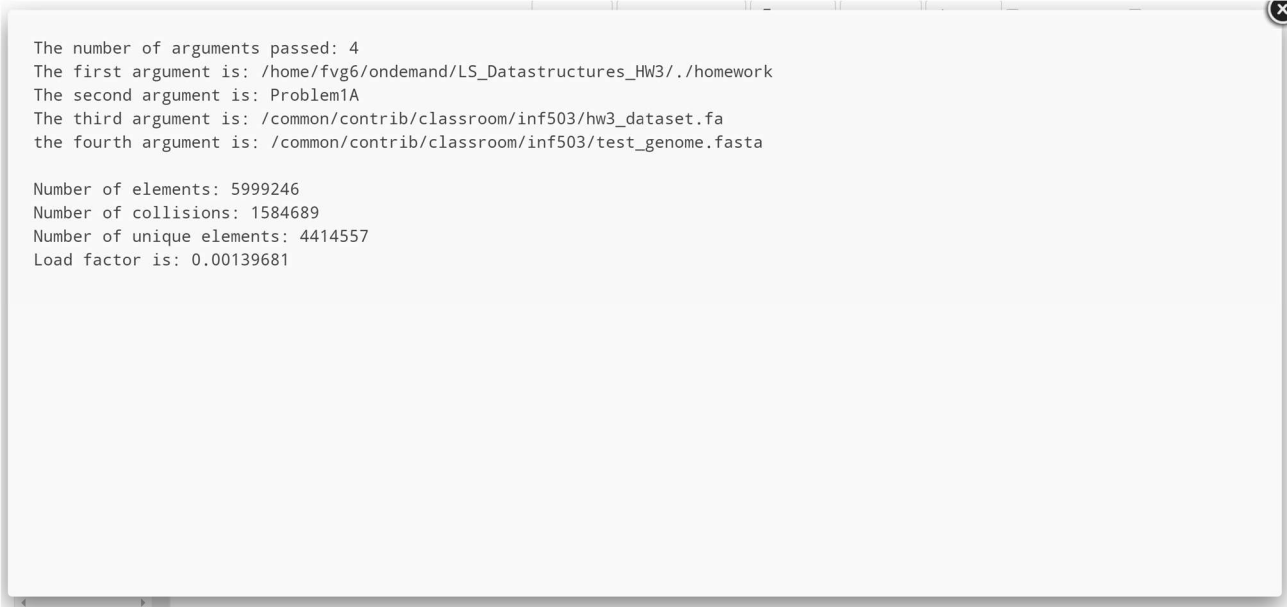


# INF-503 – Homework 3

## Problem 1

A)

A terminal window with a light gray background and a dark gray border. It contains the following text:

```
The number of arguments passed: 4
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW3/./homework
The second argument is: Problem1A
The third argument is: /common/contrib/classroom/inf503/hw3_dataset.fa
the fourth argument is: /common/contrib/classroom/inf503/test_genome.fasta

Number of elements: 5999246
Number of collisions: 1584689
Number of unique elements: 4414557
Load factor is: 0.00139681
```

As can be seen from the output, the size of the hash table is 5,999,246 elements, there were 1,584,689 collisions, and the number of unique elements were 4,414,557. As the direct access hash table has a size of  $4^{16}$ , the load factor was 0.00139681.

B)

```
The number of arguments passed: 4
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW3/./homework
The second argument is: Problem1B
The third argument is: /common/contrib/classroom/inf503/hw3_dataset.fa
the fourth argument is: /common/contrib/classroom/inf503/test_genome.fasta

Number of matching fragments found in read set: 2046995
The entire search process took: 3 seconds
```

As can be seen, 2,046,995 16-mer fragments were found in the read set. It took only 3 seconds to search through the entire read set using a direct access has table.

## Problem 2

A)

```
Size of hash table: 10000
Number of elements: 5999246
Number of collisions: 5989246
Load factor is: 599.925
Time to read in the sequence fragment file: 5

Size of hash table: 100000
Number of elements: 5999246
Number of collisions: 5899246
Load factor is: 59.9925
Time to read in the sequence fragment file: 6

Size of hash table: 1000000
Number of elements: 5999246
Number of collisions: 5051475
Load factor is: 5.99925
Time to read in the sequence fragment file: 7

Size of hash table: 10000000
Number of elements: 5999246
Number of collisions: 2618871
Load factor is: 0.599925
Time to read in the sequence fragment file: 6
```

I have constructed a table:

Table size	Collisions	Time
10,000	5,989,246	5
100,000	5,899,246	6
1,000,000	5,051,475	7
10,000,000	2,618,871	6

These results make sense to me. The number of collisions become smaller as the table size become larger, since there are now more “buckets” for the sequence fragments to be distributed in, and thus the chance that a bucket won’t contain multiple sequence fragments becomes greater. The time it takes to read in the sequence fragment file doesn’t really change however, as the same number of sequence fragments has to be read, regardless of the size of the hash table.

B)

```

The number of arguments passed: 4
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW3/./homework
The second argument is: Problem2B
The third argument is: /common/contrib/classroom/inf503/hw3_dataset.fa
the fourth argument is: /common/contrib/classroom/inf503/test_genome.fasta

Number of matching fragments found in read set: 2046995
The entire search process took: 5 seconds

```

2,046,995 matching fragments are found, the same number as in Problem 1B. This time it took 5 seconds to search through the entire read set, slightly longer than in Problem 1B. This is due to the hash table now using chaining, instead of being direct access. Thus, for this hash table, instead of finding the key and instantly knowing if the fragment is found in the hash table, now we sometimes have to search for the fragment in a short linked list, after finding the key.