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

A) Needleman Wunch(Random and Completely Random) :

a. For each of your searches (10K, 100K, and 1M), how many 'hits' with up to 2 mismatches did you find?

For Random :

a) **For 10k : Number of hits are 9864**
b) **For 100k : Number of hits are 99003**
c) **For 1M : Number of hits are 990003**

For Completely Random :

a) **For 10k : Number of hits are 8657**
b) **For 100k : Number of hits are 87005**
c) **For 1M : Number of hits are 870005**

b. For each of your searches (10K, 100K, and 1M), how long did the search take?

For Random :

a) **For 10k : Total time taken is 83765.3 seconds**
b) **For 100k : Total time taken is 85938.48 seconds**
c) **For 1M : Total time taken is 8617538.48 seconds**

For Completely Random :

d) **For 10k : Total time taken is 276091 seconds**
e) **For 100k : Total time taken is 2768202.26 seconds**
f) **For 1M : Total time taken is 27689202.26 seconds**

c. 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)

B) BLAST(Random and Completely Random):

a. For each of your searches (10K, 100K, and 1M), how many 'hits' with up to 2 mismatches did you find?

For Random :

a) For 10k : Number of hits are 21325
b) For 100k : Number of hits are 209867
c) For 1M : Number of hits are 1437699

For Completely Random :

g) For 10k : Number of hits are 19692
h) For 100k : Number of hits are 258651
i) For 1M : Number of hits are 2871541

b. For each of your searches (10K, 100K, and 1M), how long did the search take?

For Random :

a) For 10k : Total time taken is 113.82 seconds
b) For 100k : Total time taken is 246.79 seconds
c) For 1M : Total time taken is 316.57 seconds

For Completely Random :

j) For 10k : Total time taken is 120.24 seconds
k) For 100k : Total time taken is 176.4 seconds
l) For 1M : Total time taken is 373.45 seconds

https://ondemand.hpc.nau.edu/pun/sys/dashboard/files/fs//scratch/au282/HW_4/output_2.txt - Google Chrome

ondemand.hpc.nau.edu/pun/sys/dashboard/files/fs//scratch/au282/HW_4/output_2.txt

```
File size of genome data is 3057186663
File size of query dataset is 27772294
================================================================================================================
Using BLAST WAY
Generating random indices using random function
The best score is 32
Found for string TAACCCTAACCCTAAC in query buffer
Found for string TAACCCTAACCCTAAC in genome buffer
Total number of hits observed in 10000 n-mers are 21325
Total time taken for searching all hits for 10000 n-mers are 113.82 seconds
================================================================================================================
The best score is 28
Found for string GAAGATGAAGAATAAC in query buffer
Found for string GAGATGTAAGAATAAC in genome buffer
Total number of hits observed in 100000 n-mers are 209867
Total time taken for searching all hits for 100000 n-mers are 246.79 seconds
================================================================================================================
The best score is 28
Found for string TTGATTGAGGGCTCAA in query buffer
Found for string TTGTTGAGGGCTGCAA in genome buffer
Total number of hits observed in 1000000 n-mers are 1437699
Total time taken for searching all hits for 1000000 n-mers are 316.57 seconds
================================================================================================================
Generating completely random indices using custom function
The best score is 28
Found for string TGAGCAAGTTCTAGGC in query buffer
Found for string TGAGCAAGTCTCTAGC which was generated completely random
Total number of hits observed in 10000 n-mers are 19692
Total time taken for searching all hits for 10000 n-mers are 120.24 seconds
================================================================================================================
The best score is 26
Found for string CGAGACCTCTACCACT in query buffer
Found for string CAAGACGTCTACCACT which was generated completely random
Total number of hits observed in 100000 n-mers are 258651
Total time taken for searching all hits for 100000 n-mers are 176.4 seconds
================================================================================================================
The best score is 26
Found for string CCCAGCCAACGTTTGG in query buffer
Found for string CCCAGCCAACGTTTCT which was generated completely random
Total number of hits observed in 1000000 n-mers are 2871541
Total time taken for searching all hits for 1000000 n-mers are 373.45 seconds
================================================================================================================
```

c. 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):

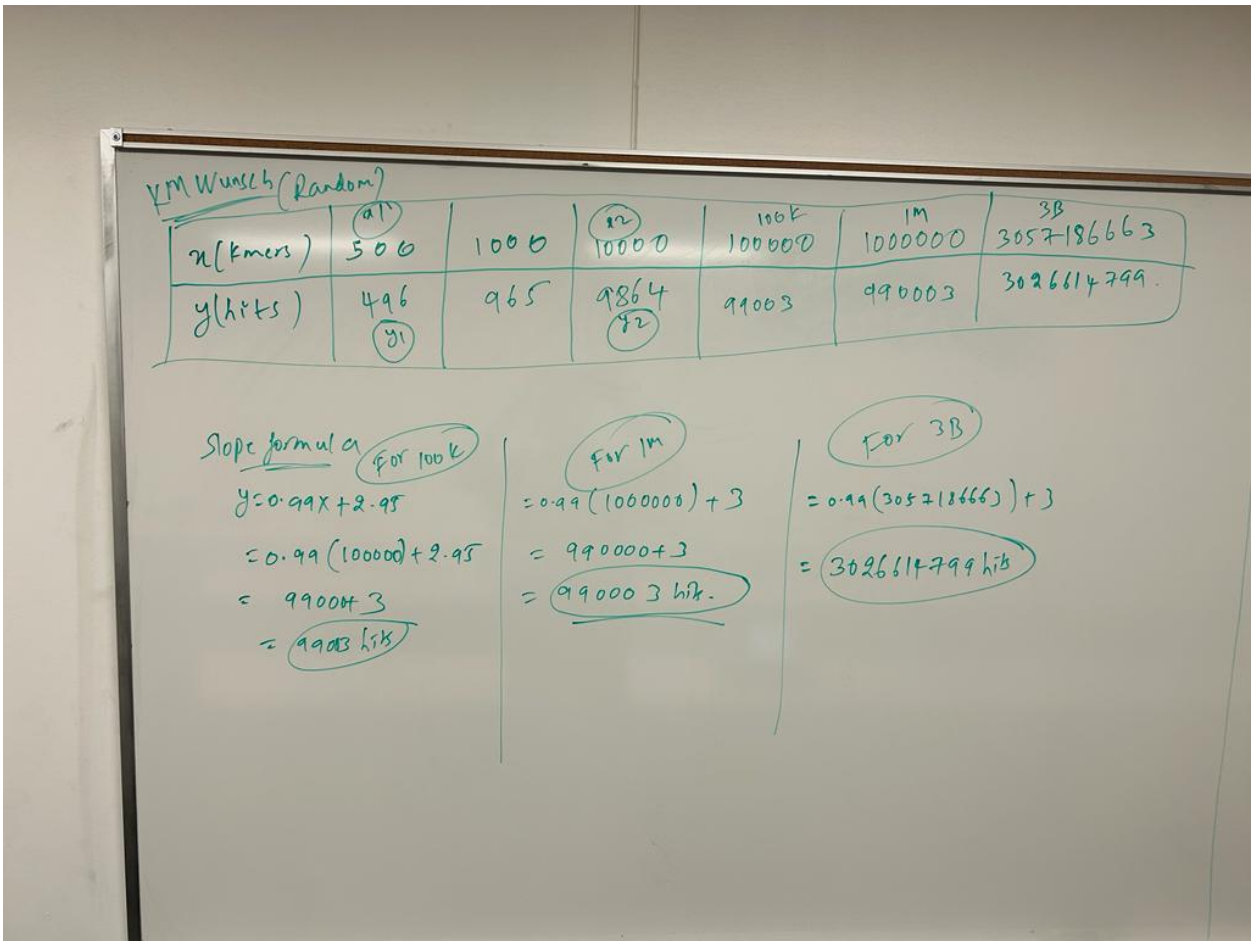
Number of hits are 8804688475
Total time taken is 9.05 days

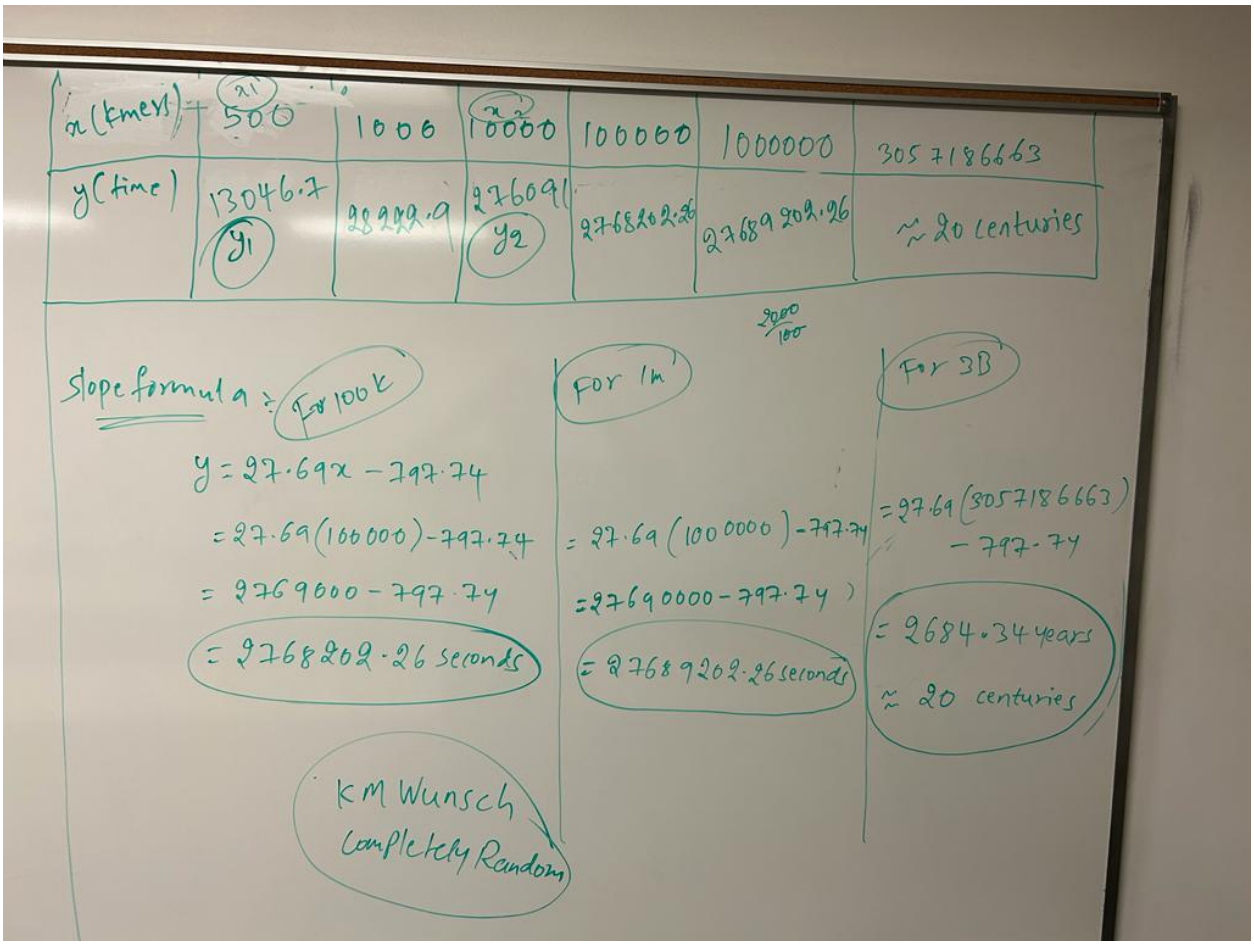
d. How does that compare with the benchmarks from problem 1, part B?

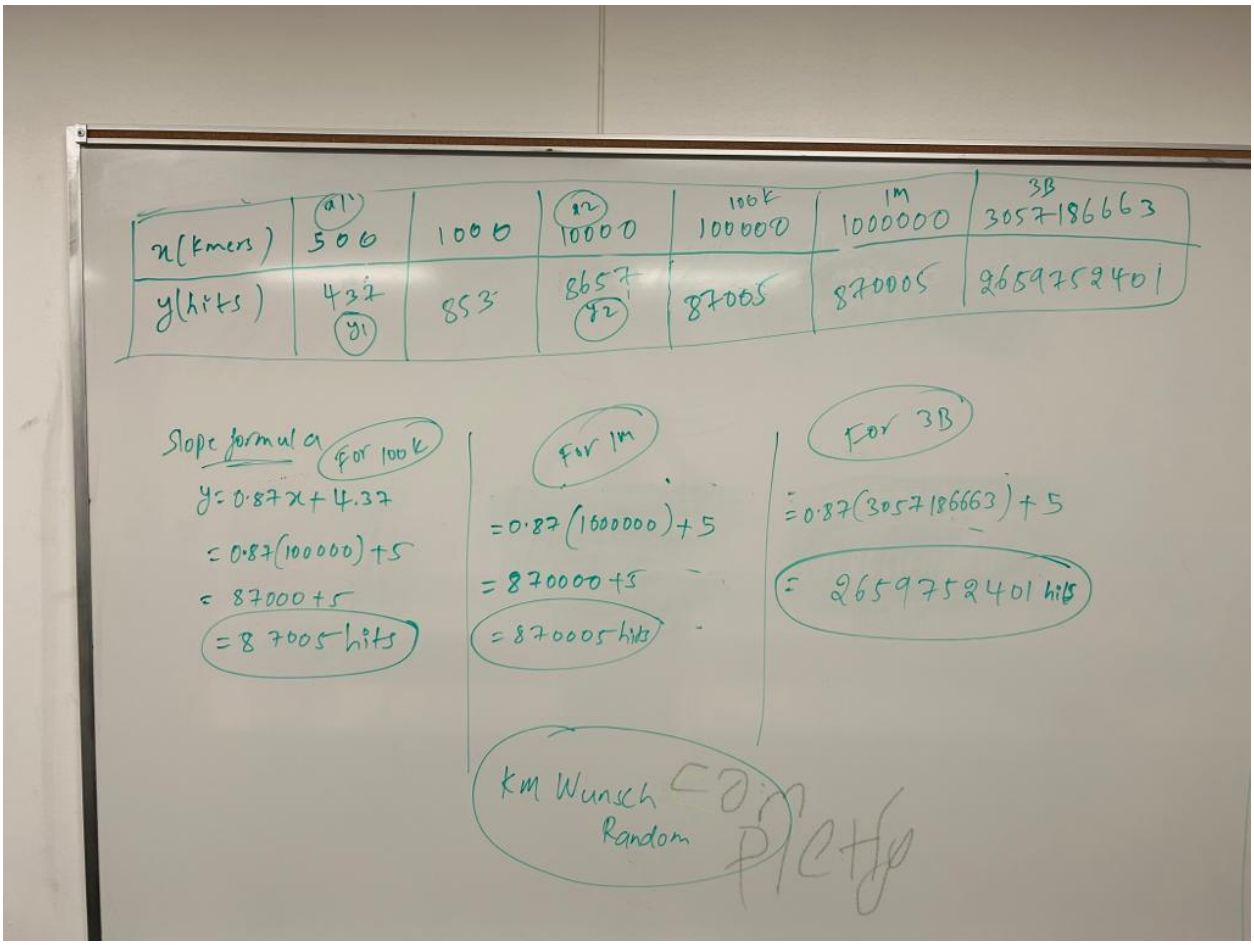
For completely random using NW it took us approximately 20 centuries but when we used BLAST for the same it took approximately 9.05 days which is a drastic change. This is because BLAST is a heuristic approach and we ignore the k-mers which never resulted to a hit sequence but we wasted our time in NW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

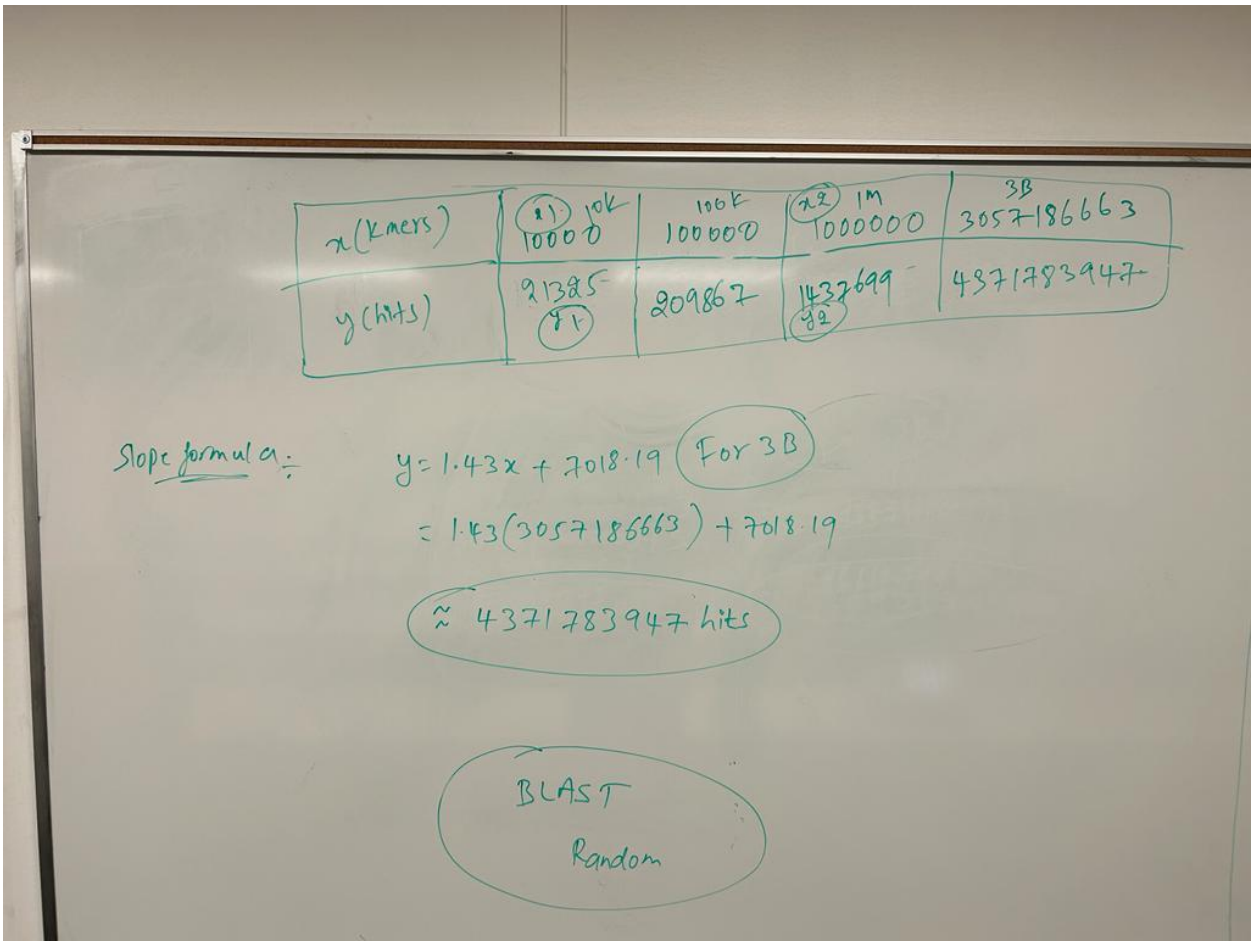
Time complexity for NW : O(N) * O(G) * O(n * n)
Time complexity for BLAST : O(N) * O(n) * O(n)

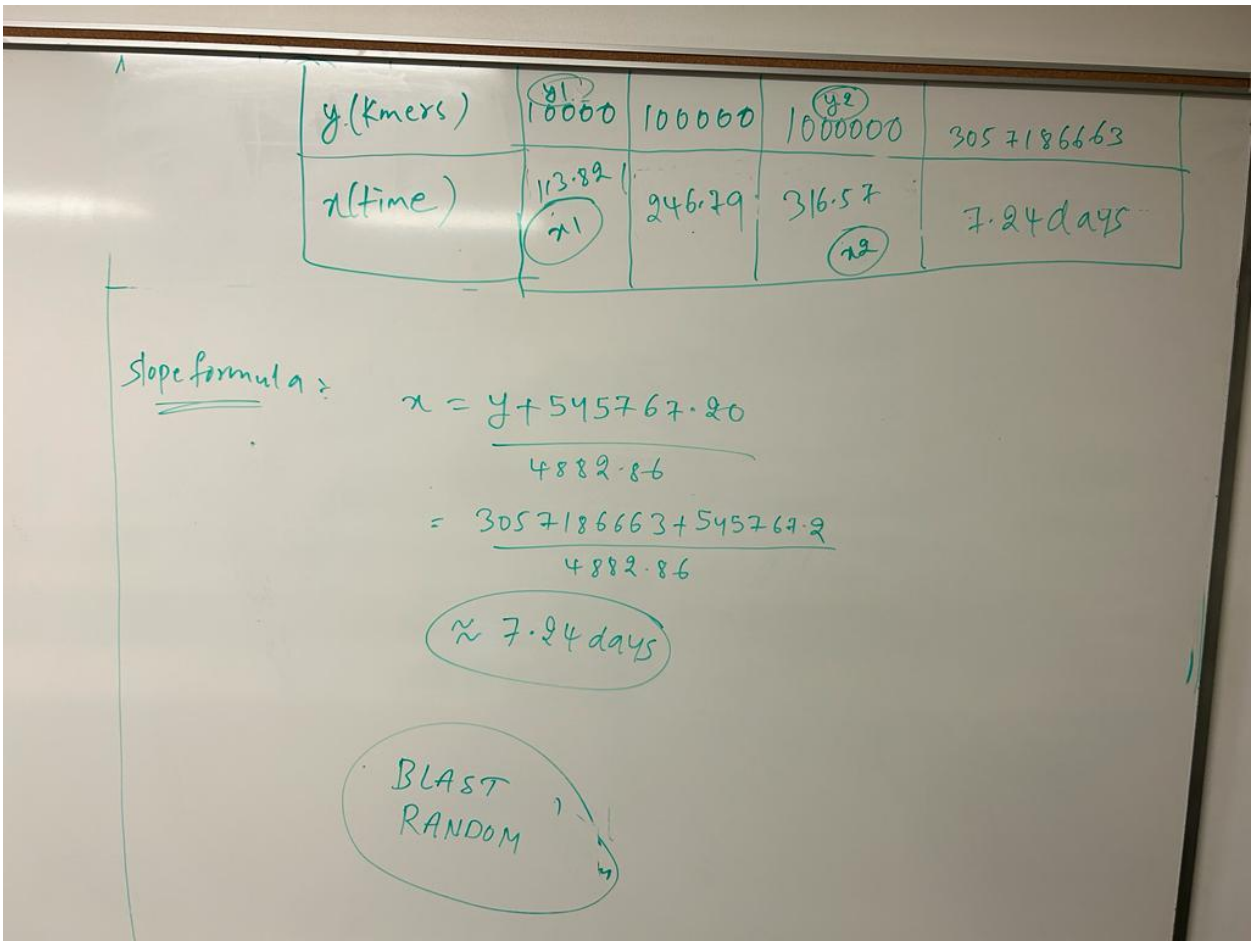
Estimation calculations are mentioned below for both NW and BLAST:

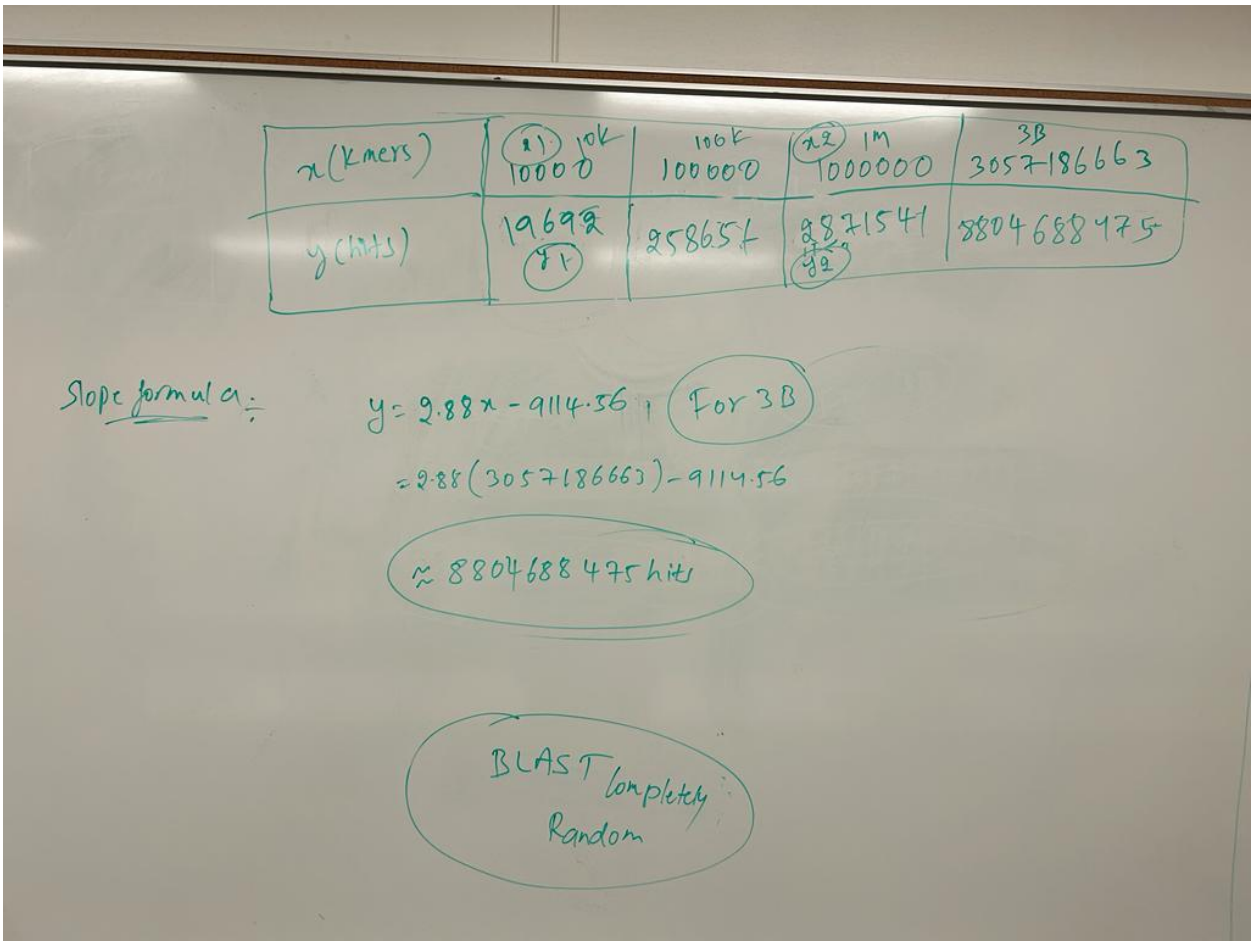

VM Wunsch (Random)
x1
x2
100K
1M
3B
x(kmers)
500
1000
10000
100000
1000000
3057186663
y(hits)
496
965
9864
99003
990003
3026614799
y1
y2
Slope formula
For 100K
y=0.99x+2.95
=0.99(100000)+2.95
=99004 3
=99003 hits
For 1M
=0.99(1000000)+3
=9900003
=990003 hits.
For 3B
=0.99(3057186663)+3
=3026614799 hits

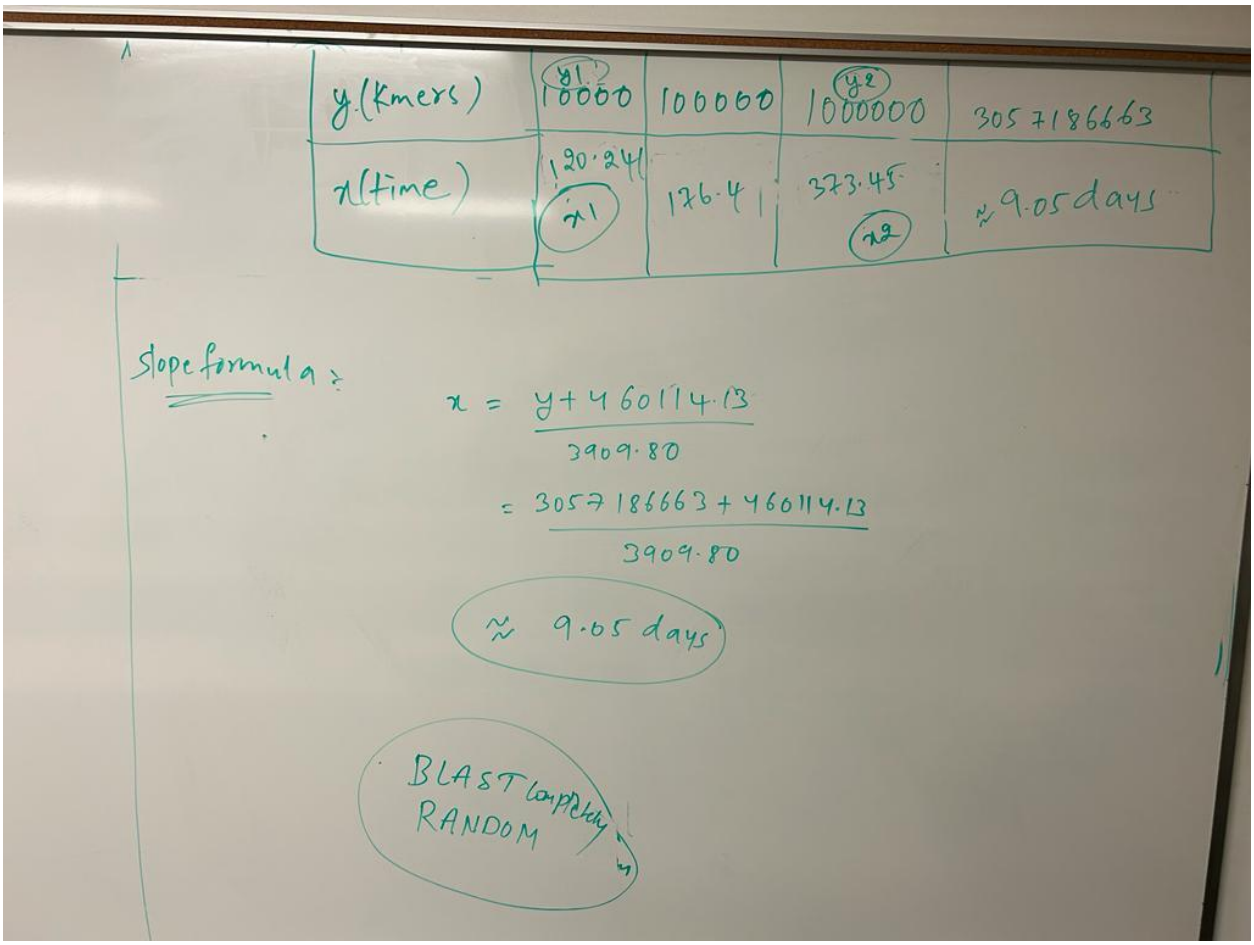

x(kmers)
500
1000
10000
100000
1000000
3057186663
y(time)
13046.7
y1
y2
27689202.26
≈ 20 centuries
Slope formula :
For 100k
y = 27.69x − 797.74
= 27.69(100000) − 797.74
= 2769000 − 797.74
= 2768202.26 seconds
For 1m
= 27.69(1000000) − 797.74
= 27690000 − 797.74
= 27689202.26 seconds
For 3B
= 27.69(3057186663) − 797.74
= 2684.34 years
≈ 20 centuries
km Wunsch
Completely Random

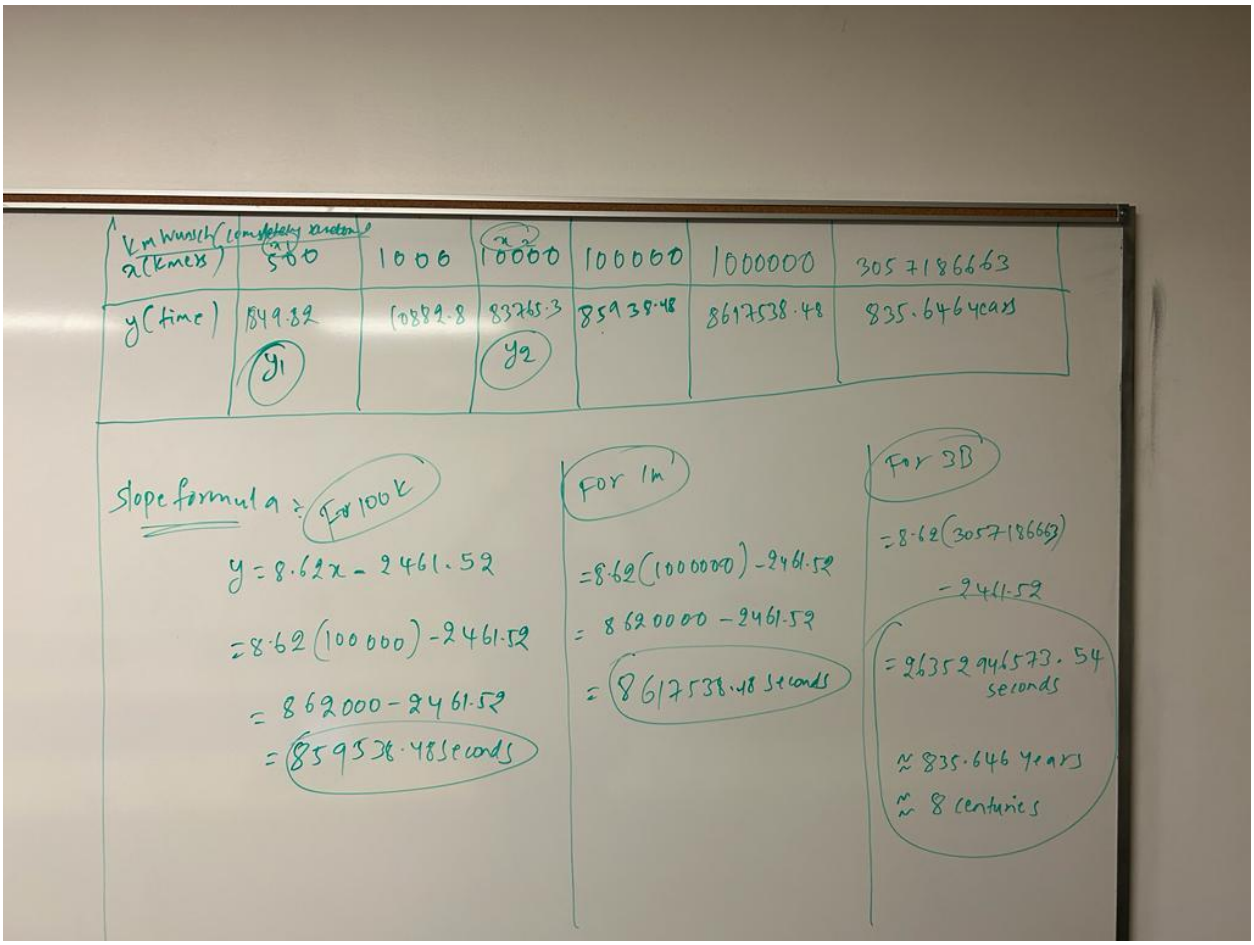

x(kmers)
y(hits)
x1
500
y1
437
1000
853
x2
10000
8657
y2
100K
100000
87005
1M
1000000
870005
3B
3057186663
2659752401
Slope formula
y = 0.87x + 4.37
= 0.87(100000) + 5
= 87000 + 5
= 87005 hits
for 100k
for 1M
= 0.87(1000000) + 5
= 870000 + 5
= 870005 hits
for 3B
= 0.87(3057186663) + 5
= 2659752401 hits
km Wunsch
Random
plot


x (kmers)
10000 10K
100K 100000
x2 1M 1000000
3B 3057186663
y (hits)
21325 y1
209867
1437699 y2
4371783947
Slope formula:
y = 1.43x + 7018.19 For 3B
= 1.43(3057186663) + 7018.19
≈ 4371783947 hits
BLAST
Random


y (Kmers)
y1
10000
100000
y2
1000000
3057186663
x (time)
113.82
x1
246.79
316.57
x2
7.24 days
Slope formula :
x = y + 545767.20 / 4882.86
= 3057186663 + 545767.2 / 4882.86
≈ 7.24 days
BLAST
RANDOM


x (kmers)
10000
100K
1000000
1M
3B
3057186663
y (hits)
25865t
8804688475
Slope formula:
y = 2.88x - 9114.56
For 3B
= 2.88(3057186663) - 9114.56
≈ 8804688475 hits
BLAST completely Random


y (kmers)
y1
10000
100000
y2
1000000
3057186663
x (time)
120.24
x1
176.4
373.45
x2
≈ 9.05 days
Slope formula :
x = y + 460114.13 / 3909.80
= 3057186663 + 460114.13 / 3909.80
≈ 9.05 days
BLAST
RANDOM


Km Wunsch (completely exact)
x(kmers)
500
1000
10000
100000
1000000
3057186663
y(time)
849.82
y1
1089.8
83765.3
y2
85938.48
8617538.48
835.646 years
Slope formula:
For 100k
y = 8.62x - 2461.52
= 8.62(100000) - 2461.52
= 862000 - 2461.52
= 859538.48 seconds
For 1m
= 8.62(1000000) - 2461.52
= 8620000 - 2461.52
= 8617538.48 seconds
For 3B
= 8.62(3057186663) - 2461.52
= 26352946573.54 seconds
≈ 835.646 years
≈ 8 centuries