

INF503_Assignment_2_mp2525 - Solutions

Problem #1 (of 2): Fun with linked lists

A. Read in the entire 36 million read set and report RAM and CPU time used to load the data into memory.

Ans:

```
[mp2525@wind ~/large-scale-data-structures/homework2]$ cat /scratch/mp2525/linkedlist_q1a.txt
g++ -o homework homework.o fasta.o
The number of arguments passed: 4
The first argument is: /home/mp2525/large-scale-data-structures/homework2/./homework
The second argument is: problem1A
The third argument is: /common/contrib/classroom/inf503/hw_dataset.fa
The fourth argument is: -1
////////////////////////////////////
Initialized 36220411 genome sequences.
#####
Time to read full dataset: 8.017 seconds.
#####
Deallocating all linkedlists..
Time to Deallocate memory: 0.437 seconds.
#####
////////////////////////////////////
rm -rf *.o homework
```

```
[mp2525@wind ~/large-scale-data-structures/homework2]$ sbatch q1a.sh
Submitted batch job 37558180
[mp2525@wind ~/large-scale-data-structures/homework2]$ jobstats -j 37558180
```

JobID	JobName	ReqMem	MaxRSS	ReqCPUS	UserCPU	Timelimit	Elapsed	State	JobEff
37558180	q1a_linkedli+	9.77G	2.70G	1	00:07.457	00:10:00	00:01:15	COMPLETED	20.07

```
=====
Memory      : 27.63%
CPU         : -
GPU         : -
Time Limit  : 12.50%
=====
Efficiency Score: 20.06
=====
```

From the first screenshot, we can see that the initialization of the entire 36 million genome sequences in the linked list took **8.017 seconds**.

Similarly, from the second screenshot, we can see that the memory used for the execution of the job is **2.70GB**.

B. Implement a destructor for your class to delete/deallocate your array data structure. How long should it take (big O notation)? Explain why (a couple of sentences at most).

Ans:

A terminal window showing the execution of a C++ program. The background is dark with a faint skull watermark. The text is as follows:

```
[mp2525@wind ~/large-scale-data-structures/homework2]$ cat /scratch/mp2525/linkedlist_q1b.txt
g++ -c homework.cpp -O3
g++ -c fasta.cpp -O3
g++ -o homework homework.o fasta.o
The number of arguments passed: 4
The first argument is: /home/mp2525/large-scale-data-structures/homework2/./homework
The second argument is: problem1B
The third argument is: /common/contrib/classroom/inf503/hw_dataset.fa
The fourth argument is: -1
////////////////////////////////////
Initialized 36220411 genome sequences.
#####
Time to read full dataset: 8.498 seconds.
#####
Deallocating all linkedlists..
Time to Deallocate memory: 0.598 seconds.
#####
////////////////////////////////////
rm -rf *.o homework
```

From the first screenshot, we can see that the deallocation of the entire 36 million genome sequences in the linked list took **0.598 seconds**. The big O notation for the deallocation of the entire dataset is **O(N)**. It is because we iterate until the linked list ends with NULL and the length of the linked list is the entire dataset count (N).

C. Implement a copy constructor and perform a deep copy of the entire FASTAreadset_LL object. How long should it take (big O notation)? Explain why.

Ans:

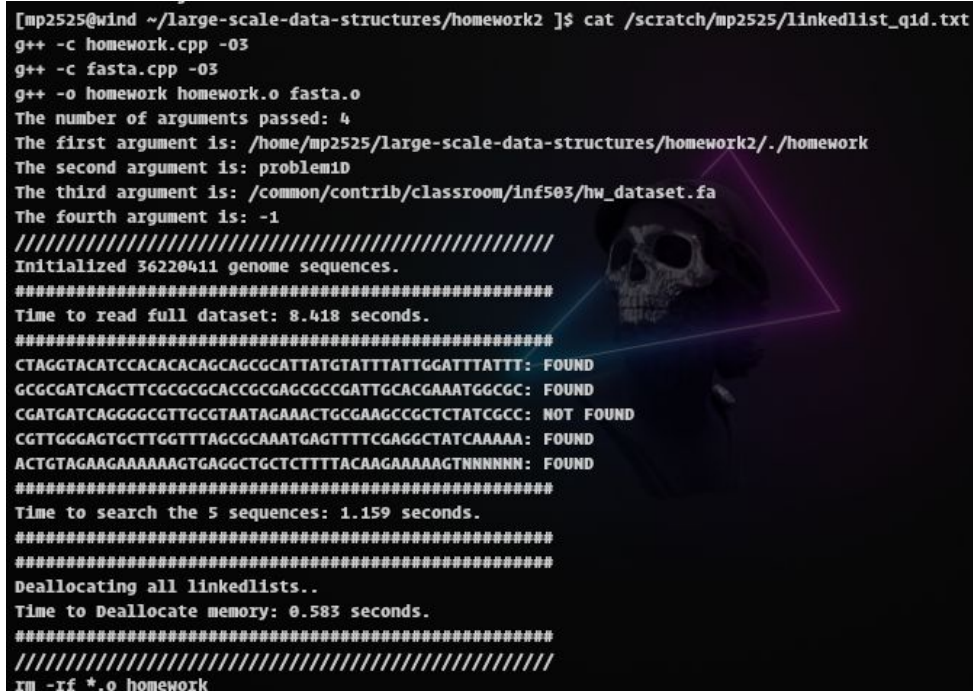
```
[mp2525@wind ~/large-scale-data-structures/homework2]$ cat /scratch/mp2525/LinkedList_q1c.txt
g++ -o homework homework.o fasta.o
The number of arguments passed: 4
The first argument is: /home/mp2525/large-scale-data-structures/homework2/./homework
The second argument is: problematic
The third argument is: /common/contrib/classroom/inf503/hw_dataset.fa
The fourth argument is: -1
////////////////////////////////////
Initialized 36220411 genome sequences.
#####
Time to read full dataset: 8.313 seconds.
#####
Time to perform a deep copy of entire dataset: 2.356 seconds.
#####
Deallocating all linkedlists..
Time to Deallocate memory: 0.502 seconds.
#####
Deallocating all linkedlists..
Time to Deallocate memory: 0.483 seconds.
#####
////////////////////////////////////
rm -rf *.o homework
```

From the first screenshot, we can see that the deep copy of the entire dataset from an object to the new object took **2.356 seconds**. The big O notation to perform the deep copy of the entire dataset is $O(N \times 50) \sim O(N)$ where characters to insert are of fixed size (50). The deep copy of the entire dataset has less execution time than the initialization of the dataset from the file. It is because we do not have to perform any file read and write operations. Moreover, the data are moved from one object to another from the memory locations.

D. Implement a search function which would take a sequence fragment (OK to assume that it will be exactly 50 characters long) and search for this fragment within the FASTAreadset_LL object. The search function should return the pointer to the node containing a match OR the NULL pointer value if a 'hit' was not found. Which of the following sequences were found in the read set:

- CTAGGTACATCCACACACAGCAGCGCATTATGTATTTATTGGATTATT
- GCGCGATCAGCTTCGCGCGCACCGCGAGCGCCGATTGCACGAAATGGCGC
- CGATGATCAGGGGCGTTGCGTAATAGAACTGCGAAGCCGCTCTATCGCC
- CGTTGGGAGTGCTTGCTTTAGCGCAAATGAGTTTTCGAGGCTATCAAAAA
- ACTGTAGAAGAAAAAAGTGAGGCTGCTCTTTTACAAGAAAAAGTNNNNNN

Ans:



```
[mp2525@wind ~/large-scale-data-structures/homework2]$ cat /scratch/mp2525/linkedlist_q1d.txt
g++ -c homework.cpp -O3
g++ -c fasta.cpp -O3
g++ -o homework homework.o fasta.o
The number of arguments passed: 4
The first argument is: /home/mp2525/large-scale-data-structures/homework2/./homework
The second argument is: problem1D
The third argument is: /common/contrib/classroom/inf503/hw_dataset.fa
The fourth argument is: -1
////////////////////////////////////
Initialized 36220411 genome sequences.
#####
Time to read full dataset: 8.418 seconds.
#####
CTAGGTACATCCACACACAGCAGCGCATTATGTATTTATTGGATTATT: FOUND
GCGCGATCAGCTTCGCGCGCACCGCGAGCGCCGATTGCACGAAATGGCGC: FOUND
CGATGATCAGGGGCGTTGCGTAATAGAACTGCGAAGCCGCTCTATCGCC: NOT FOUND
CGTTGGGAGTGCTTGCTTTAGCGCAAATGAGTTTTCGAGGCTATCAAAAA: FOUND
ACTGTAGAAGAAAAAAGTGAGGCTGCTCTTTTACAAGAAAAAGTNNNNNN: FOUND
#####
Time to search the 5 sequences: 1.159 seconds.
#####
Deallocating all linkedlists..
Time to Deallocate memory: 0.583 seconds.
#####
////////////////////////////////////
rm -rf *.o homework
```


From the first screenshot, we can see that first, second, fourth, and fifth genome sequences are found in the entire 36 million genome sequences. Node pointer is returned when the sequence is found and similarly, NULL is returned when the sequence is not found. It took **1.159 seconds** to perform the search of 5 sequences in the entire dataset. The time complexity of the algorithm is $O(N)$ in the worst case.

Problem #2 (of 2): Basic search

A. Break down the genome sequence into all 50-character long fragments contained within (shifting start location by one base each time). Store these fragments in an array or linked list data structure (your choice). How many 50-character fragments did you observe in the Genome?

Ans:

```
[mp2525@wind ~/large-scale-data-structures/homework2]$ cat /scratch/mp2525/linkedlist_q2a.txt
g++ -c homework.cpp -O3
g++ -c fasta.cpp -O3
g++ -o homework homework.o fasta.o
The number of arguments passed: 5
The first argument is: /home/mp2525/large-scale-data-structures/homework2/./homework
The second argument is: problem2A
The third argument is: /common/contrib/classroom/inf503/test_genome.fasta
The fourth argument is: -1
The fifth argument is: -1
////////////////////////////////////
#####
Initialized 5227244 50-mers data.
Time to read genome dataset: 0.381 seconds.
#####
#####
Deallocating all linkedlists..
Time to Deallocate memory: 0.064 seconds.
#####
////////////////////////////////////
rm -rf *.o homework
```



From the first screenshot, we can see that **5227244** 50-character long fragments were found in the entire dataset contained within (shifting start location by one base each time). The time complexity for the initialization is $O(N*50) \sim O(N)$ where characters to insert are of fixed size(50). Also, It took **0.381 seconds** to initialize the genome sequences. Similarly, It took **0.064 seconds** to deallocate the linked list of genome sequences.

B. Iterate through all 50-mers found in the genome, using the search function you developed in 1D to query the read set. How many genome 50-mer fragments were found in your read set? How long does it take to complete the entire search process (all 50-mers)?

Ans:

Solution 1: Naive approach

For 1000 genome sequences search:

```
[mp2525@wind ~/large-scale-data-structures/homework2] $ cat /scratch/mp2525/LinkedList_q2b.txt
g++ -c homework.cpp -O3
g++ -c fasta.cpp -O3
g++ -o homework homework.o fasta.o
The number of arguments passed: 5
The first argument is: /home/mp2525/large-scale-data-structures/homework2/./homework
The second argument is: problem2B
The third argument is: /common/contrib/classroom/inf503/test_genome.fasta
The fourth argument is: -1
The fifth argument is: 1000
////////////////////////////////////
Initialized 36220411 genome sequences.
#####
Time to read full dataset: 8.316 seconds.
#####
#####
Initialized 1000 50-mers data.
Time to read genome dataset: 0.000 seconds.
#####
797 50 mers genome sequences match were found.
#####
Time search all 50-mers genome sequences: 206.995 seconds.
#####
#####
Deallocating all linkedlists..
Time to Deallocate memory: 0.443 seconds.
#####
////////////////////////////////////
rm -rf *.o homework
```

797 out of 10000 50-characters long genome sequences are found in **206.995 seconds**

For 10000 genome sequences search:

```
[mp2525@wind ~/large-scale-data-structures/homework2]$ cat /scratch/mp2525/linkedlist_q2b.txt
g++ -c homework.cpp -O3
g++ -c fasta.cpp -O3
g++ -o homework homework.o fasta.o
The number of arguments passed: 5
The first argument is: /home/mp2525/large-scale-data-structures/homework2/./homework
The second argument is: problem2B
The third argument is: /common/contrib/classroom/inf503/test_genome.fasta
The fourth argument is: -1
The fifth argument is: 10000
//////////
Initialized 36220411 genome sequences.
#####
Time to read full dataset: 8.916 seconds.
#####
#####
Initialized 10000 50-mers data.
Time to read genome dataset: 0.001 seconds.
#####
8372 50 mers genome sequences match were found.
#####
Time search all 50-mers genome sequences: 2581.182 seconds.
#####
#####
Deallocating all linkedlists..
Time to Deallocate memory: 0.550 seconds.
#####
//////////
rm -rf *.o homework
```

8372 out of 10000 50-characters long genome sequences are found in **2581.182 seconds**

For 100000 genome sequences search

```
[mp2525@wind ~]$ cat /scratch/mp2525/linkedlist_q2b.txt
g++ -c homework.cpp -O3
g++ -c fasta.cpp -O3
g++ -o homework homework.o fasta.o
The number of arguments passed: 5
The first argument is: /home/mp2525/large-scale-data-structures/homework2/./homework
The second argument is: problem2B
The third argument is: /common/contrib/classroom/inf503/test_genome.fasta
The fourth argument is: -1
The fifth argument is: 100000
//////////
Initialized 36220411 genome sequences.
#####
Time to read full dataset: 7.979 seconds.
#####
#####
Initialized 100000 50-character long genome sequences.
Time to read genome dataset: 0.007 seconds.
#####
83188 50 mers genome sequences match were found.
#####
Time search all 50-character long genome sequences: 22877.859 seconds.
#####
#####
Deallocating all linkedlists..
Time to Deallocate memory: 0.479 seconds.
#####
//////////
```

83188 out of 100000 50-characters long genome sequences are found in **22877.859 seconds**.

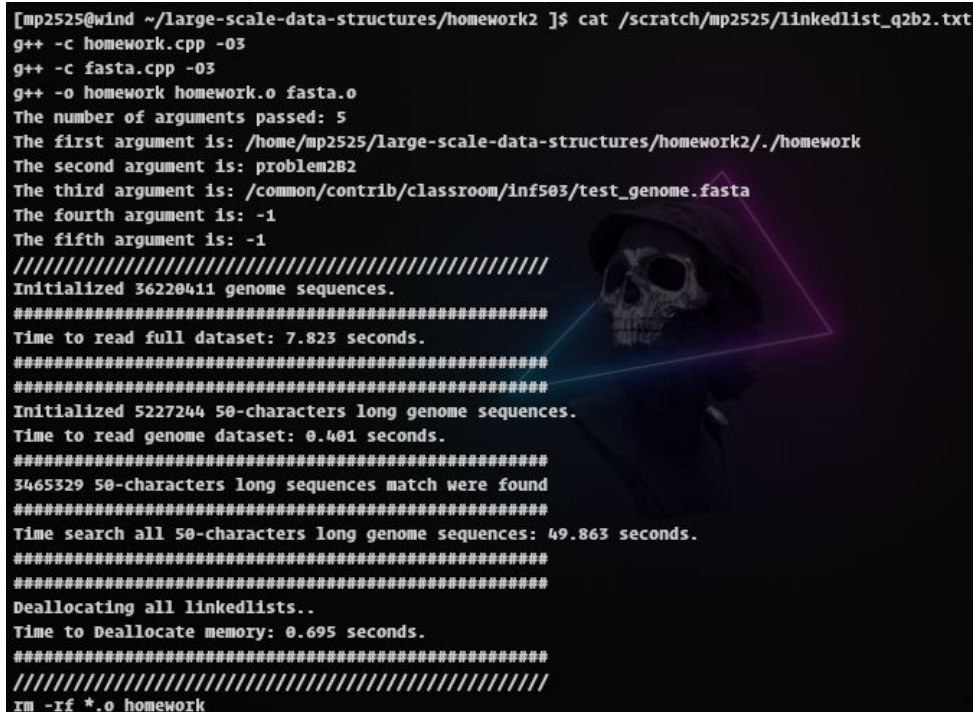
From the observation, we can see that the time complexity of the search is approximately $O(M * N)$ where M is 36 million datasets and N is 50-characters long genome sequences. So, the search of the entire 5227244 sequences will take

$(206.995 / (36220411 * 1000)) * (36220411 * 5227244)$
= 1082013.37178 seconds
= ~ **300.56 hours approximately**

Solution 2: Binary Search in sorted array data structure

In this solution, I have first stored the 36 million datasets into an array data structure and sort them using a quick sort algorithm. Then, each of the 50-character long genome sequences is searched in the sorted array using the binary search.

For 5227244 genome sequences search:



```
[mp2525@wind ~/large-scale-data-structures/homework2] $ cat /scratch/mp2525/linkedlist_q2b2.txt
g++ -c homework.cpp -O3
g++ -c fasta.cpp -O3
g++ -o homework homework.o fasta.o
The number of arguments passed: 5
The first argument is: /home/mp2525/large-scale-data-structures/homework2/./homework
The second argument is: problem2B2
The third argument is: /common/contrib/classroom/inf503/test_genome.fasta
The fourth argument is: -1
The fifth argument is: -1
////////////////////////////////////
Initialized 36220411 genome sequences.
#####
Time to read full dataset: 7.823 seconds.
#####
Initialized 5227244 50-characters long genome sequences.
Time to read genome dataset: 0.401 seconds.
#####
3465329 50-characters long sequences match were found
#####
Time search all 50-characters long genome sequences: 49.863 seconds.
#####
Deallocating all linkedlists..
Time to Deallocate memory: 0.695 seconds.
#####
////////////////////////////////////
rm -rf *.o homework
```

From the screenshot, we can see that **3465329** out of **5227244** of the 50-characters long genome sequences are found in the 36 million datasets. It took about **49.863 seconds** to perform the array sort and binary search.

The reduction of the execution time compared to solution 1 is because of the fact that we are searching sequences in the sorted array data structure using the binary search whose time complexity is $O(\log N)$.

Similarly,

For 1000 genome sequences search:

```
[mp2525@wind ~/large-scale-data-structures/homework2]$ cat /scratch/mp2525/LinkedList_q2b2.txt
g++ -o homework homework.o fasta.o
The number of arguments passed: 5
The first argument is: /home/mp2525/large-scale-data-structures/homework2/./homework
The second argument is: problem2B2
The third argument is: /common/contrib/classroom/inf503/test_genome.fasta
The fourth argument is: -1
The fifth argument is: 1000
////////////////////////////////////
Initialized 36220411 genome sequences.
#####
Time to read full dataset: 7.894 seconds.
#####
Initialized 1000 50-characters long genome sequences.
Time to read genome dataset: 0.000 seconds.
#####
797 50-characters long sequences match were found
#####
Time search all 50-characters long genome sequences: 46.095 seconds.
#####
Deallocating all linkedlists..
Time to Deallocate memory: 0.620 seconds.
#####
////////////////////////////////////
rm -rf *.o homework
```

797 out of 1000 50-characters long genome sequences are found.

For 10000 genome sequences search:

```
[mp2525@wind ~/large-scale-data-structures/homework2]$ cat /scratch/mp2525/LinkedList_q2b2.txt
g++ -c homework.cpp -O3
g++ -c fasta.cpp -O3
g++ -o homework homework.o fasta.o
The number of arguments passed: 5
The first argument is: /home/mp2525/large-scale-data-structures/homework2/./homework
The second argument is: problem2B2
The third argument is: /common/contrib/classroom/inf503/test_genome.fasta
The fourth argument is: -1
The fifth argument is: 10000
////////////////////////////////////
Initialized 36220411 genome sequences.
#####
Time to read full dataset: 7.847 seconds.
#####
Initialized 10000 50-character long genome sequences.
Time to read genome dataset: 0.001 seconds.
#####
8372 50 characters long sequences match were found
#####
Time search all 50-character long genome sequences: 35.402 seconds.
#####
Deallocating all linkedlists..
Time to Deallocate memory: 0.575 seconds.
#####
////////////////////////////////////
rm -rf *.o homework
```

8372 out of 10000 50-characters long genome sequences are found.

For 100000 genome sequences search:

```
[mp2525@wind ~/large-scale-data-structures/homework2] $ cat /scratch/mp2525/LinkedList_q2b2.txt
g++ -c homework.cpp -O3
g++ -c fasta.cpp -O3
g++ -o homework homework.o fasta.o
The number of arguments passed: 5
The first argument is: /home/mp2525/large-scale-data-structures/homework2/./homework
The second argument is: problem2B2
The third argument is: /common/contrib/classroom/Inf503/test_genome.fasta
The fourth argument is: -1
The fifth argument is: 100000
////////////////////////////////////
Initialized 36220411 genome sequences.
#####
Time to read full dataset: 8.260 seconds.
#####
Initialized 100000 50-characters long genome sequences.
Time to read genome dataset: 0.007 seconds.
#####
83188 50-characters long sequences match were found
#####
Time search all 50-characters long genome sequences: 29.218 seconds.
#####
Deallocating all linkedlists..
Time to Deallocate memory: 0.442 seconds.
#####
////////////////////////////////////
rm -rf *.o homework
```

83188 out of 10000 50-characters long genome sequences are found.

The results found in solution 2 are the same as solution 1.