

LSDS Assignment 3

Command for running Problem A : `srun homework_3 1 human.txt human_reads_trimmed.fa`

Command for running Problem B : `srun homework_3 2 human.txt human_reads_trimmed.fa`

A) Assess the impact of the hash table size :

- For each of your 4 hash table sizes, how many collisions did you observe while populating the hash?

Number of collisions observed while populating the hash table are mentioned in the screenshot below.

```
https://ondemand.hpc.nau.edu/pun/sys/dashboard/files/fs//scratch/au282/HW_3/output_1.txt - Google Chrome
ondemand.hpc.nau.edu/pun/sys/dashboard/files/fs//scratch/au282/HW_3/output_1.txt

File size of genome data is 3057186663
File size of query dataset is 27772294
=====
Total time taken for building the hash table of size 1000000 is 35.0744 seconds
Total time taken for building the hash table of size 10000000 is 36.0569 seconds
Total time taken for building the hash table of size 30000000 is 36.7541 seconds
Total time taken for building the hash table of size 60000000 is 37.3024 seconds
=====
Number of collisions occurred for an hashtable of size 1000000 are 27510406
=====
Number of collisions occurred for an hashtable of size 10000000 are 25869924
=====
Number of collisions occurred for an hashtable of size 30000000 are 22972624
=====
Number of collisions occurred for an hashtable of size 60000000 are 19946973
=====
```

- For each of your 4 hash table sizes, how long did it take you to populate the hash table? Do the timing results make sense? Explain.

There is no major difference here in time for populating the hash table of sizes 1 M till 60 M. This is because while populating the hash table I am inserting the nodes at the beginning of the linked list which optimizes things. As attached in the above screenshot, the time taken for populating the hash table are as follows:

For 1 Million : 35.0744 seconds

For 10 Million : 36.0569 seconds

For 30 Million : 36.7541 seconds

For 60 Million : 37.3024 seconds

B) Searching speed:

- a. How long did it take to search for every possible 16-character long fragment of the subject dataset within the query dataset?

Total time taken for searching every fragment of the subject dataset in the query dataset is 4882.16 seconds.

- b. How many such fragments did you find?

Total number of fragment hits are : 542456454

- c. Print the first 10 fragments of the subject dataset that you found within the Query_HT.

The first 10 fragments of the subject dataset found in Query_HT is mentioned in the below screenshot.

```
https://ondemand.hpc.nau.edu/pun/sys/dashboard/files/fs//scratch/au282/HW_3/output_2.txt - Google Chrome
ondemand.hpc.nau.edu/pun/sys/dashboard/files/fs//scratch/au282/HW_3/output_2.txt

File size of genome data is 3057186663
File size of query dataset is 27772294
=====
Total time taken for building the hash table of size 1000000 is 28.2375 seconds
Total time taken for building the hash table of size 10000000 is 29.637 seconds
Total time taken for building the hash table of size 30000000 is 30.2246 seconds
Total time taken for building the hash table of size 60000000 is 30.6313 seconds
=====
First 10 matching fragments in hashtable of size 60000000 are :

1 Matching fragment in genome data at index 0 is : CTAACCCCTAACCCCTAA
2 Matching fragment in genome data at index 1 is : TAACCCCTAACCCCTAAC
3 Matching fragment in genome data at index 2 is : AACCCCTAACCCCTAACCC
4 Matching fragment in genome data at index 3 is : ACCCTAACCCCTAACCCCT
5 Matching fragment in genome data at index 4 is : CCCTAACCCCTAACCCCTA
6 Matching fragment in genome data at index 5 is : CCTAACCCCTAACCCCTAA
7 Matching fragment in genome data at index 6 is : CTAACCCCTAACCCCTAA
8 Matching fragment in genome data at index 7 is : TAACCCCTAACCCCTAAC
9 Matching fragment in genome data at index 8 is : AACCCCTAACCCCTAACCC
10 Matching fragment in genome data at index 9 is : ACCCTAACCCCTAACCC
=====
Total number of fragment hits are 542456454
=====
Total time taken for finding all the fragments of genome data in hashtable of size 60000000 are 4882.16 seconds
```