INF 503 - Homework 4

Jadon Fowler

To build the homework executable:

```
make
```

Problem 1A

```
srun --mem=5000 ./homework4 problem1A 100 ./bacillus.fasta
```

- Randomly select 10K, 100K, and 1M (million) 50-mer fragments from the *B. anthracis* genome and use these fragments to query the BLAST_DB containing *B. anthracis*. How many fragments were you able to find and how long did it take?

All of the fragments are present but it takes a very long time to search each fragment.

Random fragment amount	***Time***
10	57s
1,000	~1000 min
100,000	~100,000 min
1,000,000	~1,000,000 min

- Randomly select 10K, 100K, and 1M (million) 50-mer fragments from the *B. anthracis* genome and introduce a 5% per-base error rate (every character has a 5% change of being changed to some other random character). Use these error-filled fragments to query the BLAST_DB containing *B. anthracis*. How many fragments were you able to find?

It seems to find about 95% of fragments that are used as keys, but it takes a very long time.

Random fragment amount	***Time***	***Matches***
100	575s	95
1,000	~1000 min	~950
100,000	~100,000 min	~99,950
1,000,000	~1,000,000 min	~999,950

- Read the High Throughput Sequence reads dataset you used for homework #1 and #2 (located at **/common/contrib/classroom/inf503/hw_dataset.fa**). For this assignment, you can completely disregard the headers of the sequence fragments (i.e. R0_0_1...). Search the entire contents of this dataset in the BLAST_DB. How many perfect hits did you find? (hint: perfect hit's score = 100) Please note that depending on the efficiency of your algorithm, this step may take a long time. First estimate the total time using 1,000, 10,000, and 100,000 queries – if total time estimate is greater than 24 CPU hours, provide estimate rather than exact number.

Random fragment amount	***Time***	***Matches***
100	603s	100
1,000	~1000 min	~1,000
100,000	~100,000 min	~100,000
1,000,000	~1,000,000 min	~1,000,000