Large Scale Data Structures & Organization

Assignment 03: Homework #3

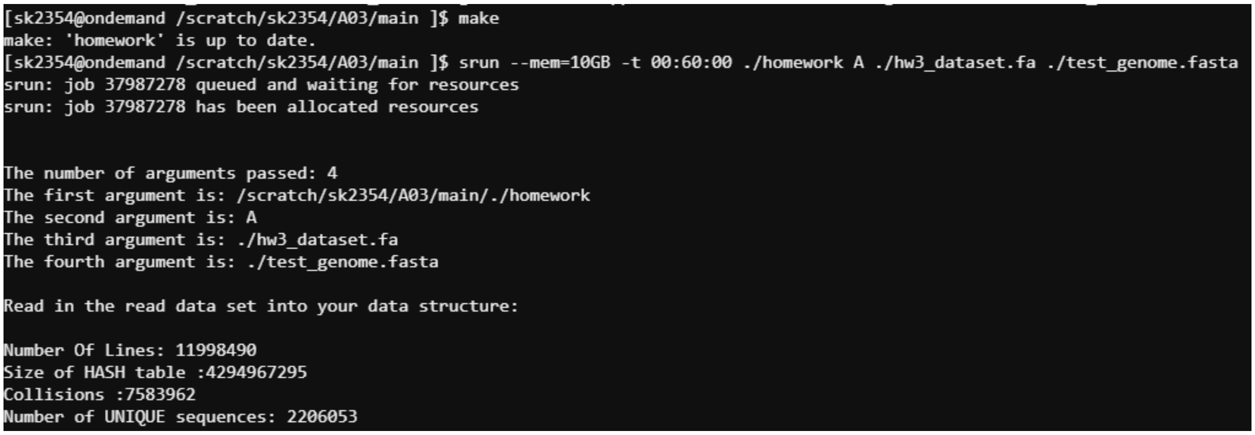
By Saurabh Jawahar Kakade

sk2354@nau.edu

Part 01/a/Solution:

Output:

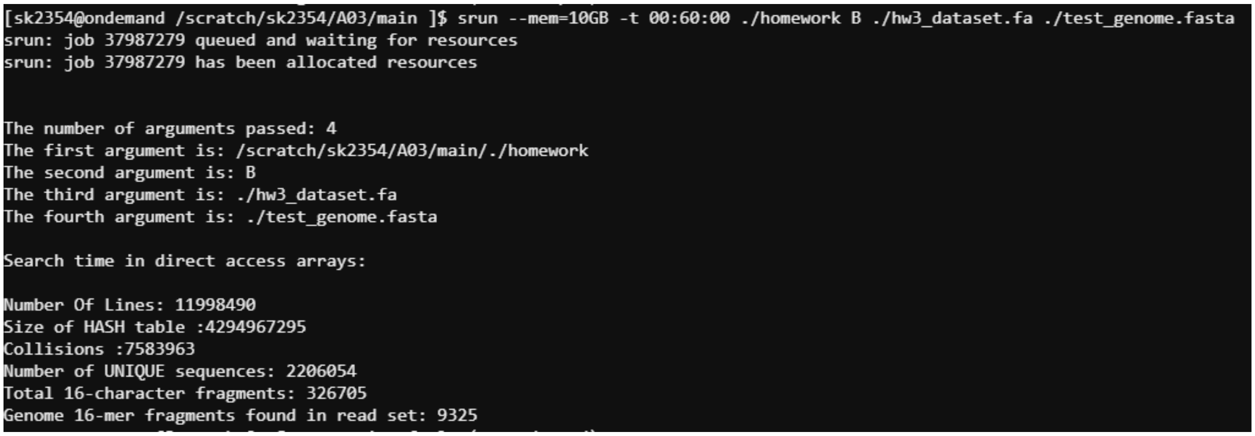
- Size of hash table: 4294967295
- Number of collisions: 7583962
- Unique sequence: 2206053
- Load (αT) in hash table = Unique sequence / Size of hash table
  = 2206053 / 4294967295
  = 0.00051363674



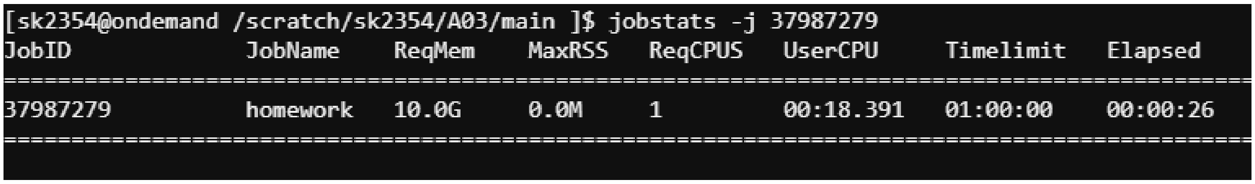
```
[sk2354@ondemand /scratch/sk2354/A03/main ]$ make
make: 'homework' is up to date.
[sk2354@ondemand /scratch/sk2354/A03/main ]$ srun --mem=10GB -t 00:60:00 ./homework A ./hw3_dataset.fa ./test_genome.fasta
srun: job 37987278 queued and waiting for resources
srun: job 37987278 has been allocated resources


The number of arguments passed: 4
The first argument is: /scratch/sk2354/A03/main/./homework
The second argument is: A
The third argument is: ./hw3_dataset.fa
The fourth argument is: ./test_genome.fasta

Read in the read data set into your data structure:

Number Of Lines: 11998490
Size of HASH table :4294967295
Collisions :7583962
Number of UNIQUE sequences: 2206053
```

Part 01/b/Solution:

- Output: Genome 16-mer fragments found are: 9325 and time 26 sec.



```
[sk2354@ondemand /scratch/sk2354/A03/main ]$ srun --mem=10GB -t 00:60:00 ./homework B ./hw3_dataset.fa ./test_genome.fasta
srun: job 37987279 queued and waiting for resources
srun: job 37987279 has been allocated resources


The number of arguments passed: 4
The first argument is: /scratch/sk2354/A03/main/./homework
The second argument is: B
The third argument is: ./hw3_dataset.fa
The fourth argument is: ./test_genome.fasta

Search time in direct access arrays:

Number Of Lines: 11998490
Size of HASH table :4294967295
Collisions :7583963
Number of UNIQUE sequences: 2206054
Total 16-character fragments: 326705
Genome 16-mer fragments found in read set: 9325
```

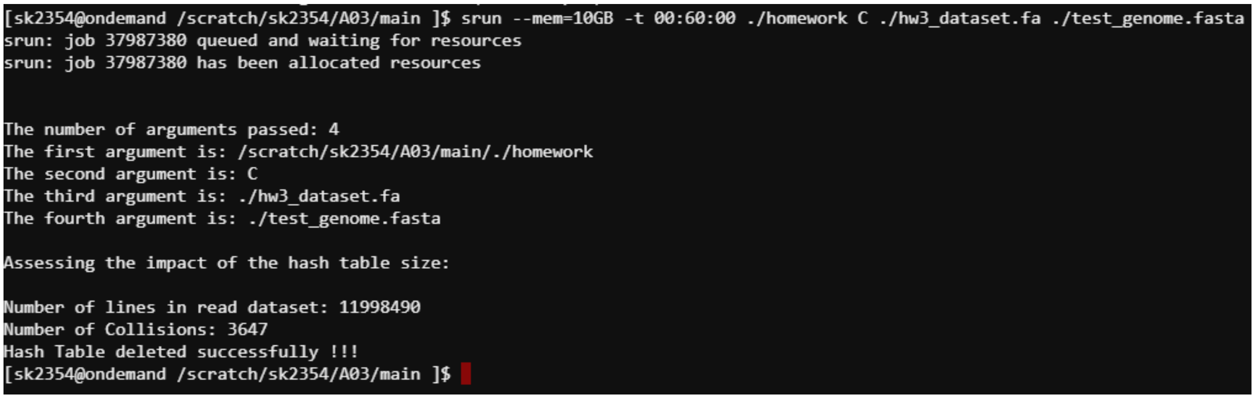


```
[sk2354@ondemand /scratch/sk2354/A03/main ]$ jobstats -j 37987279
JobID           JobName    ReqMem    MaxRSS   ReqCPUS   UserCPU    Timelimit   Elapsed
=====================================================================================
37987279        homework   10.0G     0.0M     1         00:18.391  01:00:00    00:00:26
=====================================================================================
```

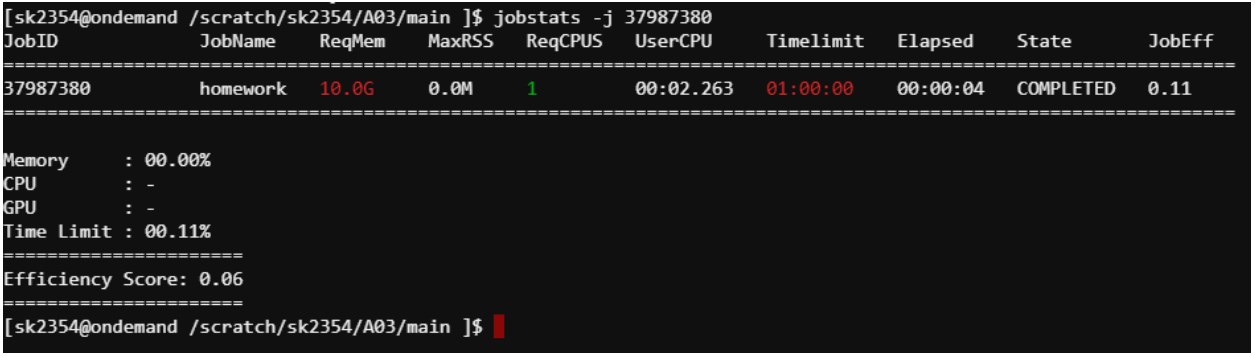
Part 02/a/Solution:

Output:

- For 10,000: Collisions: 3647. Time: 04 sec.

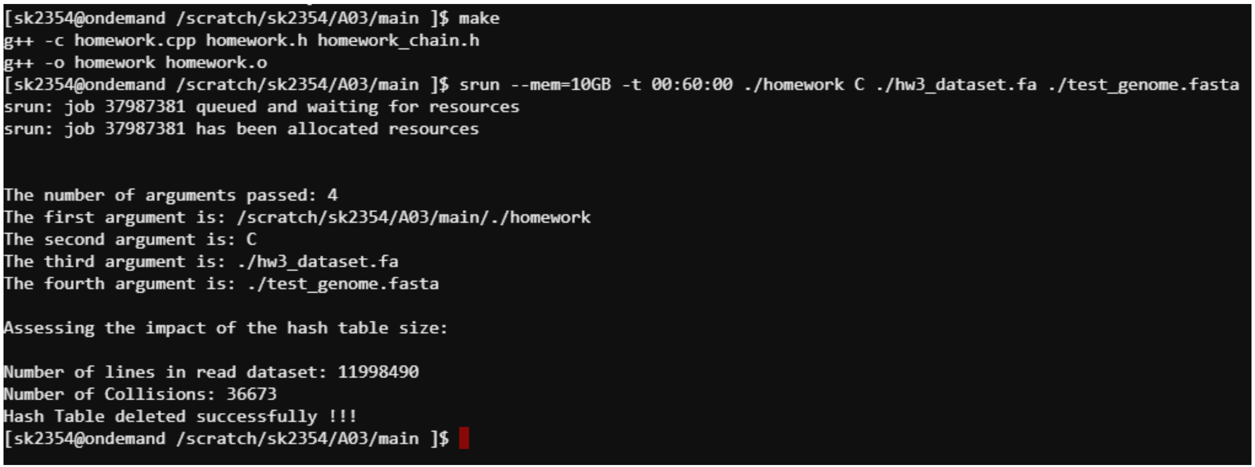


```
[sk2354@ondemand /scratch/sk2354/A03/main ]$ srun --mem=10GB -t 00:60:00 ./homework C ./hw3_dataset.fa ./test_genome.fasta
srun: job 37987380 queued and waiting for resources
srun: job 37987380 has been allocated resources


The number of arguments passed: 4
The first argument is: /scratch/sk2354/A03/main/./homework
The second argument is: C
The third argument is: ./hw3_dataset.fa
The fourth argument is: ./test_genome.fasta

Assessing the impact of the hash table size:

Number of lines in read dataset: 11998490
Number of Collisions: 3647
Hash Table deleted successfully !!!
[sk2354@ondemand /scratch/sk2354/A03/main ]$
```

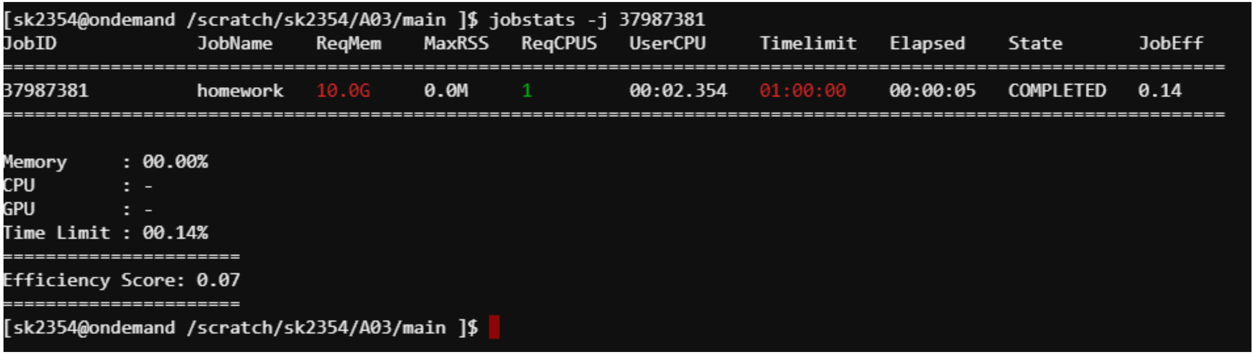


```
[sk2354@ondemand /scratch/sk2354/A03/main ]$ jobstats -j 37987380
JobID            JobName    ReqMem    MaxRSS   ReqCPUS   UserCPU     Timelimit   Elapsed    State       JobEff
==========================================================================================================
37987380         homework   10.0G     0.0M     1         00:02.263   01:00:00    00:00:04   COMPLETED   0.11
==========================================================================================================

Memory     : 00.00%
CPU        : -
GPU        : -
Time Limit : 00.11%
=====================
Efficiency Score: 0.06
=====================
[sk2354@ondemand /scratch/sk2354/A03/main ]$
```

- For 100,000: Collisions: 36673, Time: 05 sec.

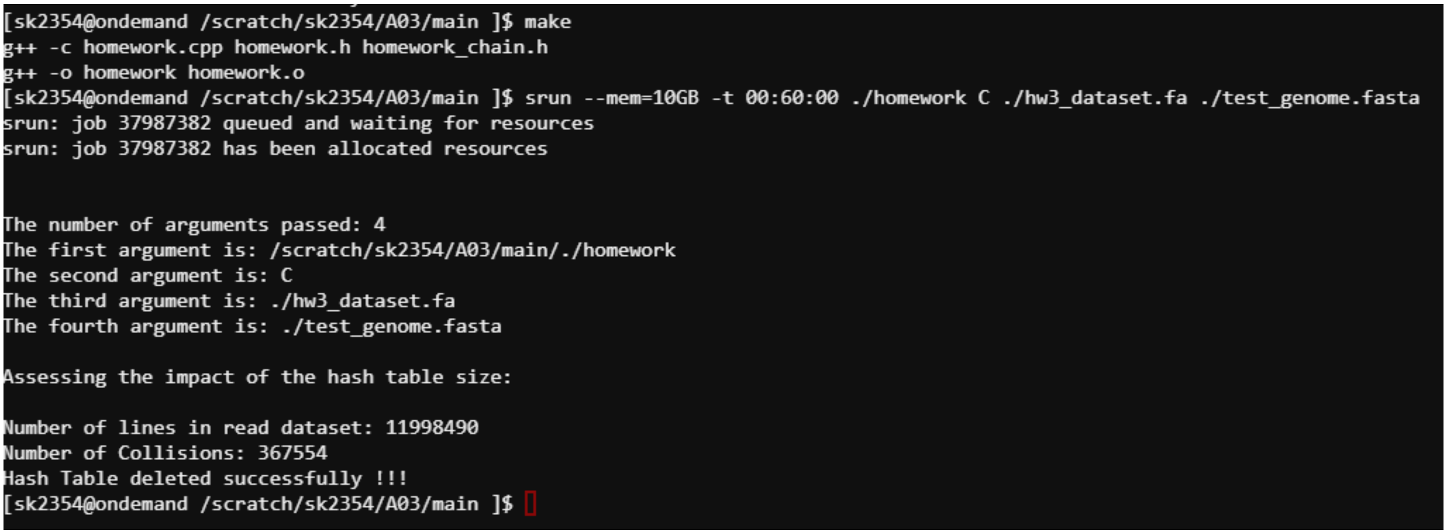


```
[sk2354@ondemand /scratch/sk2354/A03/main ]$ make
g++ -c homework.cpp homework.h homework_chain.h
g++ -o homework homework.o
[sk2354@ondemand /scratch/sk2354/A03/main ]$ srun --mem=10GB -t 00:60:00 ./homework C ./hw3_dataset.fa ./test_genome.fasta
srun: job 37987381 queued and waiting for resources
srun: job 37987381 has been allocated resources


The number of arguments passed: 4
The first argument is: /scratch/sk2354/A03/main/./homework
The second argument is: C
The third argument is: ./hw3_dataset.fa
The fourth argument is: ./test_genome.fasta

Assessing the impact of the hash table size:

Number of lines in read dataset: 11998490
Number of Collisions: 36673
Hash Table deleted successfully !!!
[sk2354@ondemand /scratch/sk2354/A03/main ]$
```

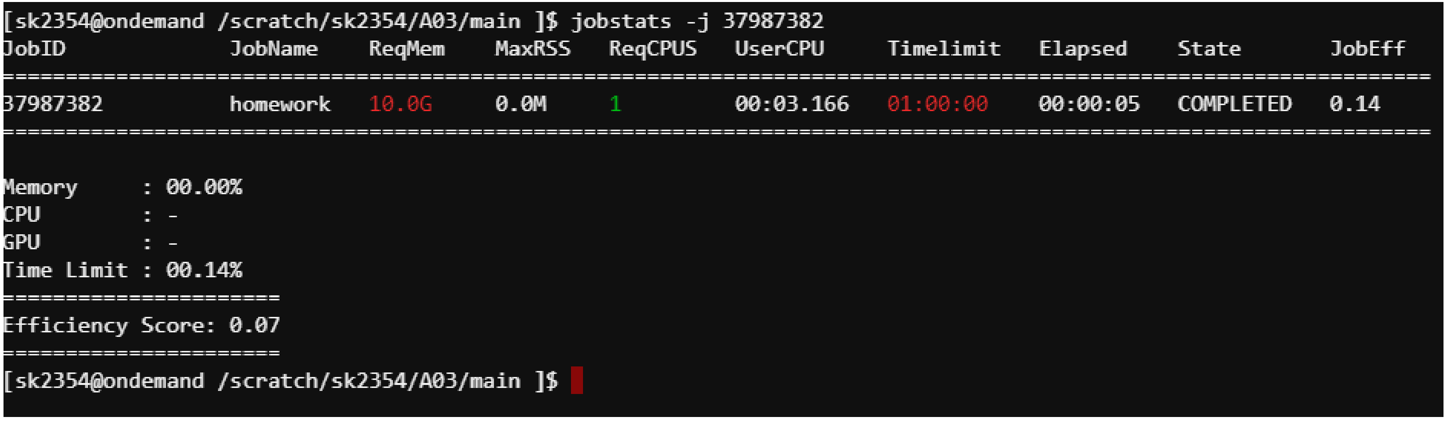


```
[sk2354@ondemand /scratch/sk2354/A03/main ]$ jobstats -j 37987381
JobID            JobName    ReqMem    MaxRSS   ReqCPUS   UserCPU     Timelimit   Elapsed    State       JobEff
==========================================================================================================
37987381         homework   10.0G     0.0M     1         00:02.354   01:00:00    00:00:05   COMPLETED   0.14
==========================================================================================================

Memory     : 00.00%
CPU        : -
GPU        : -
Time Limit : 00.14%
=====================
Efficiency Score: 0.07
=====================
[sk2354@ondemand /scratch/sk2354/A03/main ]$
```

- For 1,000,000: Collisions: 367554, Time: 05 sec.

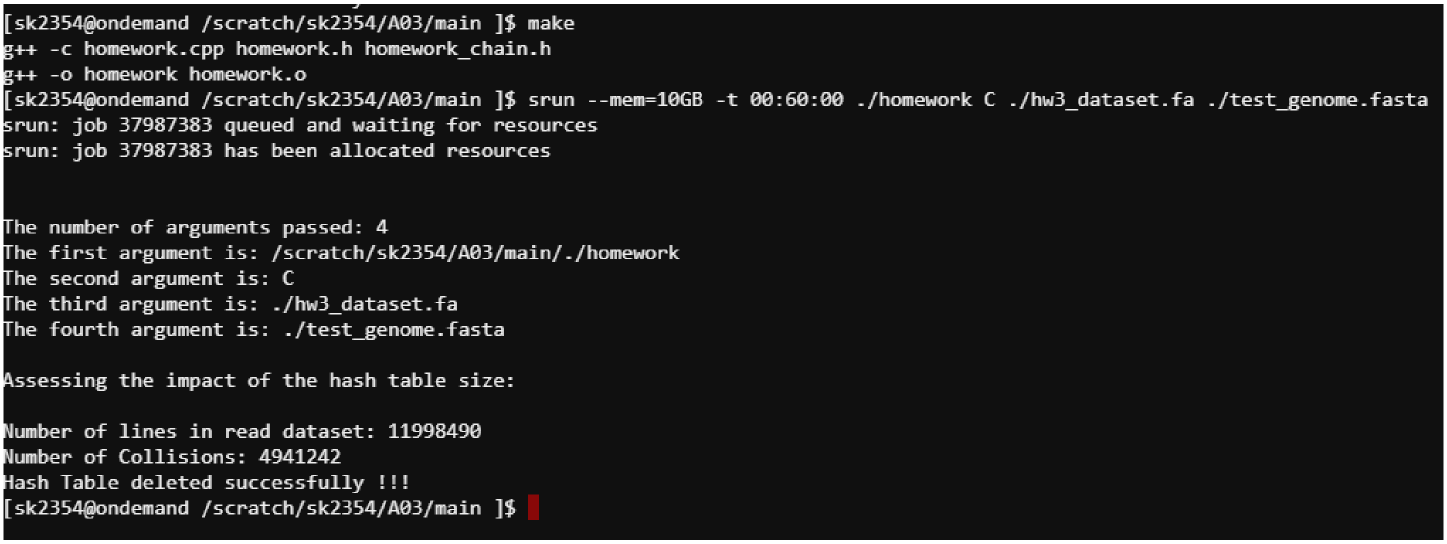


```
[sk2354@ondemand /scratch/sk2354/A03/main ]$ make
g++ -c homework.cpp homework.h homework_chain.h
g++ -o homework homework.o
[sk2354@ondemand /scratch/sk2354/A03/main ]$ srun --mem=10GB -t 00:60:00 ./homework C ./hw3_dataset.fa ./test_genome.fasta
srun: job 37987382 queued and waiting for resources
srun: job 37987382 has been allocated resources


The number of arguments passed: 4
The first argument is: /scratch/sk2354/A03/main/./homework
The second argument is: C
The third argument is: ./hw3_dataset.fa
The fourth argument is: ./test_genome.fasta

Assessing the impact of the hash table size:

Number of lines in read dataset: 11998490
Number of Collisions: 367554
Hash Table deleted successfully !!!
[sk2354@ondemand /scratch/sk2354/A03/main ]$
```

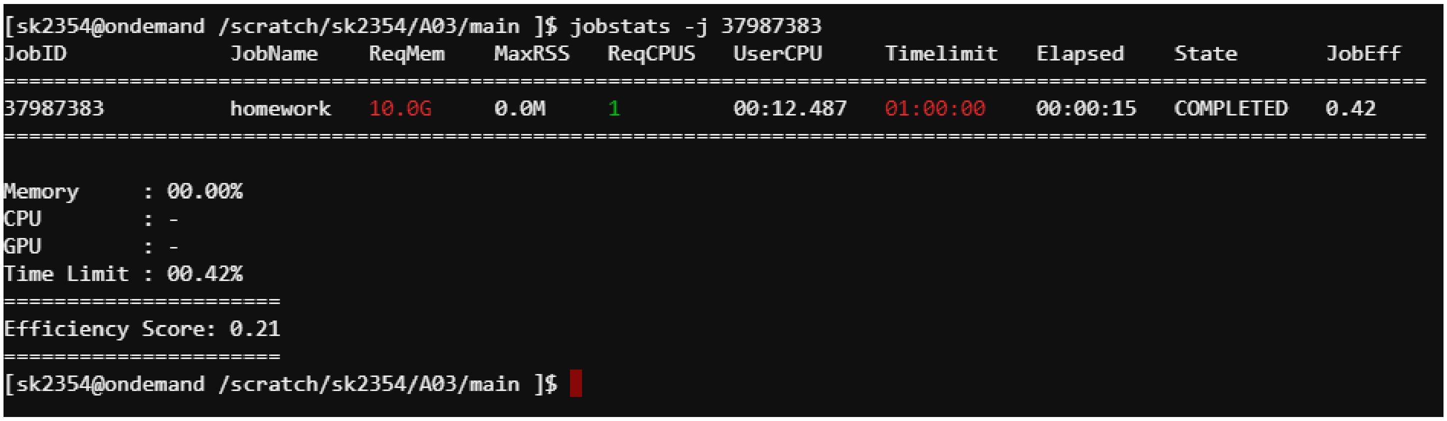


```
[sk2354@ondemand /scratch/sk2354/A03/main ]$ jobstats -j 37987382
JobID              JobName    ReqMem    MaxRSS   ReqCPUS   UserCPU     Timelimit   Elapsed    State       JobEff
===================================================================================================================
37987382           homework   10.0G     0.0M     1         00:03.166   01:00:00    00:00:05   COMPLETED   0.14
===================================================================================================================

Memory     : 00.00%
CPU        : -
GPU        : -
Time Limit : 00.14%
======================
Efficiency Score: 0.07
======================
[sk2354@ondemand /scratch/sk2354/A03/main ]$
```

- For 10,000,000: Collisions: 4941242, Time: 15 sec



```
[sk2354@ondemand /scratch/sk2354/A03/main ]$ make
g++ -c homework.cpp homework.h homework_chain.h
g++ -o homework homework.o
[sk2354@ondemand /scratch/sk2354/A03/main ]$ srun --mem=10GB -t 00:60:00 ./homework C ./hw3_dataset.fa ./test_genome.fasta
srun: job 37987383 queued and waiting for resources
srun: job 37987383 has been allocated resources


The number of arguments passed: 4
The first argument is: /scratch/sk2354/A03/main/./homework
The second argument is: C
The third argument is: ./hw3_dataset.fa
The fourth argument is: ./test_genome.fasta

Assessing the impact of the hash table size:

Number of lines in read dataset: 11998490
Number of Collisions: 4941242
Hash Table deleted successfully !!!
[sk2354@ondemand /scratch/sk2354/A03/main ]$
```

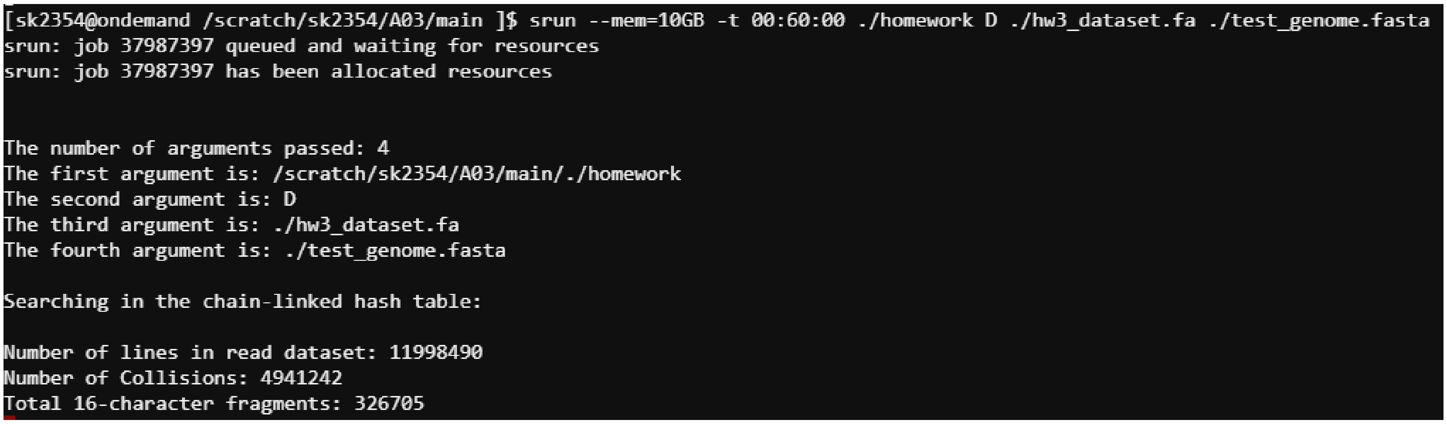


```
[sk2354@ondemand /scratch/sk2354/A03/main ]$ jobstats -j 37987383
JobID            JobName    ReqMem    MaxRSS    ReqCPUS    UserCPU     Timelimit    Elapsed     State       JobEff
=============================================================================================================
37987383         homework   10.0G     0.0M      1          00:12.487   01:00:00     00:00:15    COMPLETED   0.42
=============================================================================================================

Memory     : 00.00%
CPU        : -
GPU        : -
Time Limit : 00.42%
=====================
Efficiency Score: 0.21
=====================
[sk2354@ondemand /scratch/sk2354/A03/main ]$
```

- The result makes a good sense.
- Size of hash table is inversely proportional to number of collisions.
- Higher the hash table size, more area is available for data causing less collisions and vice versa.

Part 02/b/Solution:

Output:



```
[sk2354@ondemand /scratch/sk2354/A03/main ]$ srun --mem=10GB -t 00:60:00 ./homework D ./hw3_dataset.fa ./test_genome.fasta
srun: job 37987397 queued and waiting for resources
srun: job 37987397 has been allocated resources


The number of arguments passed: 4
The first argument is: /scratch/sk2354/A03/main/./homework
The second argument is: D
The third argument is: ./hw3_dataset.fa
The fourth argument is: ./test_genome.fasta

Searching in the chain-linked hash table:

Number of lines in read dataset: 11998490
Number of Collisions: 4941242
Total 16-character fragments: 326705
```

- Total 16 character fragments found are: 326705
- On monsoon it took long time.
- As compare with problem 1b, we can see that time complexity is O(1) and for hash chaining due to linked list it is O(n)