INF 503 - Homework 3

Jadon Fowler

To build the homework3 executable:

```
make
```

Problem 1A

```
srun --mem=5000 ./homework3 problem1A 10000
/common/contrib/classroom/inf503/test_genome.fasta
```

```
Loaded Bacillus in 0.021426 seconds.
5227370 fragments can be made from the Bacillus dataset.
...processed 0 fragments in 3e-06s with 0 collisions
...processed 1000000 fragments in 0.172779s with 91654 collisions
...processed 2000000 fragments in 0.365696s with 231827 collisions
...processed 3000000 fragments in 0.575689s with 455078 collisions
...processed 4000000 fragments in 0.802484s with 751312 collisions
...processed 5000000 fragments in 1.04551s with 1118135 collisions
Done reading in FASTA data in 1.10302s
Found 1209638 collisions.
Done!
```

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

Table size	Collisions	Time
10,000	5,217,370	123.232s
100,000	5,127,370	13.626s
1,000,000	4,233,276	2.30687s
10,000,000	1,209,638	1.10302s

These timings make sense because collisions require traversing a LinkedList which is slower than accessing the array when not colliding.

Problem 1B

```
srun --mem=5000 ./homework3 problem1B 10000000
/common/contrib/classroom/inf503/test_genome.fasta
```

```
Reading in FASTA data...
...skipping line: >NC_003997.3
Loaded Bacillus in 0.020563 seconds.
5227370 fragments can be made from the Bacillus dataset.
...processed 0 fragments in 3e-06s with 0 collisions
...processed 1000000 fragments in 0.17291s with 91654 collisions
...processed 2000000 fragments in 0.36621s with 231827 collisions
...processed 3000000 fragments in 0.576413s with 455078 collisions
...processed 4000000 fragments in 0.803565s with 751312 collisions
...processed 5000000 fragments in 1.04714s with 1118135 collisions
Done reading in FASTA data in 1.10483s
Picking 1 million random lines from Bacillus...
Found 1000000 random Bacillus lines in the dataset in 0.190948s.
Generating 1 million random lines...
Found 1165 generated lines in the dataset in 0.489673s.
Done!
```

- Generate 1 million random 16-mers from the *B. anthracis* genome (pick a random start somewhere in the genome and read 16 characters from that point). Search these within your hash table. How many of these fragments did you find and how long did it take?

All 1 million of the fragments randomly picked from the genome were found in the table. This makes sense because the table was filled with all possible fragments from the genome. It took `0.190948` seconds

- Generate 1 million completely random 16-mers (not from the genome): Search these within your hash table. How many of these fragments did you find and how long did it take? Do the timing results make sense? Explain.

`1165` generated lines were found in the table and it took `0.489673` seconds. Since the table size is 10 million, there aren't a lot of collisions. This made the computation really quick.

- Without implementing anything, explain how would the timing for above 2 tasks change if **m** was equal to 10,000.

There would be many more collisions when loading the lines from the genome, which would have caused the time to increase by a couple orders of magnitude.

Problem 1C

```
srun --mem=5000 ./homework3 problem1C 10000000
/common/contrib/classroom/inf503/test_genome.fasta
```

```
Reading in FASTA data...
...skipping line: >NC_003997.3
Loaded Bacillus in 0.020601 seconds.
5227370 fragments can be made from the Bacillus dataset.
...processed 0 fragments in 4e-06s with 0 collisions
...processed 1000000 fragments in 0.172982s with 91654 collisions
...processed 2000000 fragments in 0.366536s with 231827 collisions
...processed 3000000 fragments in 0.577966s with 455078 collisions
...processed 4000000 fragments in 0.805149s with 751312 collisions
...processed 5000000 fragments in 1.04813s with 1118135 collisions
Done reading in FASTA data in 1.10562s
Picking 1 million random lines from Bacillus...
Using 5227370 fragments from the Bacillus dataset as keys...
Found 5227370 Bacillus lines in the dataset in 0.501046s.
Applying 1% error to Bacillus...
Using 5227370 fragments from the Bacillus dataset as keys...
Found 4635532 mutated Bacillus lines in the dataset in 0.524403s.
Done!
```

- Iterate through the *B. anthracis* genome again, generating all possible 16-mers. How many of these 16-mer fragments were found in your hash table?

All `5227370` fragments were found in the table.

- Iterate through the *B. anthracis* genome, introducing a 1% per-base error rate (every character in the 16-mer has a 1% chance to change to something else). (**10 pts** of extra credit if you use an appropriate distribution during simulation). How many of these fragments did you find?

`4635532` fragments were found.