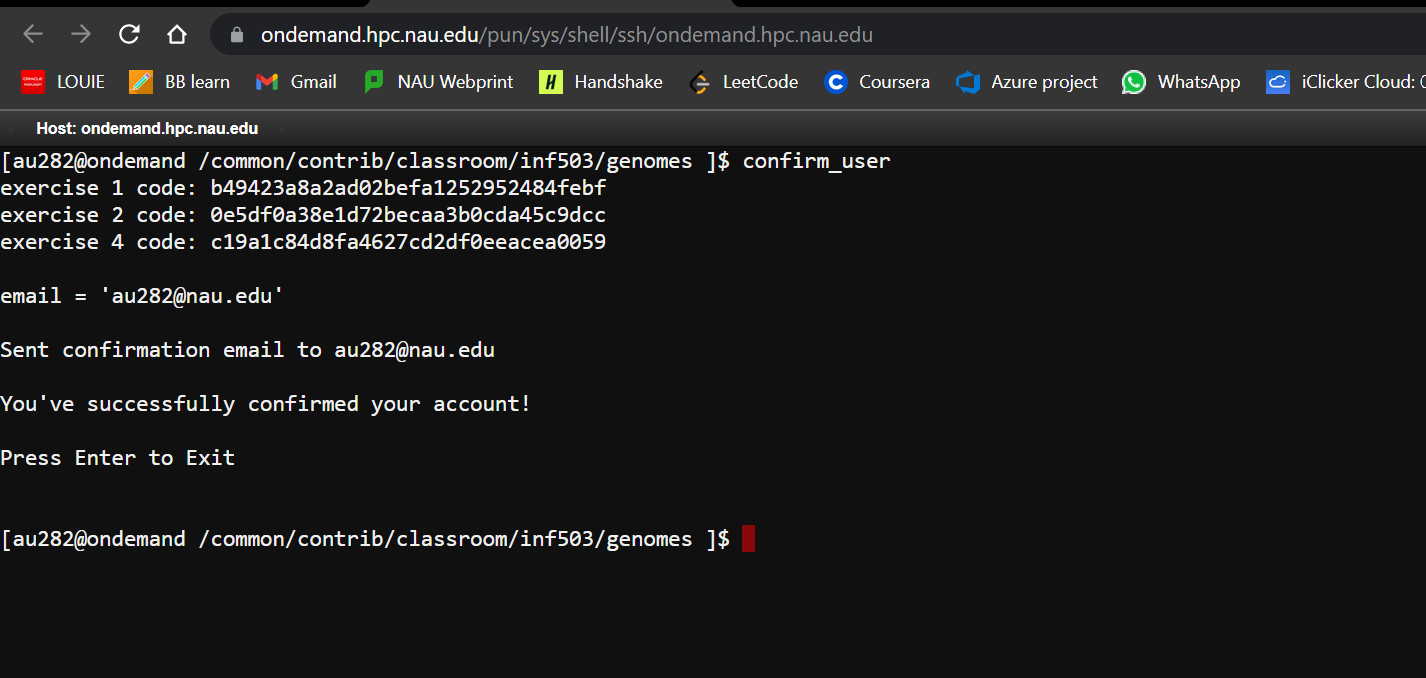
**Large scale Data Structures**

**Homework – 1**

**Problem 1**



**Problem 2**

1. **srun homework\_1 human.txt 1**

human.txt is a genome file. ‘1’ here denotes the flag to execute the read\_genome() function. “homework\_1” is the executable file.

• How many scaffolds were there?

There were 607 scaffolds

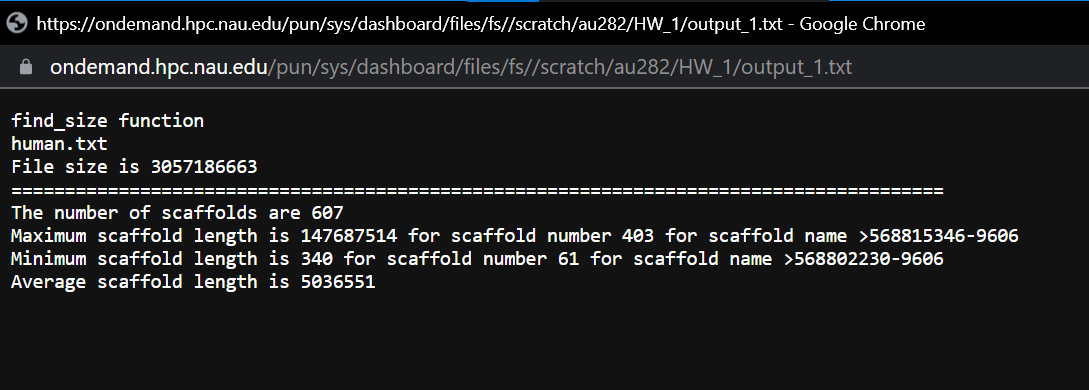
• What was the longest and shortest scaffold? Provide names of scaffolds and lengths.

Maximum scaffold length is 147687514 for scaffold number 403 for scaffold name >568815346-9606

Minimum scaffold length is 340 for scaffold number 61 for scaffold name >568802230-9606

• What was the average scaffold length?

Average scaffold length is 5036551



**b) srun homework\_1 human.txt 2**

human.txt is a genome file. ‘2’ here denotes the flag which finds the number of A,C,G,T and GC content. “homework\_1” is the executable file.

• What is the ‘big O’ notation of your search (linear / quadratic / cubic / etc)?

The time complexity is O(n) where ‘n’ is the size of human genome.

• How long does it take (in seconds) to execute this function? Hint: You will need to use system time within your code to get accurate time estimates.

Total time taken for program execution is 52.6248 seconds

• What was the GC content of the human genome (percent of C’s and G’s in the genome)?

The number of A's are 897004549

The number of G's are 625451943

The number of C's are 622850383

The number of T's are 899663937

The percentage of C in entire human genome is 20.3733%

The percentage of G in entire human genome is 20.4584%

The percentage of GC in entire human genome is 40.8317%

