**LSDS Assignment 4**

**Command for running Problem A : srun homework\_4 1 human.txt human\_reads\_trimmed.fa**

**Command for running Problem B : srun homework\_4 2 human.txt human\_reads\_trimmed.fa**

1. **Needleman Wunch(Random and Completely Random) :**
   1. For each of your searches (10K, 100K, and 1M), how many ‘hits’ with up to 2 mismatches did you find?

**For Random :**

* + 1. **For 10k : Number of hits are 9864**
    2. **For 100k : Number of hits are 99003**
    3. **For 1M : Number of hits are 990003**

**For Completely Random :**

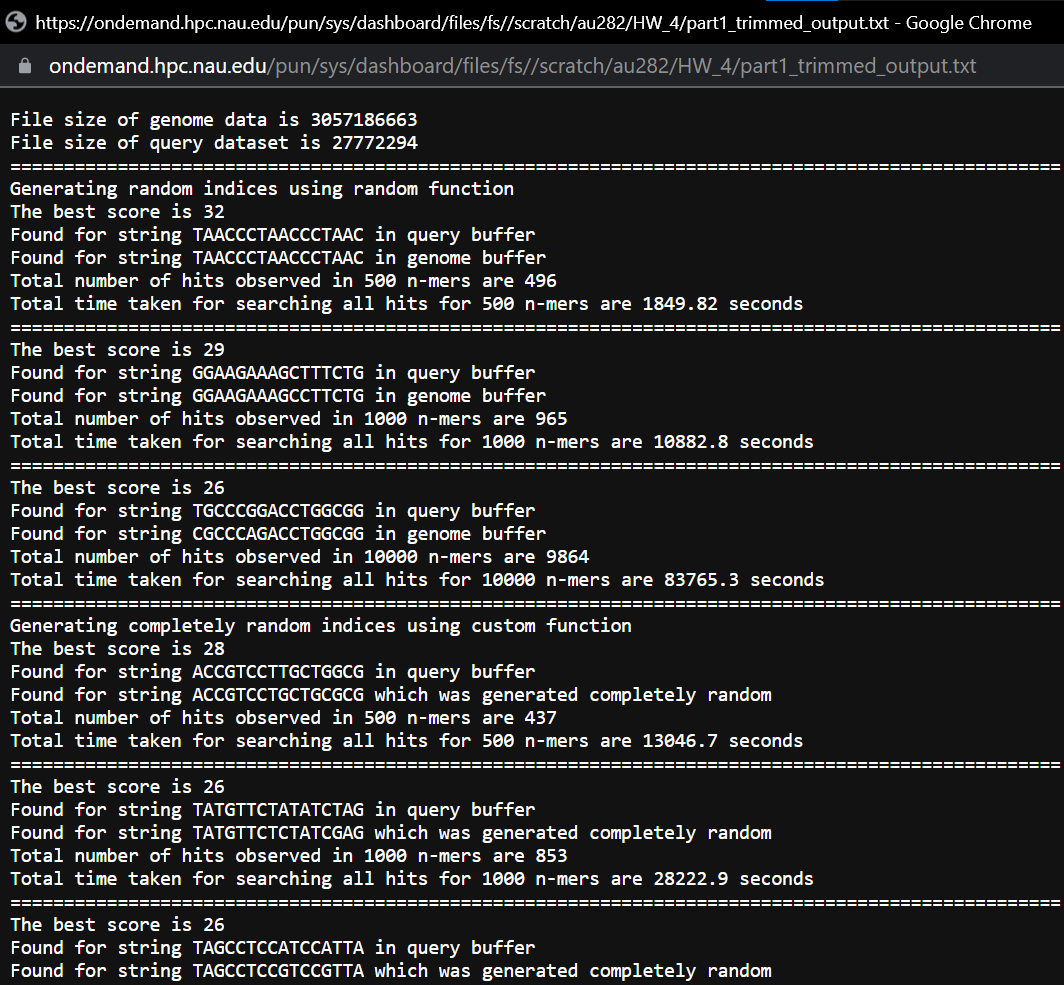
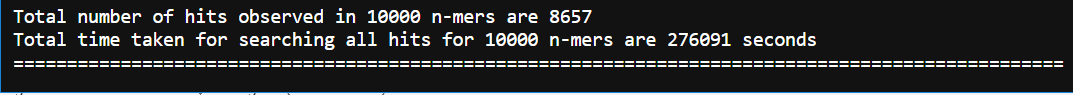
1. **For 10k : Number of hits are 8657**
2. **For 100k : Number of hits are 87005**
3. **For 1M : Number of hits are 870005**
   1. For each of your searches (10K, 100K, and 1M), how long did the search take?

**For Random :**

* + 1. **For 10k : Total time taken is 83765.3 seconds**
    2. **For 100k : Total time taken is 85938.48 seconds**
    3. **For 1M : Total time taken is 8617538.48 seconds**

**For Completely Random :**

1. **For 10k : Total time taken is 276091 seconds**
2. **For 100k : Total time taken is 2768202.26 seconds**
3. **For 1M : Total time taken is 27689202.26 seconds**

* 1. How long would the search take for the entire subject dataset?

**For Random -> human genome(approx. 3 billion):**

**Number of hits are 2659752401**

**Total time taken is 835.646 years(approx. 8 centuries)**

**For Completely Random -> human genome(approx. 3 billion):**

**Number of hits are 2659752401**

**Total time taken is 2684.34 years(approx. 20 centuries)**

1. **BLAST(Random and Completely Random):**
   1. For each of your searches (10K, 100K, and 1M), how many ‘hits’ with up to 2 mismatches did you find?

**For Random :**

* + 1. **For 10k : Number of hits are 21325**
    2. **For 100k : Number of hits are 209867**
    3. **For 1M : Number of hits are 1437699**

**For Completely Random :**

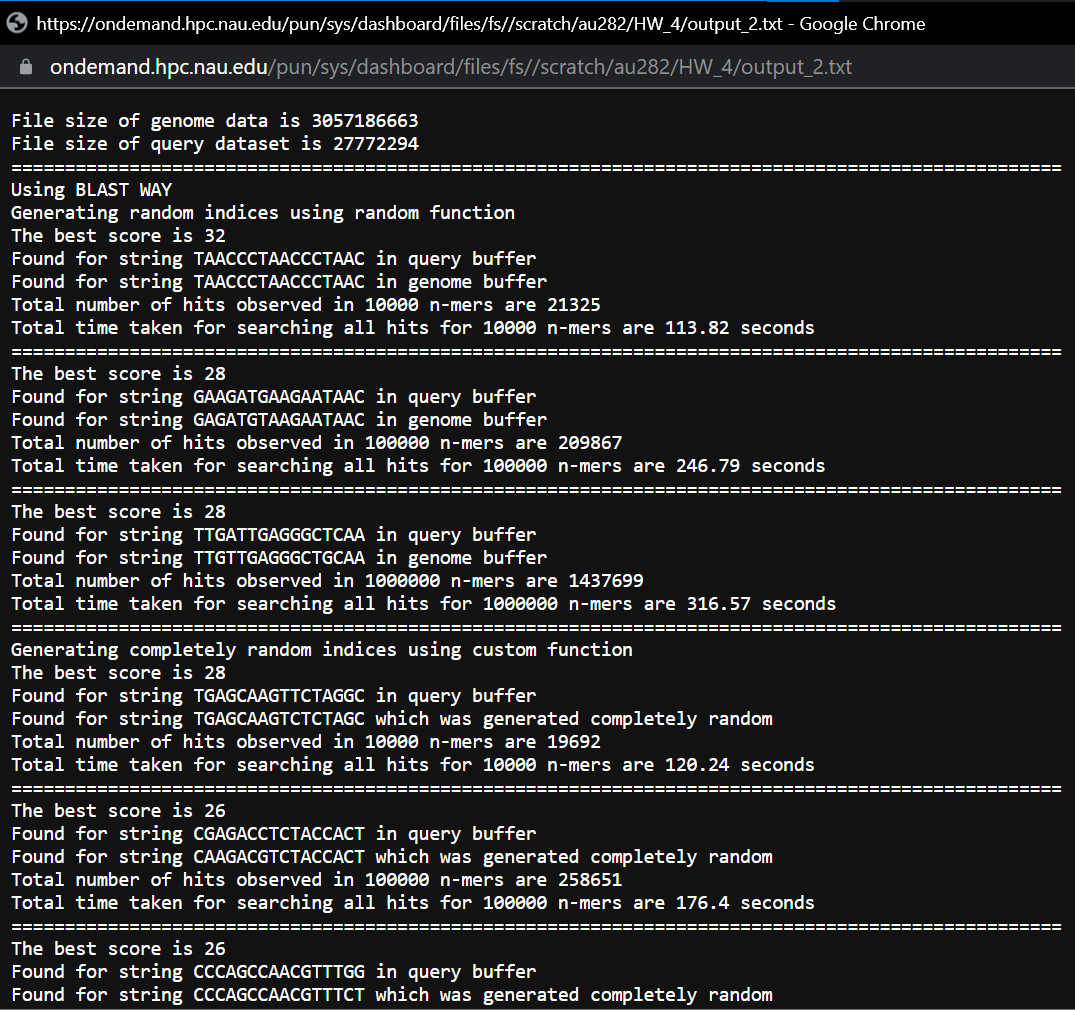
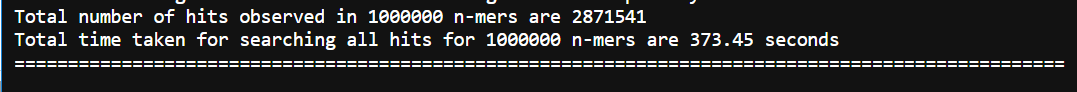
1. **For 10k : Number of hits are 19692**
2. **For 100k : Number of hits are 258651**
3. **For 1M : Number of hits are 2871541**
   1. For each of your searches (10K, 100K, and 1M), how long did the search take?

**For Random :**

* + 1. **For 10k : Total time taken is 113.82 seconds**
    2. **For 100k : Total time taken is 246.79 seconds**
    3. **For 1M : Total time taken is 316.57 seconds**

**For Completely Random :**

1. **For 10k : Total time taken is 120.24 seconds**
2. **For 100k : Total time taken is 176.4 seconds**
3. **For 1M : Total time taken is 373.45 seconds**

* 1. How long would the search take for the entire subject dataset?

**For Random -> human genome(approx. 3 billion):**

**Number of hits are 4371783947**

**Total time taken is 7.24 days**

**For Completely Random -> human genome(approx. 3 billion):**

**Number of hits are 8804688475**

**Total time taken is 9.05 days**

* 1. How does that compare with the benchmarks from problem 1, part B?

**For completely random using NW it took us approximately 20 centuries but when we used BLAST for the same it took approximately 9.05 days which is a drastic change. This is because BLAST is a heuristic approach and we ignore the k-mers which never resulted to a hit sequence but we wasted our time in NW still evaluating for it.**

**Time complexity for NW : O(N) \* O(G) \* O(n \* n)**

**Time complexity for BLAST : O(N) \* O(n) \* O(n)**

**Estimation calculations are mentioned below for both NW and BLAST:**

