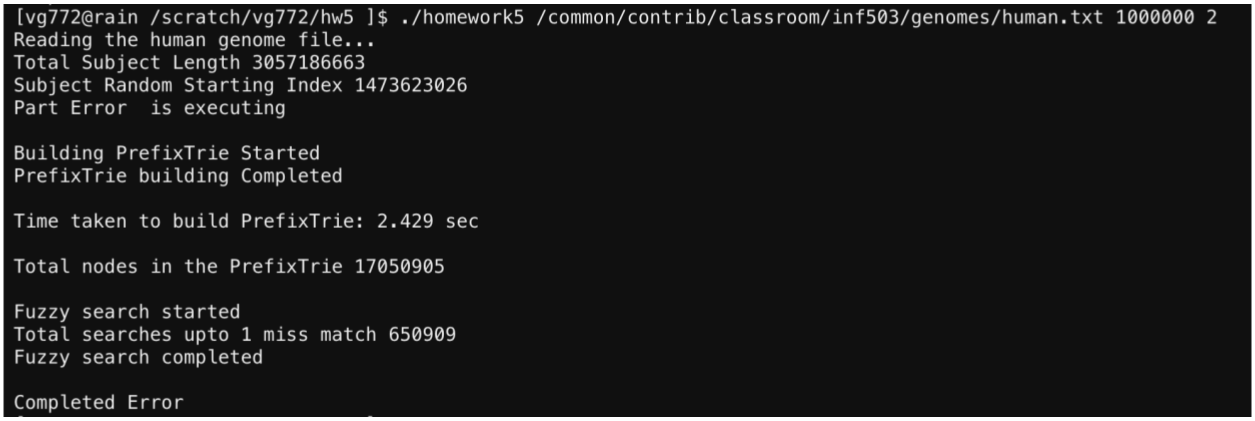
```
[vg772@rain /scratch/vg772/hw5 ]$ ./homework5 /common/contrib/classroom/inf503/genomes/human.txt 50000 2
Reading the human genome file...
Total Subject Length 3057186663
Subject Random Starting Index 610252794
Part Error  is executing

Building PrefixTrie Started
PrefixTrie building Completed

Time taken to build PrefixTrie: 0.117 sec

Total nodes in the PrefixTrie 1325627

Fuzzy search started
Total searches upto 1 miss match 40429
Fuzzy search completed

Completed Error
```

- 100k



```
[vg772@rain /scratch/vg772/hw5 ]$ ./homework5 /common/contrib/classroom/inf503/genomes/human.txt 100000 2
Reading the human genome file...
Total Subject Length 3057186663
Subject Random Starting Index 1078995192
Part Error  is executing

Building PrefixTrie Started
PrefixTrie building Completed

Time taken to build PrefixTrie: 0.234 sec

Total nodes in the PrefixTrie 2433641

Fuzzy search started
Total searches upto 1 miss match 79061
Fuzzy search completed

Completed Error
```

- 1M



```
[vg772@rain /scratch/vg772/hw5 ]$ ./homework5 /common/contrib/classroom/inf503/genomes/human.txt 1000000 2
Reading the human genome file...
Total Subject Length 3057186663
Subject Random Starting Index 1473623026
Part Error  is executing

Building PrefixTrie Started
PrefixTrie building Completed

Time taken to build PrefixTrie: 2.429 sec

Total nodes in the PrefixTrie 17050905

Fuzzy search started
Total searches upto 1 miss match 650909
Fuzzy search completed

Completed Error
```

Compared to part A, part B has a higher number of nodes due to the introduction of a 5% error in the generated sequence. This error introduces unique characters into the prefix trie, which leads to an increase in the number of nodes in the trie. As the error rate introduces variations in the sequences, more unique branches are created, resulting in a more larger trie structure.

b) Iterate through all possible 36-mers in segment, 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.

- 5k - 4155
- 50k - 40429
- 100k - 79061
- 1M – 650909

The match counts show a clear trend when comparing sequences with and without a 5% error rate. For sequences without error, the matches are 5,522 (5k), 32,062 (50k), 44,007 (100k), and 54,460 (1M). Introducing a 5% error results in 4,155 (5k), 40,429 (50k), 79,061 (100k), and 650,909 (1M) matches. This increase in match counts with errors is due to the prefix trie's ability to handle near-matches, allowing up to one mismatch per 36-mer. As the error rate introduces variations, more near-matches are found, particularly with larger query sizes, explaining the higher match counts. This demonstrates the trie's effectiveness in accommodating errors, leading to more matches as the error rate and query size increase.

Part - C
(20 pts) Explain the following in a video recording of duration for at most 8 minutes.

· Using your write-up (.pdf format) that you have submitted, please provide a detailed explanation of your approach to solving the problem. You should cover the following points within a maximum time limit of 3 minutes:

i. Provide a detailed explanation of the answers submitted in the write-up document for both Part A and Part B. Elaborate on why the results obtained are logical, and present your conclusions based on those results.

In Part A, the size of the tries grows with the dataset size, though not in a linear fashion due to the presence of shared prefixes. For smaller datasets, the trie sizes are modest, but they increase significantly as the dataset grows, though not proportionally. Similarly, the number of 36-mers with up to one mismatch that have matches also grows with dataset size.

In Part B, the trie sizes are much larger, reflecting a more diverse set of datasets. As the dataset size increases, the trie sizes become considerably larger. The number of matches for 36-mers with up to one mismatch also increases notably. The larger trie sizes and greater match counts in Part B indicate less redundancy and greater diversity in the sequences compared to Part A.

i. Describe the differences observed in the results between Part A and Part B?
Additionally, explain why these results make sense and why these differences
might have been observed(if any).
ii. Describe the specific bugs and issues you encountered while solving this
assignment. These bugs could be from any part of your code for this homework.
Provide detailed explanations of these challenges, avoiding trivial errors such as
"missing a semicolon in the code."
iii. Highlight at least one specific optimization you made to improve the code's
efficiency or readability.
NOTE:
1. If the video recording duration is less than 3 minutes or exceeds 8 minutes, a
deduction of 50%
will be applied to Part C. This means that 10 points will be deducted from Part C.
2. For the video explanation, you must include both audio and video components while
presenting
your screen. Failure to include these elements will result in receiving 0 points for Part C.
Ensure
your video submissions include your face and code demonstrations.
3. Please utilize Canvas / Zoom / Microsoft Teams for recording your video. Once
recorded, submit

the video link in the write-up file you submit. Ensure that you grant the instructor viewing

```
srun make

srun ./homework5 /common/contrib/classroom/inf503/genomes/human.txt 5000 1
```

Commands

make

5k Basic Prefix

```
./homework5 /common/contrib/classroom/inf503/genomes/human.txt 5000 1
```

50k Basic Prefix

```
./homework5 /common/contrib/classroom/inf503/genomes/human.txt 50000 1
```

100k Basic Prefix

```
./homework5 /common/contrib/classroom/inf503/genomes/human.txt 100000 1
```

1M Basic Prefix

```
./homework5 /common/contrib/classroom/inf503/genomes/human.txt 1000000 1
```

5k Error Prefix

```
./homework5 /common/contrib/classroom/inf503/genomes/human.txt 5000 2
```

50k Error Prefix

```
./homework5 /common/contrib/classroom/inf503/genomes/human.txt 50000 2
```

100k Error Prefix

```
./homework5 /common/contrib/classroom/inf503/genomes/human.txt 100000 2
```

1M Error Prefix

```
./homework5 /common/contrib/classroom/inf503/genomes/human.txt 1000000 2
```

5k Both

```
./homework5 /common/contrib/classroom/inf503/genomes/human.txt 5000 3
```

Video Presentation Link

https://nau.zoom.us/rec/share/cLcrA1mREiJnNFaW0frQ1tdML3Xx5Z32swY2QTvEyTF_oibSDGnEnXMuTypEQwBO.HtKqsUBL8k56dFjG?startTime=1722840208000