Name: Parth Patel
ID: 6293553

HW5:
Prefix Tries
Usage:
./main --genome <genome_file_path> --fragments <fragment_count> --search <search_type> --mismatch <mismatch_tolerance> --error <error_rate>

Arguments:
–genome : Genome File path
–fragments : Fragment count
–search : Search Type : basic/ mutate

Optional Arguments:
–mismatch: mismatch tolerance (default 1)
–error: Mutation rate (default 5%)

Example: ./main --genome genome.txt --fragments 5000 --search basic
Example: ./main --genome genome.txt --fragments 5000 --mismatch 1 --search mutate --error 5

Part A:

Number of nodes in the tree.

5k:

Nodes: 143915
Matches: 4821

50k:

Nodes: 874361
Matches: 32132

100k:

Nodes: 1234420
Matches: 43217

1M:

Nodes: 1415195
Matches: 49965

Observation on size of the trie:

The size of the trie initially increases as we push more fragments from the genome into the trie, But as we increase the number of fragments more than 50K then size increase is nominal, which indicates that the random fragments taken from the genome are repeating and repeated fragments won't be added to the trie, which is true because we are generating 50k, 100K and 1M fragments from a genome of size 50k, which can only have 49,965 unique fragments, which in turn means that we are pushing duplicates into the trie and it does not increase the size of the trie.

Observations on the number of matches:

The number of matches closely follows the number of queries we inserted into the trie up until we reach the threshold of 49964, this is because we guarantee a match for the queries that were randomly selected and inserted into the trie with up to 1 mismatch, but as we know with randomness we do no guarantee unique guesses each time hence the lower number of matches in my result suggests that we did not push a new query each time when we were adding the queries to the trie, and hence while searching we find less matches for 5k, 50k and 100k, because of random chance some of the 49965 fragments were never pushed inside the trie, but when we push 1M queries, we just improved the probability of each query being picked and pushed in the trie and hence we find all the 49965 queries.

Part B:
Here we inserted queries after introducing mutations(errors) at a 5% rate.
Number of nodes in the tree.

5k:

Nodes: 150996
Matches: 3006

50k:

Nodes: 1312658
Matches: 22952

100k:

Nodes: 2388005
Matches: 35279

1M:

Nodes: 16389635
Matches: 49965

Observation on size of the trie:
Here in contrast to the normal queries, we can see that the number of nodes keeps increasing This is related to the mutation rate, as there is a 1/20(5%) chance of mutation for each fragment. We change the nucleotide to a different one than what was present at one of the 36 spots in a fragment, now what this does is add a variable to the already randomly picked fragments. So, this time even if the same fragment is picked multiple times there is a chance that there will be mutations in the fragment and hence they constitute a branch in the prefix trie and an addition to the number of nodes.

Observations on the number of matches:
Her we can see the same trend in the number of matches as it was in the basic search because And it's not surprising results because even though we are adding mutations to the fragments, The fuzzy matching logic will be able to match the original fragment from which the new mutated query was generated as long as only 1 mutation occurred or else the match will fail, this can be seen in the case of 5k, 50k and 100k, as the all found less than 49965 matches, this can be because of two reasons one because of repeated guesses while adding the fragments and other because of mutations, but when we look at 1M queries pushed, we find all the 49965 fragments, because the probability of each query being picked and pushed was increased drastically and hence all the unique queries were infact present in the trie, in the original or with 5% mutation.

Sample output:
Basic queries:

```
[pp594@wind ~/hw5 ]$ make runa
./main --genome /scratch/pp594/data/human.txt  --fragments 5000 --search basic --mismatch 1
---------------------------------------------------------------
Genome Length: 3057186663
Chosen index: 1954722979
Search Type: basic
Mismatch tolerance: 1

Number of fragments: 5000
Number of nodes: 143915
Number of matches: 4821
---------------------------------------------------------------
./main --genome /scratch/pp594/data/human.txt  --fragments 50000 --search basic --mismatch 1
---------------------------------------------------------------
Genome Length: 3057186663
Chosen index: 906138856
Search Type: basic
Mismatch tolerance: 1

Number of fragments: 50000
Number of nodes: 874361
Number of matches: 32132
---------------------------------------------------------------
./main --genome /scratch/pp594/data/human.txt  --fragments 100000 --search basic --mismatch
1
---------------------------------------------------------------
Genome Length: 3057186663
Chosen index: 348525216
Search Type: basic
Mismatch tolerance: 1

Number of fragments: 100000
Number of nodes: 1234420
Number of matches: 43217
---------------------------------------------------------------
./main --genome /scratch/pp594/data/human.txt  --fragments 1000000 --search basic
--mismatch 1
---------------------------------------------------------------
Genome Length: 3057186663
Chosen index: 1150849705
Search Type: basic
Mismatch tolerance: 1
```

Number of fragments: 1000000
Number of nodes: 1415195
Number of matches: 49964

Mutated queries:
[pp594@wind ~/hw5]$ make runb
./main --genome /scratch/pp594/data/human.txt --fragments 5000 --search mutate --mismatch 1

Genome Length: 3057186663
Chosen index: 1357884313
Search Type: mutate
Mismatch tolerance: 1
Error rate: 5%

Number of fragments: 5000
Number of nodes: 150996
Number of matches: 3006

./main --genome /scratch/pp594/data/human.txt --fragments 50000 --search mutate --mismatch 1

Genome Length: 3057186663
Chosen index: 1088585161
Search Type: mutate
Mismatch tolerance: 1
Error rate: 5%

Number of fragments: 50000
Number of nodes: 1312658
Number of matches: 22952

./main --genome /scratch/pp594/data/human.txt --fragments 100000 --search mutate --mismatch 1

Genome Length: 3057186663
Chosen index: 1130536751
Search Type: mutate
Mismatch tolerance: 1
Error rate: 5%

Number of fragments: 100000
Number of nodes: 2388005
Number of matches: 35279

./main --genome /scratch/pp594/data/human.txt --fragments 1000000 --search mutate --mismatch 1

Genome Length: 3057186663
Chosen index: 85944720
Search Type: mutate
Mismatch tolerance: 1
Error rate: 5%

Number of fragments: 1000000
Number of nodes: 16389635
Number of matches: 49964
