LSDS Assignment 5

Command for running Problem : srun homework_5 human_50k_trimmed.txt

A) Basic Prefix trie :

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

For prefix trie of 5000 36-mers number of nodes are 145219

For prefix trie of 50000 36-mers number of nodes are 811901

For prefix trie of 100000 36-mers number of nodes are 396333

For prefix trie of 1000000 36-mers number of nodes are 64907

The pattern I observed here is as the size of the trie kept on increasing the number of nodes in the trie started decreasing this is because there are so many fragments which had similar prefix.

b. 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.

For prefix trie of 5000 36-mers number of hits are 4863

For prefix trie of 50000 36-mers number of hits are 33508

For prefix trie of 100000 36-mers number of hits are 47905

For prefix trie of 1000000 36-mers number of hits are 50288

As the size of the prefix trie increased the number of hits kept increased. This is because we generated the trie from the genome buffer and then now we are trying to iterate the genome buffer and search in the trie. So obviously when the size of trie increases the number of fragments would be repeated a lot which will increase the hits.

B) Impact of error rate on trie structure:

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.

For prefix trie of 5000 36-mers number of nodes are 140513

For prefix trie of 50000 36-mers number of nodes are 1289484

For prefix trie of 100000 36-mers number of nodes are 2212388

For prefix trie of 1000000 36-mers number of nodes are 16471466

The number of nodes in Part B are more compared to part A. This is because we introduced an error rate which would make the fragments not follow the similar prefix. In this case the nodes would be different and we need to add them to the trie which resulted in increase of nodes.

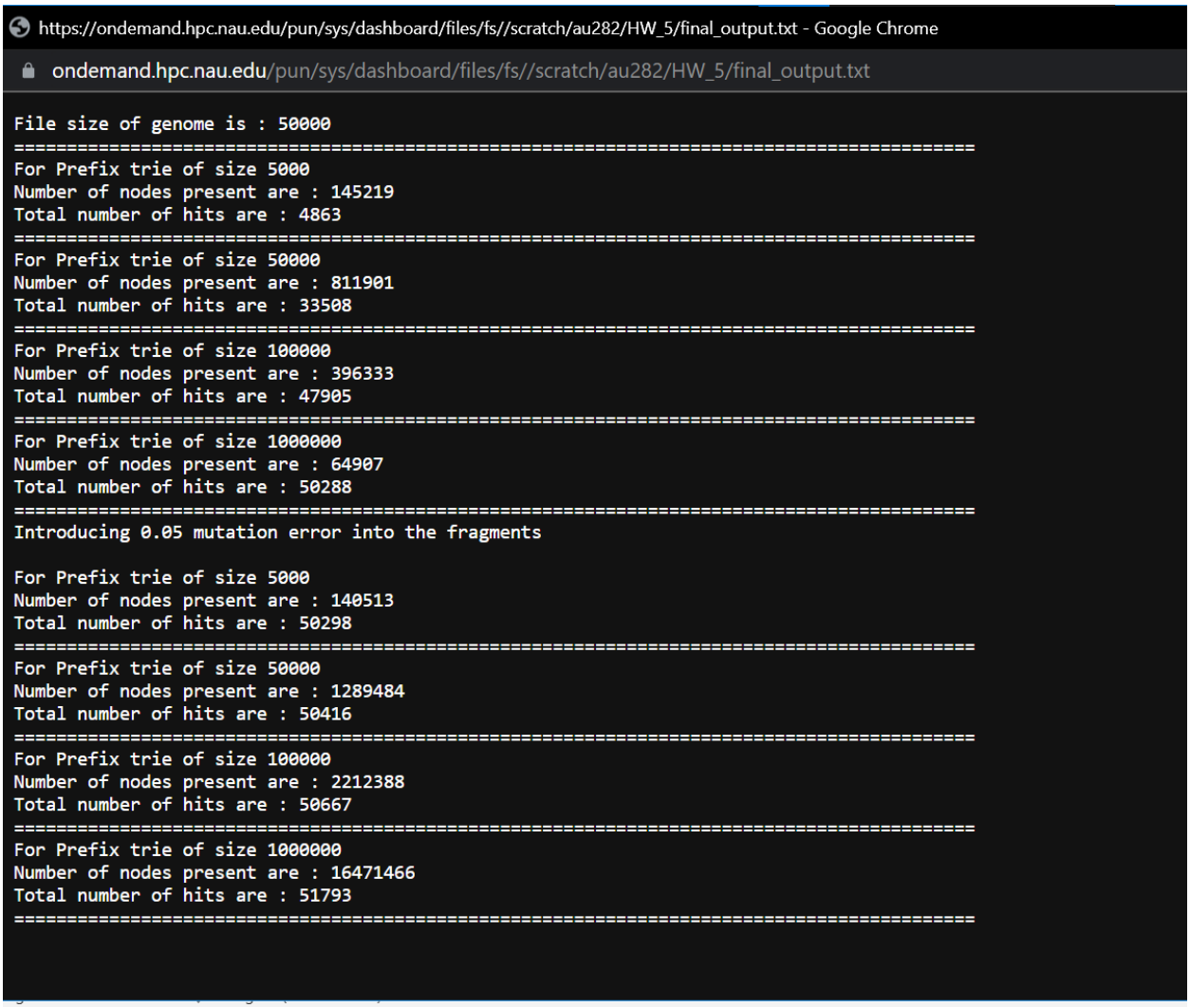
b. Iterate through all possible 36-mers in segment, using each to search / traverse the prefix trie with up to 1 mismatch. How many of your 36-mers had a match? Does it make sense? Explain why.

For prefix trie of 5000 36-mers number of hits are 50298

For prefix trie of 50000 36-mers number of hits are 50416

For prefix trie of 100000 36-mers number of hits are 50667

For prefix trie of 1000000 36-mers number of hits are 51793


https://ondemand.hpc.nau.edu/pun/sys/dashboard/files/fs//scratch/au282/HW_5/final_output.txt - Google Chrome
ondemand.hpc.nau.edu/pun/sys/dashboard/files/fs//scratch/au282/HW_5/final_output.txt
File size of genome is : 50000
For Prefix trie of size 5000
Number of nodes present are : 145219
Total number of hits are : 4863
For Prefix trie of size 50000
Number of nodes present are : 811901
Total number of hits are : 33508
For Prefix trie of size 100000
Number of nodes present are : 396333
Total number of hits are : 47905
For Prefix trie of size 1000000
Number of nodes present are : 64907
Total number of hits are : 50288
Introducing 0.05 mutation error into the fragments
For Prefix trie of size 5000
Number of nodes present are : 140513
Total number of hits are : 50298
For Prefix trie of size 50000
Number of nodes present are : 1289484
Total number of hits are : 50416
For Prefix trie of size 100000
Number of nodes present are : 2212388
Total number of hits are : 50667
For Prefix trie of size 1000000
Number of nodes present are : 16471466
Total number of hits are : 51793