INF 503 - Homework 5

Jadon Fowler

My homework 5 & 6 are bundled into one executable.
To build the homework executable:

```
make
```

Problem 1A

```
srun --mem=5000 ./homework5 5A 1000000 covid.txt
```

```
Building trie from SARS-COV2...
The trie contains 867211 nodes.
The trie contains 29867 terminal nodes.
Using 29867 fragments from the SARS-COV2 dataset as keys...
Found 29867 SARS-COV2 lines in the dataset in 0.010097s.
Done!
```

- For each of the 36-mer datasets, what are the sizes of the trie (# of nodes)? Explain the pattern that you observed.
- Iterate through all possible 36-mers in the SARS-CoV2 genome, using each to search / traverse the prefix trie with up to 1 mismatch. How many of your 36-mers had a match? Does it make sense? Explain why.

Table size	Trie nodes	Matches
5,000	140,345	4,607
50,000	710,773	24,348
100,000	837,377	28,812
1,000,000	867,211	29,867

These matches make sense since there are 29,867 possible fragments in the SARS-COV2 dataset.

Problem 1B

```
srun --mem=5000 ./homework5 5B 1000000 covid.txt
```

- For each of the 36-mer datasets, what are the sizes of the trie (# of nodes)? Explain differences (if any) between the trie sizes in partA and part B.
- Iterate through all possible 36-mers in the SARS-CoV2 genome, using each to search / traverse the prefix trie with up to 1 mismatch. How many of your 36-mers had a match? Does it make sense? Explain why.

Table size	Trie nodes	Matches
5,000	139,990	1,185
50,000	707,346	6193
100,000	839,819	7255
1,000,000	868,737	7942

The trie node sizes are almost the same which makes sense since it's storing the same amount of data.

These matches are much lower than the previous ones because we have introduced errors into the genome.