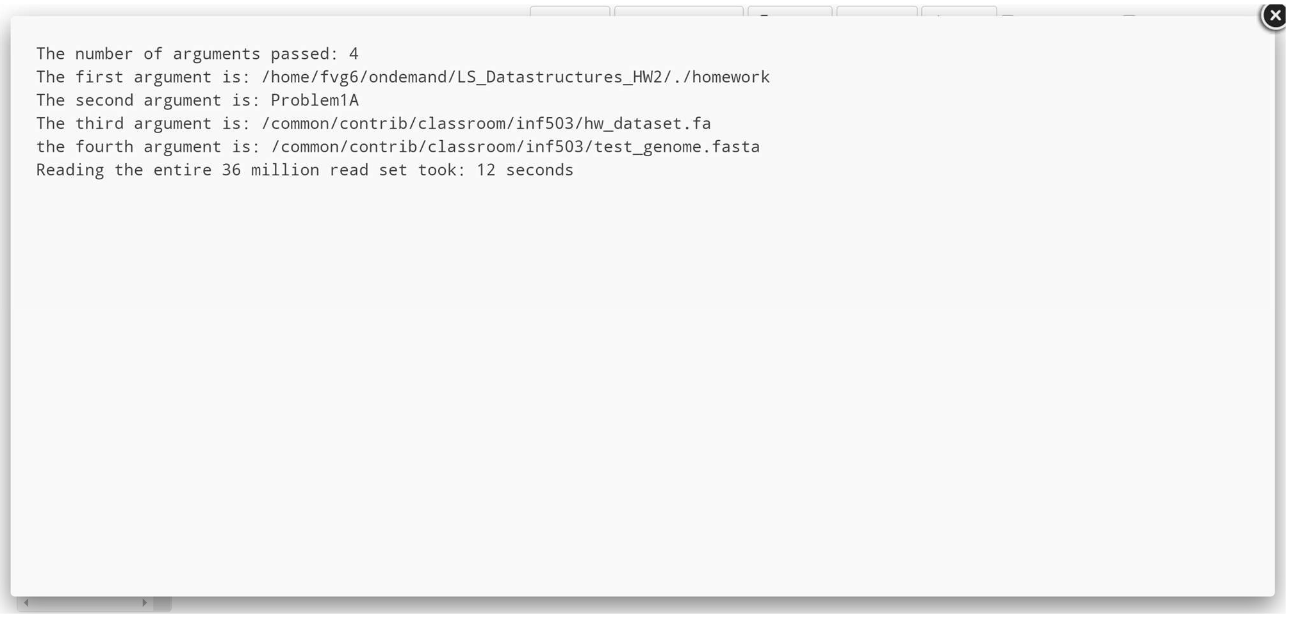
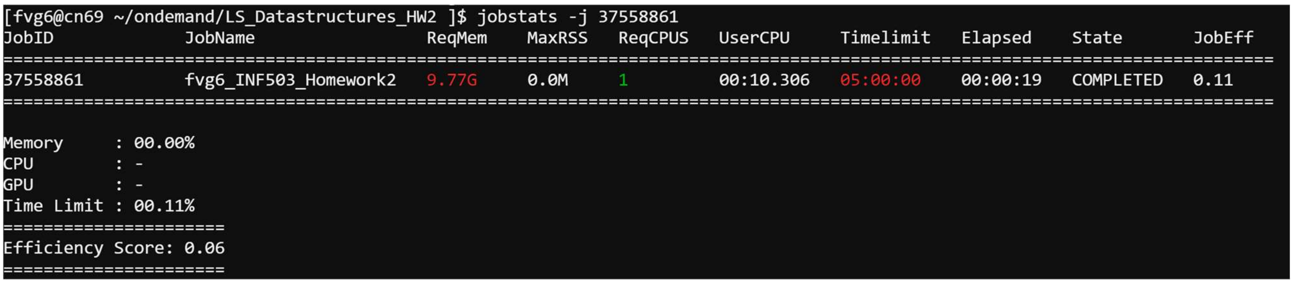
INF-503 – Homework 2

Problem 1

A)



```
The number of arguments passed: 4
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW2/./homework
The second argument is: Problem1A
The third argument is: /common/contrib/classroom/inf503/hw_dataset.fa
the fourth argument is: /common/contrib/classroom/inf503/test_genome.fasta
Reading the entire 36 million read set took: 12 seconds
```



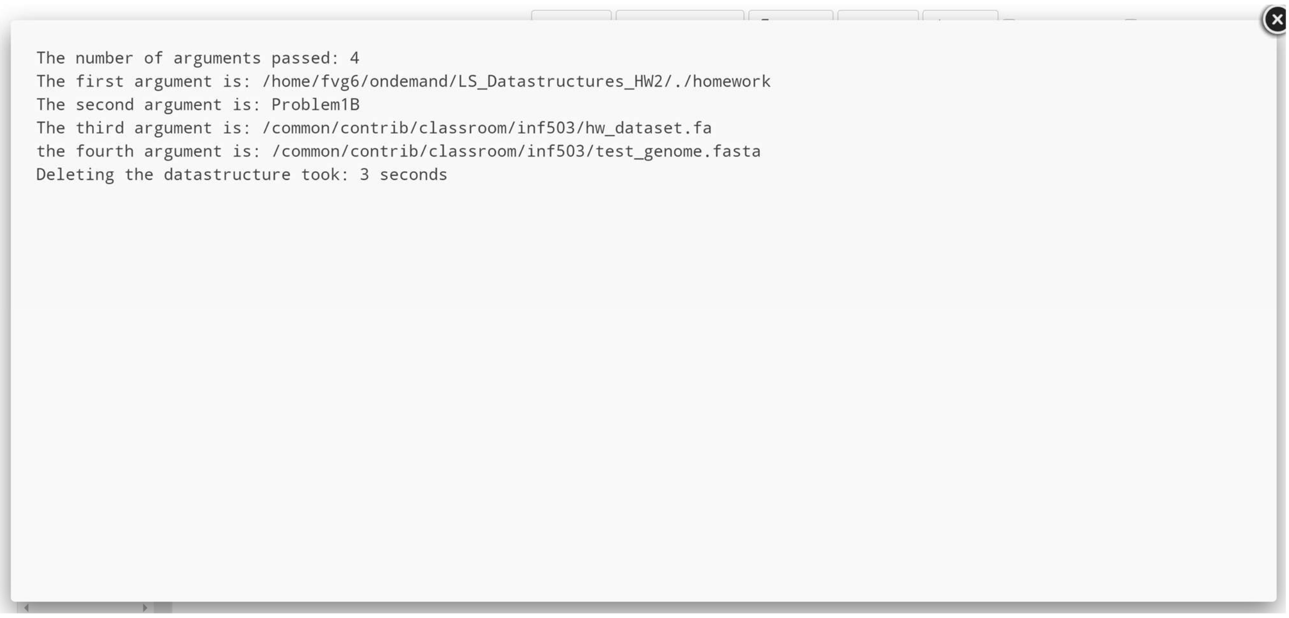
```
[fvg6@cn69 ~/ondemand/LS_Datastructures_HW2 ]$ jobstats -j 37558861
JobID            JobName                ReqMem    MaxRSS    ReqCPUS    UserCPU     Timelimit   Elapsed    State       JobEff
=============================================================================================================================
37558861         fvg6_INF503_Homework2  9.77G     0.0M      1          00:10.306   05:00:00    00:00:19   COMPLETED   0.11
=============================================================================================================================

Memory     : 00.00%
CPU        : -
GPU        : -
Time Limit : 00.11%
=====================
Efficiency Score: 0.06
=====================
```

As can be seen, the total time required to read the entire read set was 12 seconds, measured with the <ctime> library, and 19 seconds measured with the jobstats command.

According to the jobstats command, Monsoon claims that MaxRSS (the maximum amount of memory used to load the data into memory) was 0 Mb.

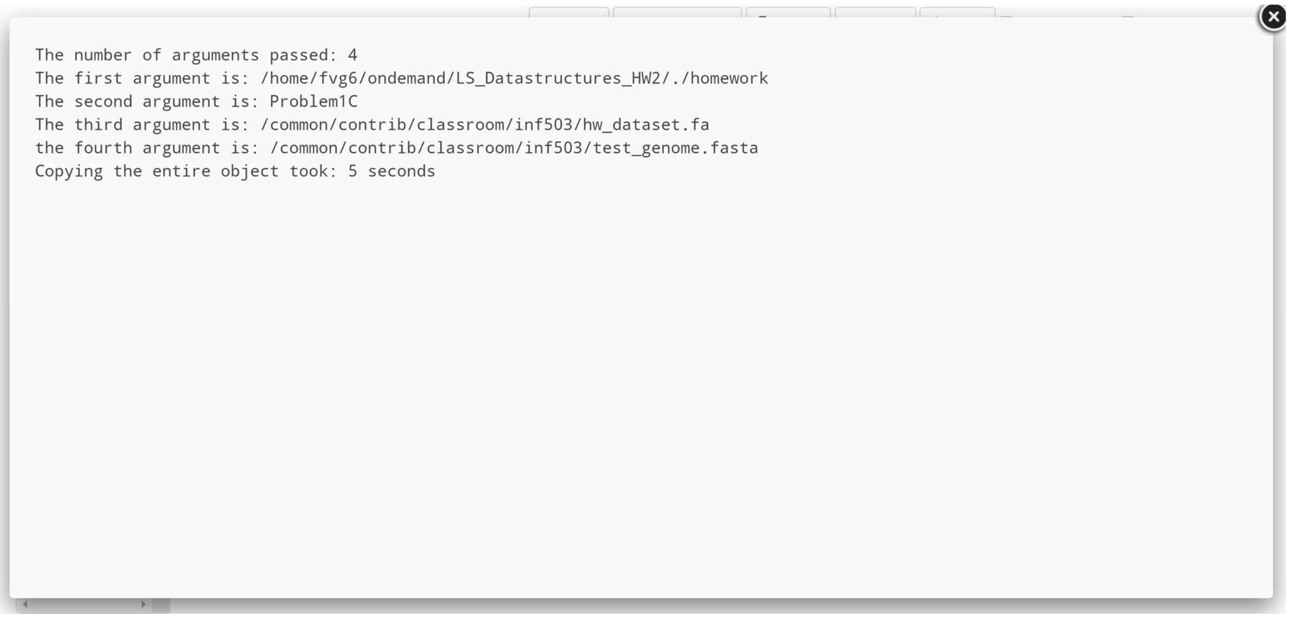
B)



```
The number of arguments passed: 4
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW2/./homework
The second argument is: Problem1B
The third argument is: /common/contrib/classroom/inf503/hw_dataset.fa
the fourth argument is: /common/contrib/classroom/inf503/test_genome.fasta
Deleting the datastructure took: 3 seconds
```

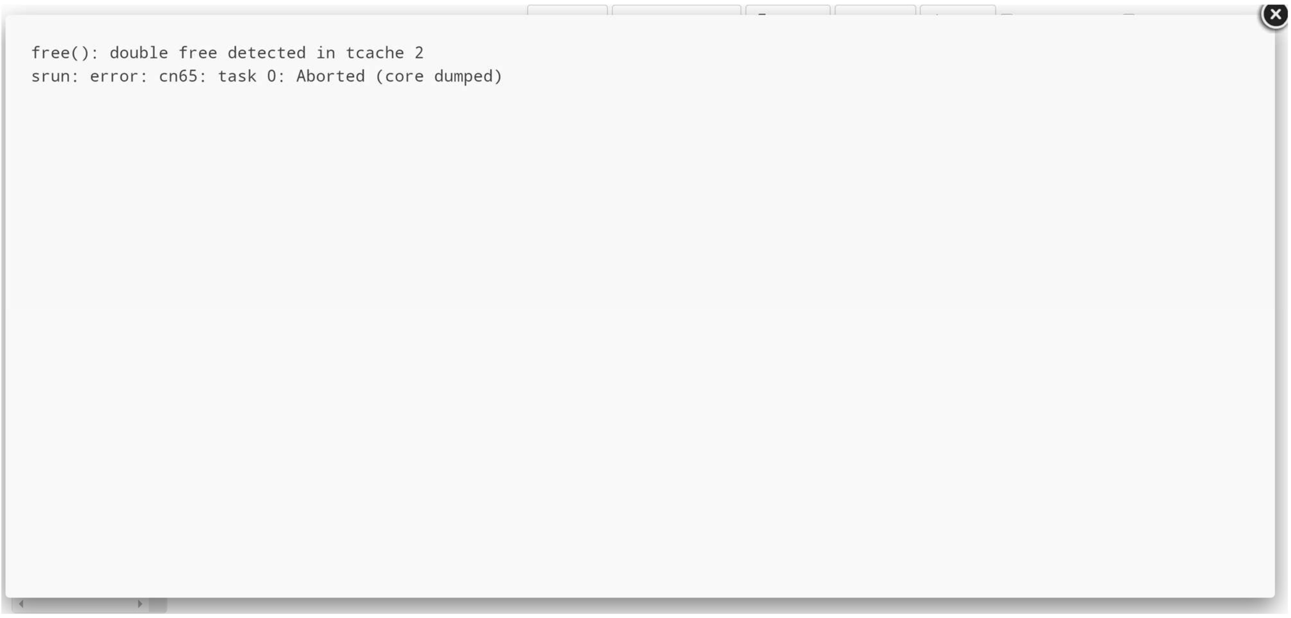
As can be seen, it took 3 seconds for the data to be deallocated. The linked list is deleted from the head down, one node at a time. Thus, the big O notation for this process is O(n).

C)



```
The number of arguments passed: 4
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW2/./homework
The second argument is: Problem1C
The third argument is: /common/contrib/classroom/inf503/hw_dataset.fa
the fourth argument is: /common/contrib/classroom/inf503/test_genome.fasta
Copying the entire object took: 5 seconds
```

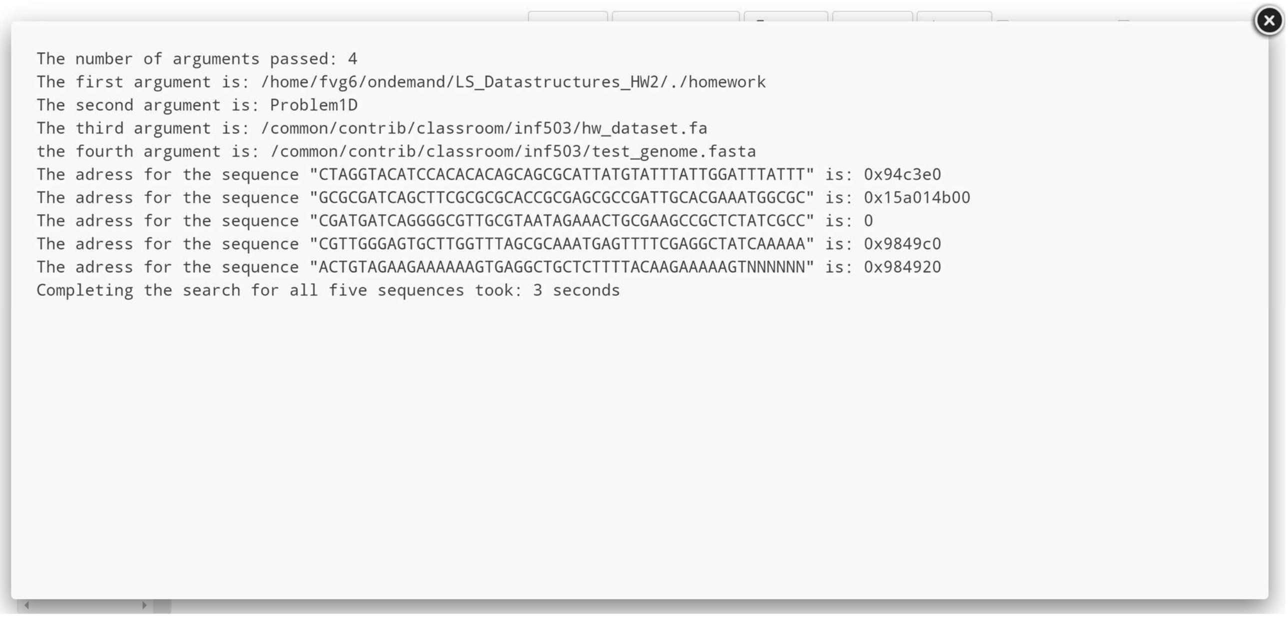
As can be seen, it took 5 seconds to copy the entire read set. The copy constructor copies each node from the original into the copy, one node at a time, and thus the big O notation for the copy constructor is O(n).



```
free(): double free detected in tcache 2
srun: error: cn65: task 0: Aborted (core dumped)
```

In solving this problem, I got an error that I'm not familiar with. It doesn't seem like it affected the program however.

D)

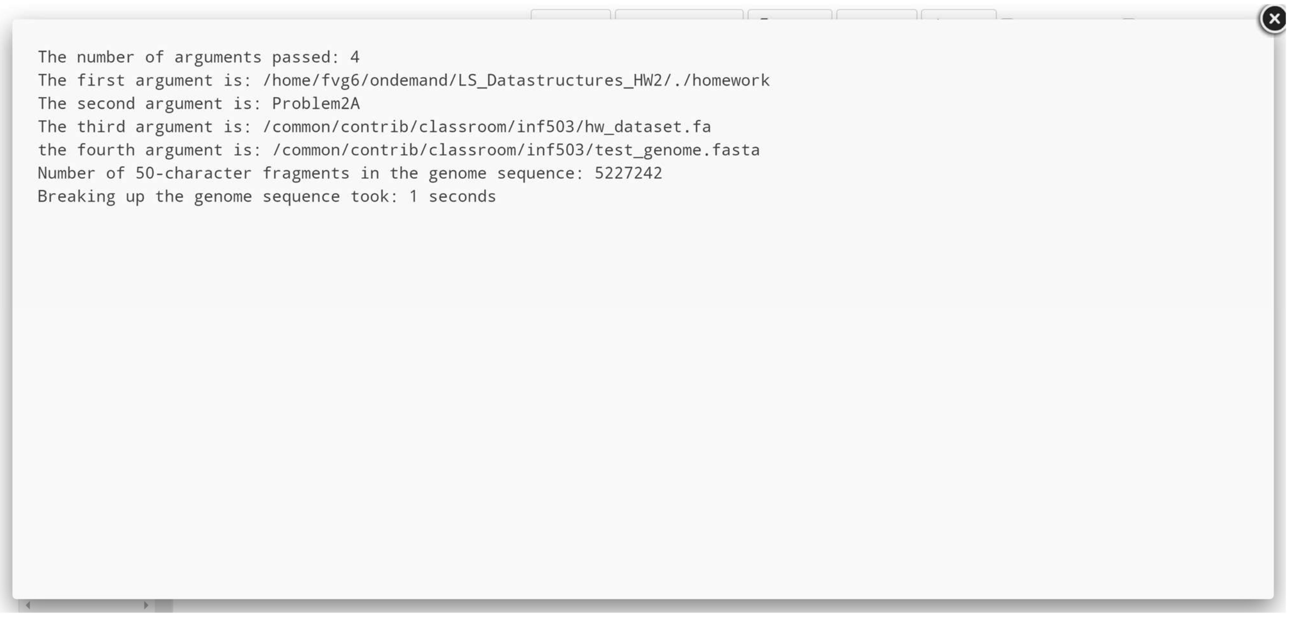


```
The number of arguments passed: 4
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW2/./homework
The second argument is: Problem1D
The third argument is: /common/contrib/classroom/inf503/hw_dataset.fa
the fourth argument is: /common/contrib/classroom/inf503/test_genome.fasta
The adress for the sequence "CTAGGTACATCCACACACAGCAGCGCATTATGTATTTATTGGATTTATTT" is: 0x94c3e0
The adress for the sequence "GCGCGATCAGCTTCGCGCGCACCGCGAGCGCCGATTGCACGAAATGGCGC" is: 0x15a014b00
The adress for the sequence "CGATGATCAGGGGCGTTGCGTAATAGAAACTGCGAAGCCGCTCTATCGCC" is: 0
The adress for the sequence "CGTTGGGAGTGCTTGGTTTAGCGCAAATGAGTTTTCGAGGCTATCAAAAA" is: 0x9849c0
The adress for the sequence "ACTGTAGAAGAAAAAAGTGAGGCTGCTCTTTTACAAGAAAAAGTNNNNNN" is: 0x984920
Completing the search for all five sequences took: 3 seconds
```

As can be seen, the search function was implemented correctly. The only sequence not found in the read set is the third one, according to the program.

Problem 2

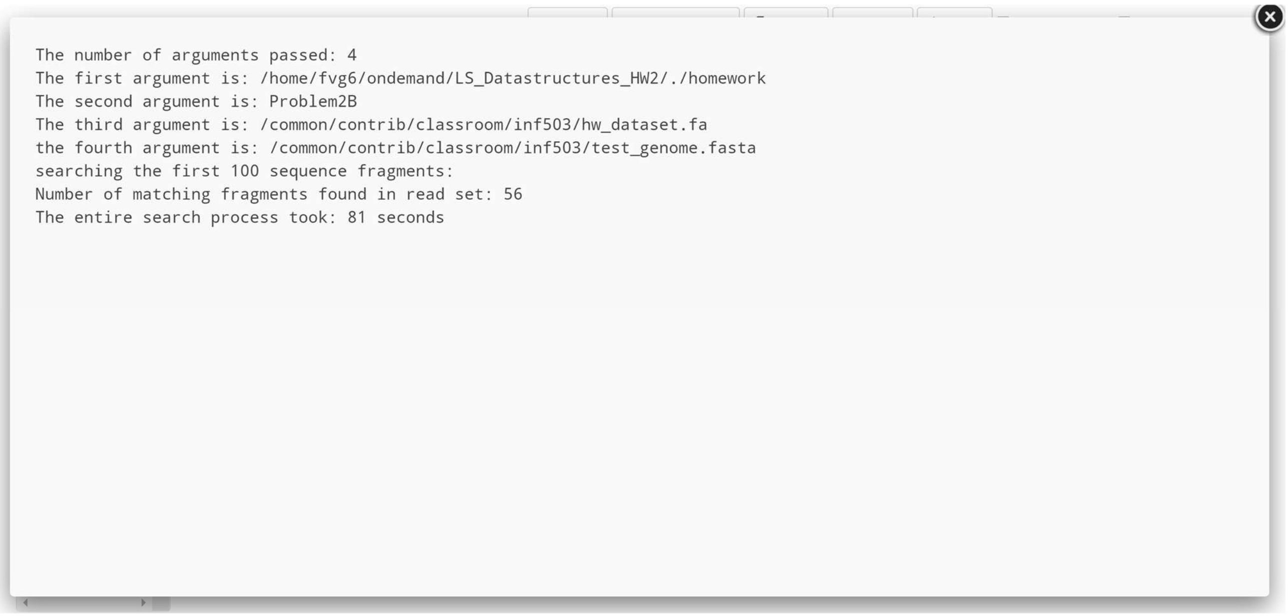
A)



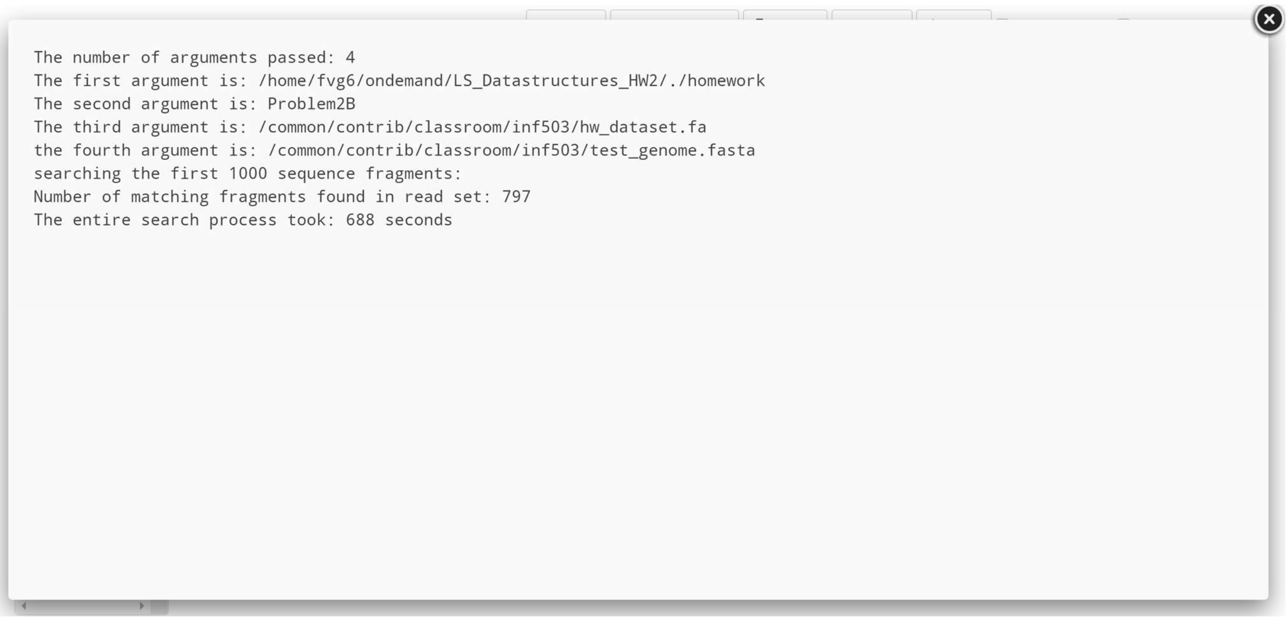
```
The number of arguments passed: 4
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW2/./homework
The second argument is: Problem2A
The third argument is: /common/contrib/classroom/inf503/hw_dataset.fa
the fourth argument is: /common/contrib/classroom/inf503/test_genome.fasta
Number of 50-character fragments in the genome sequence: 5227242
Breaking up the genome sequence took: 1 seconds
```

As can be seen, according to the program, the genome can be split up into 5,227,242 sequence fragments. My program however reads the last character in the sequence twice, and thus it appears that there's an extra character, and thus an extra sequence in the genome. The actual number of sequence fragments that the genome can be split up into is thus 5,227,241. Additionally, one extra sequence fragment shouldn't affect the search process in the next problem too much, especially when the entire genome can't be searched within a reasonable time for this assignment.

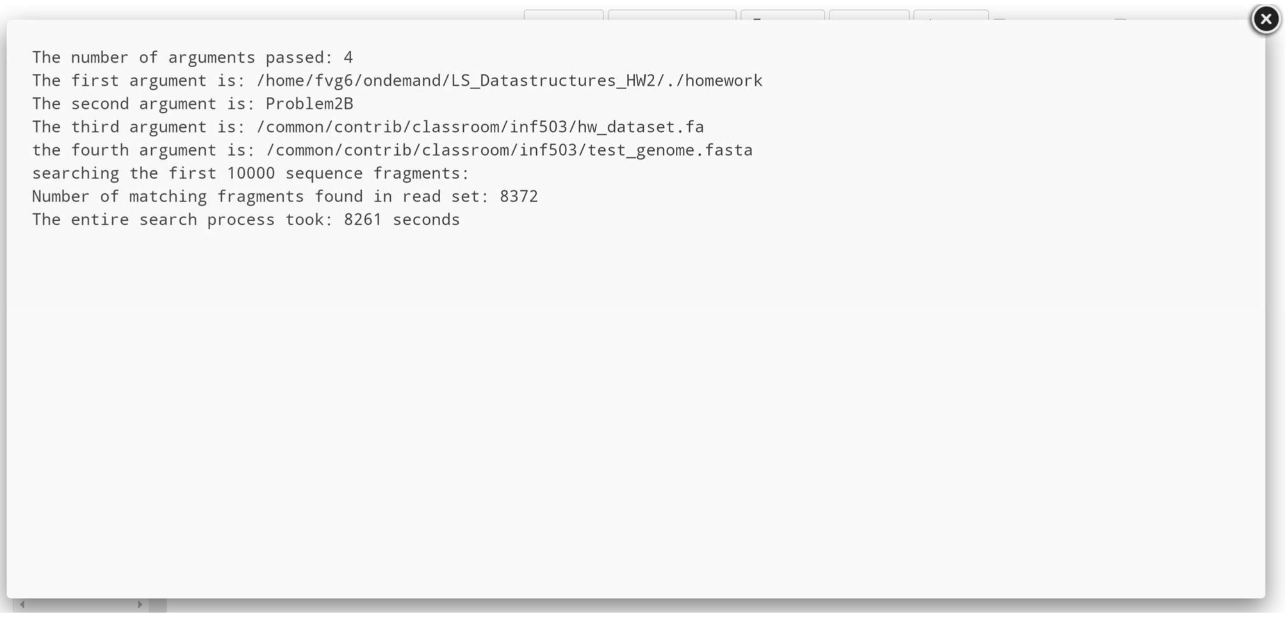
B)



```
The number of arguments passed: 4
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW2/./homework
The second argument is: Problem2B
The third argument is: /common/contrib/classroom/inf503/hw_dataset.fa
the fourth argument is: /common/contrib/classroom/inf503/test_genome.fasta
searching the first 100 sequence fragments:
Number of matching fragments found in read set: 56
The entire search process took: 81 seconds
```



```
The number of arguments passed: 4
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW2/./homework
The second argument is: Problem2B
The third argument is: /common/contrib/classroom/inf503/hw_dataset.fa
the fourth argument is: /common/contrib/classroom/inf503/test_genome.fasta
searching the first 1000 sequence fragments:
Number of matching fragments found in read set: 797
The entire search process took: 688 seconds
```



```
The number of arguments passed: 4
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW2/./homework
The second argument is: Problem2B
The third argument is: /common/contrib/classroom/inf503/hw_dataset.fa
the fourth argument is: /common/contrib/classroom/inf503/test_genome.fasta
searching the first 10000 sequence fragments:
Number of matching fragments found in read set: 8372
The entire search process took: 8261 seconds
```

I ran the program to search for the first 100, 1,000, and 10,000 sequence fragments in the read set, and I got the following results:

Out of the first 100 sequence fragments, 56 were found in the read set. It took a total of 81 seconds to search the entire read set for these 100 fragments.

Out of the first 1,000 sequence fragments, 797 were found in the read set. It took a total of 688 seconds (11 minutes and 28 seconds) to search the entire read set for these 1,000 fragments.

Out of the first 10,000 sequence fragments, 8,372 were found in the read set. It took a total of 8,261 seconds (roughly 2 hours and 18 minutes) to search the entire read set for these 10,000 fragments.

All 5,227,241 sequence fragments could not be searched for in a reasonable time. Judging from the runtime of the program when searching for the first 100, 1,000, and 10,000 fragments, the runtime seems to increase in a roughly linear fashion for larger numbers of fragments (~8 seconds for every 10 sequence fragments). Thus, searching for all 5,227,241 fragments would take around 4,000,000 seconds, or around 6-7 weeks.