INF 503 - Homework 2

Jadon Fowler

To build the homework2 executable:

```
make
```

Problem 1A

```
./homework2 problem1A /common/contrib/classroom/inf503/hw_dataset.fa
```

- *Which of the following sequences were found in the read set?*

`CTAGGTACATCCACACACAGCAGCGCATTATGTATTTATTGGATTTATTT` **IS** in the dataset.
`GCGCGATCAGCTTCGCGCGCACCGCGAGCGCCGATTGCACGAAATGGCGC` **IS** in the dataset.
`CGATGATCAGGGGCGTTGCGTAATAGAAACTGCGAAGCCGCTCTATCGCC` is **NOT** in the dataset.
`CGTTGGGAGTGCTTGGTTTAGCGCAAATGAGTTTTCGAGGCTATCAAAAA` **IS** in the dataset.
`ACTGTAGAAGAAAAAAGTGAGGCTGCTCTTTTACAAGAAAAAGTNNNNNN` **IS** in the dataset.

- *Would sorting the linked list help speed up the search (on average and in the worst case)? Explain why or why not?*

No, sorting the list would not speed up the search because we still need to traverse the entire list to search. We can't do binary search like we can in a contiguous array.

Problem 1B

```
./homework2 problem1B /common/contrib/classroom/inf503/hw_dataset.fa
/common/contrib/classroom/inf503/test_genome.fasta
```

- *How many 50 character fragments can you make from the B. anthracis genome?*

5,227,336 fragments can be made from the Bacillus dataset.

- *What is the overlap between the genome's 50-mers and the ~36 million fragments you've stored in the FASTAreadset_LL object? 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.*

1,000 queries: 257.542s with 757 overlapping fragments
3,000 queries: 751.194s with 2393 overlapping fragments
5,000 queries: 1142.62s with 4111 overlapping fragments
10,000 queries: 2244.06s with 8323 overlapping fragments
100,000 queries: ~25,000s or ~7 hours with ~75,000 overlapping fragments

- You've iterated through all 50-mers found in the genome and used them to search within the *query read set*. Would it have been faster to flip the problem – i.e. store the genome's fragments in a data structure and iterate through the query read set? Explain why or why not.

I think that would be slower since we'd have to allocate a lot more memory for the Bacillus dataset and traversing the Linked List takes more time than shifting the pointer to an array.