Tyler Thatcher

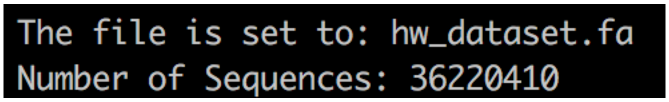
INF503

10/05/2018

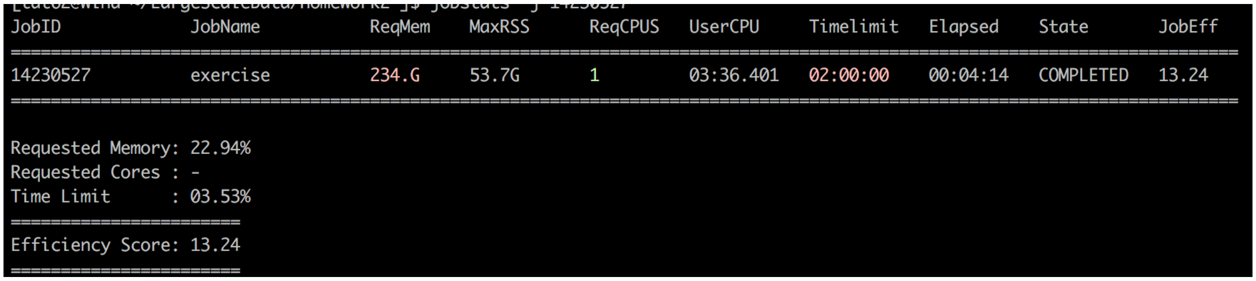
Homework #2

Problem 1A

- The below screenshots show that the entirety of the dataset has been read into memory and how much time and memory it took to read.



```
The file is set to: hw_dataset.fa
Number of Sequences: 36220410
```

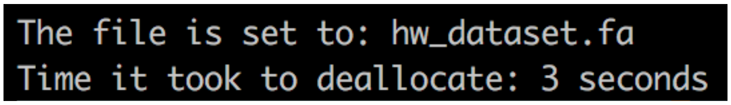


```
JobID            JobName          ReqMem    MaxRSS    ReqCPUS   UserCPU    Timelimit  Elapsed   State      JobEff
===================================================================================================================
14230527         exercise         234.G     53.7G     1         03:36.401  02:00:00   00:04:14  COMPLETED  13.24
===================================================================================================================

Requested Memory: 22.94%
Requested Cores : -
Time Limit      : 03.53%
=====================
Efficiency Score: 13.24
=====================
```

Problem 1B

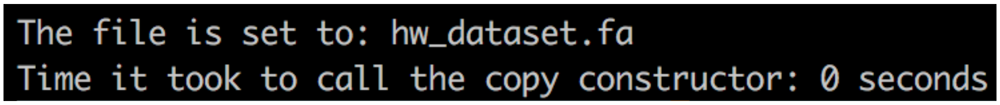
- It took 3 seconds to destroy or delete the 36 million nodes in my linked list. It makes sense because the Destructor doesn't have to do a whole lot except deallocate memory which takes little time.



```
The file is set to: hw_dataset.fa
Time it took to deallocate: 3 seconds
```

Problem 1C

- It took 0 seconds meaning it took milliseconds if not shorter to call the copy constructor. This makes sense because all it has to do is copy the address which I think is just O(1) which is the fastest it can go.

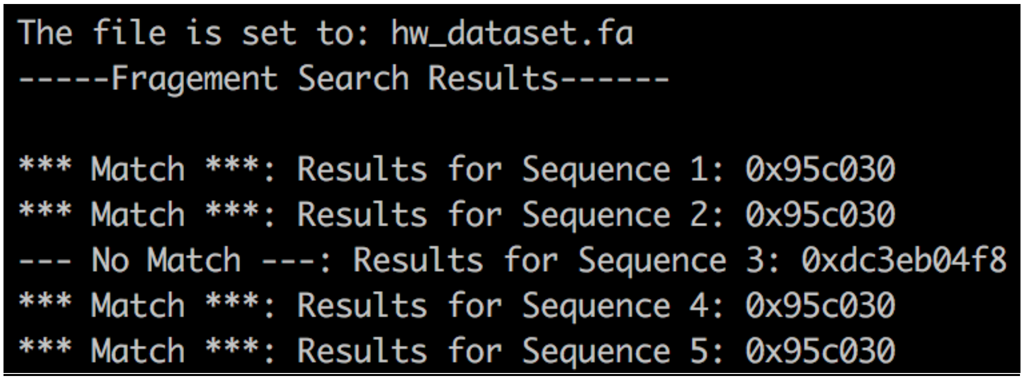


```
The file is set to: hw_dataset.fa
Time it took to call the copy constructor: 0 seconds
```

Problem 1D

1. CTAGGTACATCCACACACAGCAGCGCATTATGTATTTATTGGATTTATTT
2. GCGCGATCAGCTTCGCGCGCACCGCGAGCGCCGATTGCACGAAATGGCGC
3. CGATGATCAGGGGCGTTGCGTAATAGAAACTGCGAAGCCGCTCTATCGCC
4. CGTTGGGAGTGCTTGGTTTAGCGCAAATGAGTTTTCGAGGCTATCAAAAA
5. ACTGTAGAAGAAAAAAGTGAGGCTGCTCTTTTACAAGAAAAAGTNNNNNN

- The below screenshot shows that all sequences except number 3 were found in the dataset. Now, our searches don’t account for ‘N’ at all so that means that anything with ‘NNN’ at the end will result in a match.

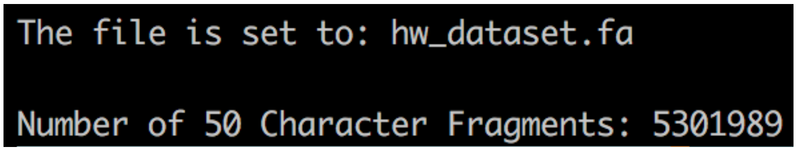


```
The file is set to: hw_dataset.fa
-----Fragement Search Results------

*** Match ***: Results for Sequence 1: 0x95c030
*** Match ***: Results for Sequence 2: 0x95c030
--- No Match ---: Results for Sequence 3: 0xdc3eb04f8
*** Match ***: Results for Sequence 4: 0x95c030
*** Match ***: Results for Sequence 5: 0x95c030
```

Problem 2A

- The number of 50mers found in this genome is shown in the screenshot below.
- The number of 50mers was: **5,301,989.**

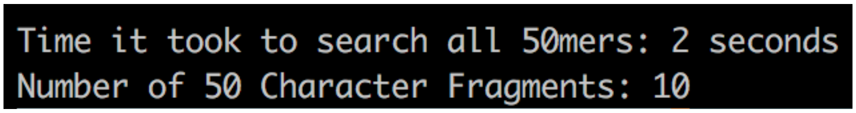


```
The file is set to: hw_dataset.fa

Number of 50 Character Fragments: 5301989
```

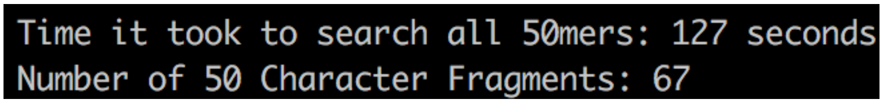
Problem 2B

10 Searches



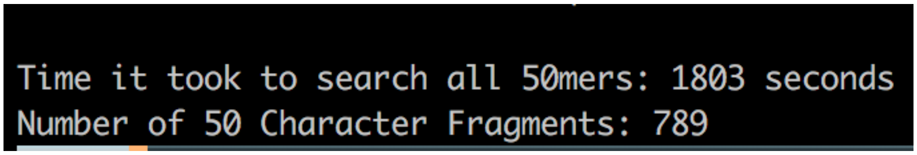
```
Time it took to search all 50mers: 2 seconds
Number of 50 Character Fragments: 10
```

100 Searches



```
Time it took to search all 50mers: 127 seconds
Number of 50 Character Fragments: 67
```

1000 Searches



```
Time it took to search all 50mers: 1803 seconds
Number of 50 Character Fragments: 789
```

- The time increase as N increases is:
    - 127 / 2 = 63.5
    - From 10 – 100 it took 63.5 times longer.
    - 1803 / 127 = 14.2
    - From 100 – 1000 it took 14.2 times longer.

- Amount found as N increases:
    - 67 / 10 = 6.7
    - From 10 – 100 there was 6.7 times more matches
    - 789 / 67 = 11.8
    - From 100 – 1000 there was 11.8 times more matches

- I estimate that the one with 10000 searches would take 1803 * 10 = 18003 seconds ~ 5 hours. Given that logic because 5.2 million / 10000 = 5200. That means that the search could take up to 18000 * 5200. This is obviously the worst case, and that is because of the O(N) search time that happens with a Linked List. I estimate this could take around 1083 days to complete based on the math provided above, but with greater sample sizes maybe it might be closer to O(lg(N)), because you told us it could take about 24-36 hours.