SUBJECT: LARGE SCALE DATA STRUCTURE AND ORGANISATION
Assignment: HOMEWORK 05

By: Saurabh Jawahar Kakade (sk2354@nau.edu)

1. Part 01 – A – 5k random 36-mers

```
[sk2354@ondemand /scratch/sk2354/A05 ]$ make
g++ -c homework.cpp homework.h
g++ -o homework homework.o
[sk2354@ondemand /scratch/sk2354/A05 ]$ srun --mem=10GB -t 00:60:00 ./homework A ./test_genome.fasta 5000 0


The number of arguments passed: 5
The Main file is: /scratch/sk2354/A05/./homework
The special flag is: A
The SARS-COV2 Genome file is: ./test_genome.fasta
The user random size is: 5000
The user error percent is (0 or 5): 0

Copy Constructor initialized !!!
Copy Construction Completed !!!

Problem 01 - Part A:

Number of Prefix Trie Nodes are: 139914
Total matches found are: 4595
Time measured: 0 seconds.
Successfully de-allocated the memory used for Trie!!
Successfully de-allocated the memory used for Trie!!
[sk2354@ondemand /scratch/sk2354/A05 ]$
```

- Size of trie = 139914
- Total 36-mers matched = 4595

2. Part 01 – A – 50k random 36-mers

```
[sk2354@ondemand /scratch/sk2354/A05 ]$ srun --mem=10GB -t 00:60:00 ./homework A ./test_genome.fasta 50000 0


The number of arguments passed: 5
The Main file is: /scratch/sk2354/A05/./homework
The special flag is: A
The SARS-COV2 Genome file is: ./test_genome.fasta
The user random size is: 50000
The user error percent is (0 or 5): 0

Copy Constructor initialized !!!
Copy Construction Completed !!!

Problem 01 - Part A:

Number of Prefix Trie Nodes are: 708473
Total matches found are: 24271
Time measured: 0 seconds.
Successfully de-allocated the memory used for Trie!!
Successfully de-allocated the memory used for Trie!!
[sk2354@ondemand /scratch/sk2354/A05 ]$
```

- Size of trie = 708473
- Total 36-mers matched = 24271

3. Part 01 – A – 100k random 36-mers

```
[sk2354@ondemand /scratch/sk2354/A05 ]$ srun --mem=10GB -t 00:60:00 ./homework A ./test_genome.fasta 100000 0


The number of arguments passed: 5
The Main file is: /scratch/sk2354/A05/./homework
The special flag is: A
The SARS-COV2 Genome file is: ./test_genome.fasta
The user random size is: 100000
The user error percent is (0 or 5): 0

Copy Constructor initialized !!!
Copy Construction Completed !!!

Problem 01 - Part A:

Number of Prefix Trie Nodes are: 837601
Total matches found are: 28820
Time measured: 0 seconds.
Successfully de-allocated the memory used for Trie!!
Successfully de-allocated the memory used for Trie!!
[sk2354@ondemand /scratch/sk2354/A05 ]$
```

- Size of trie = 837601
- Total 36-mers matched = 28820

Total Comparison for number of prefix trie nodes:

- As the number of random 36-mers are increasing from 5k to 50k, the prefix trie nodes are also increasing from 139914 to 708473, respectively. The difference of increase in number of nodes are 708473 – 139914 = 568559.
- When the number of random 36-mers are increasing from 50k to 100k, the prefix trie nodes are also increasing from 708473 to 837601, respectively. The difference of increase in number of nodes are 837601 - 708473 = 129128.
- From the above two points we can see that as we increase the random 36-mers, a smaller number of new nodes are created which are having new sequences than previous.
- At one point, the number of prefix trie nodes will become stable when all of the sequences are covered in prefix trie and it won't increase further.
- Yes, the above points make sense till we achieve the max value of prefix trie and after that it does not makes sense as collisions/redundancy/already-present condition is occurred.

Total Comparison for number of matches found by SARS-CoV2 genome:

- As the number of random 36-mers are increasing from 5k to 50k, the matches found are also increasing from 4595 to 24271, respectively. The difference of increase in number of matched found are 24271– 4595 = 19676.
- When the number of random 36-mers are increasing from 50k to 100k, the matches found are also increasing from 24271 to 28820, respectively. The difference of increase in number of matched found are 28820 - 24271 = 4549.
- From the above two points we can see that as we increase the random 36-mers, a smaller number of new matches were found which are having new sequences than previous.
- At one point, the number of matching nodes will become stable when all the sequences are covered in prefix trie and it will not find any new further.

- Yes, the above points make sense till we achieve the max matches in prefix trie and after that it does not makes sense as no new sequence will be found.

4. Part 01 – B – 1k random 36-mers with 5% error

```
[sk2354@ondemand /scratch/sk2354/A05 ]$ srun --mem=10GB -t 00:60:00 ./homework B ./test_genome.fasta 1000 5


The number of arguments passed: 5
The Main file is: /scratch/sk2354/A05/./homework
The special flag is: B
The SARS-COV2 Genome file is: ./test_genome.fasta
The user random size is: 1000
The user error percent is (0 or 5): 5

Copy Constructor initialized !!!
Copy Construction Completed !!!

Problem 01 - Part B:

Number of Prefix Trie Nodes are: 31367
Total matches found are: 804
Time measured: 0 seconds.
Successfully de-allocated the memory used for Trie!!
Successfully de-allocated the memory used for Trie!!
[sk2354@ondemand /scratch/sk2354/A05 ]$
```

- Size of trie = 31367
- Total 36-mers matched = 804

5. Part 01 – B – 50k random 36-mers with 5% error

```
[sk2354@ondemand /scratch/sk2354/A05 ]$ srun --mem=10GB -t 00:60:00 ./homework B ./test_genome.fasta 50000 5


The number of arguments passed: 5
The Main file is: /scratch/sk2354/A05/./homework
The special flag is: B
The SARS-COV2 Genome file is: ./test_genome.fasta
The user random size is: 50000
The user error percent is (0 or 5): 5

Copy Constructor initialized !!!
Copy Construction Completed !!!

Problem 01 - Part B:

Number of Prefix Trie Nodes are: 1054005
Total matches found are: 21685
Time measured: 0 seconds.
Successfully de-allocated the memory used for Trie!!
Successfully de-allocated the memory used for Trie!!
[sk2354@ondemand /scratch/sk2354/A05 ]$
```

- Size of trie = 1054005
- Total 36-mers matched = 21685

6. Part 01 – B – 100k random 36-mers with 5% error

```
[sk2354@ondemand /scratch/sk2354/A05 ]$ srun --mem=10GB -t 00:60:00 ./homework B ./test_genome.fasta 100000 5


The number of arguments passed: 5
The Main file is: /scratch/sk2354/A05/./homework
The special flag is: B
The SARS-COV2 Genome file is: ./test_genome.fasta
The user random size is: 100000
The user error percent is (0 or 5): 5

Copy Constructor initialized !!!
Copy Construction Completed !!!

Problem 01 - Part B:

Number of Prefix Trie Nodes are: 1690949
Total matches found are: 27536
Time measured: 0 seconds.
Successfully de-allocated the memory used for Trie!!
Successfully de-allocated the memory used for Trie!!
[sk2354@ondemand /scratch/sk2354/A05 ]$
```

- Size of trie = 1690949
- Total 36-mers matched = 27536

Total Comparison for number of prefix trie nodes with 5% error:

- As the number of random 36-mers are increasing from 1k to 50k, the prefix trie nodes are also increasing from 31367 to 1054005, respectively. The difference of increase in number of nodes are 1054005 – 31367 = 1022638.
- When the number of random 36-mers are increasing from 50k to 100k, the prefix trie nodes are also increasing from 1054005 to 1690949, respectively. The difference of increase in number of nodes are 1690949 - 1054005 = 636944.
- From the above two points we can see that as we increase the random 36-mers, a smaller number of new nodes are created which are having new sequences than previous.
- At one point, the number of prefix trie nodes will become stable when all of the sequences are covered in prefix trie and it won't increase further.
- Yes, the above points make sense till we achieve the max value of prefix trie and after that it does not makes sense as collisions/redundancy/already-present condition is occurred.

Total Comparison for number of matches found by SARS-CoV2 genome with 5% error:

- As the number of random 36-mers are increasing from 1k to 50k, the matches found are also increasing from 804 to 21685, respectively. The difference of increase in number of matched found are 21685 – 804 = 20881.
- When the number of random 36-mers are increasing from 50k to 100k, the matches found are also increasing from 21685 to 28820, respectively. The difference of increase in number of matched found are 28820 - 27536= 1284.
- From the above two points we can see that as we increase the random 36-mers, a smaller number of new matches were found which are having new sequences than previous.
- At one point, the number of matching nodes will become stable when all the sequences are covered in prefix trie and it will not find any new further.

- Yes, the above points make sense till we achieve the max matches in prefix trie and after that it does not makes sense as no new sequence will be found.