Subject: Large Scale Data Structures

Assignment: 04

By Saurabh Kakade

Sk2354@nau.edu

1. Part 01: A (OUTPUT)

```
[sk2354@ondemand /scratch/sk2354/A04/Part01 ]$ srun --mem=05GB -t 00:60:00 ./homework A ./hw3_dataset.fa ./test_genome.fasta


The number of arguments passed: 4
The first argument is: /scratch/sk2354/A04/Part01/./homework
The second argument is: A
The third argument is: ./hw3_dataset.fa
The fourth argument is: ./test_genome.fasta

Function in C++ that implements the Smith-Waterman alignment between two genomic sequences.:

Number Of Read Query Lines (reads): 30

Total 70-character fragments: 427

**************************************************

Sequence 01 and Sequence 02:

AGGGTTCAGGAAAGAGTTTGGAAGGATGCTGAACATCTTGAATAGGAGAC
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA

Max score in the matrix is: 18

Alignment of sequence 01 and sequence 02: (traceback_array)

TGGAAGGATGCTGAACATC
| x||x|x|xxx x||x|x
T_TAAAGGTTTA_TACCTT

**************************************************

**************************************************

Sequence 01 and Sequence 02:

AGGGTTCAGGAAAGAGTTTGGAAGGATGCTGAACATCTTGAATAGGAGAC
ACAAATAGACATATAAAGATAGACAAACATACAAACAGACAAAGATAAATAAAAATATAAAAATAAAAAC

Max score in the matrix is: 20

Alignment of sequence 01 and sequence 02: (traceback_array)

AAAGAG_TTTGGAAGGAT_GCTGAACATCTTGAATAG
x||xxx xxxxxx|xxxx xxxx||xxx |xx||xxx
CAAATAGACATATAAAGATAGACAAACA_TACAAACA
```

```
**************************************************

Sequence 01 and Sequence 02:

AGGGTTCAGGAAAGAGTTTGGAAGGATGCTGAACATCTTGAATAGGAGAC
CGCACGCGCCCACCCGCTCCCACACCCACTCCCTCTCACACACGCACTCGCGCCCACCCTCTCGCTCGCG

Max score in the matrix is: 11

Alignment of sequence 01 and sequence 02: (traceback_array)

GCTGAACATCTTG
xxxxx||xxxx|x
CGCCCACCCTCTC

**************************************************

**************************************************

Sequence 01 and Sequence 02:

AGGGTTCAGGAAAGAGTTTGGAAGGATGCTGAACATCTTGAATAGGAGAC
GTGTGGGAGAGCGAGGGCGCGCGTGAGTGGGTGGGTGTGCGAGTGCGAGAGAGCGGGTGTGCGGGGGAGT

Max score in the matrix is: 23

Alignment of sequence 01 and sequence 02: (traceback_array)

TTGGAAGGATGCTGAACATCTTGAATAGGAG
|xx|xxx||xxxx|xx|xxxxxxx|xx||xx
TGTGGGAGAGCGAGGGCGCGCGTGAGTGGGT

**************************************************

**************************************************

Sequence 01 and Sequence 02:

AGGGTTCAGGAAAGAGTTTGGAAGGATGCTGAACATCTTGAATAGGAGAC
TGTCTTTGTGTTTATGTCTATGTATATCTTTCTGTATATGTGTCTATTTTTCTATGTTTATCTGTGTTTC

Max score in the matrix is: 19

Alignment of sequence 01 and sequence 02: (traceback_array)

GAGTTTGGAAGGATGCTGAACATCTTGAATAGGAG
xxxx||xxxxxxxxxx|xx|xx|x|xxxxxxxxxx
TGTGTTTATGTCTATGTATATCTTTCTGTATATGT
```

```
**************************************************

Sequence 01 and Sequence 02:

AGGGTTCAGGAAAGAGTTTGGAAGGATGCTGAACATCTTGAATAGGAGAC
TATTTATGTTTTTATCTGTGTCTGTCTCTGTATTTCTTTATATATGTTTCTATTTTTTTGTATCTTTTTA

Max score in the matrix is: 19

Alignment of sequence 01 and sequence 02: (traceback_array)

TTTGGAAGGATGCTGAACATCTTGA_ATAGG
x|xx|xxx|xx xxxxxxx|xx|xx xxxxx
CTGTGTCTGTC_TCTGTATTTCTTTATATAT

**************************************************

**************************************************

Sequence 01 and Sequence 02:

AGGGTTCAGGAAAGAGTTTGGAAGGATGCTGAACATCTTGAATAGGAGAC
ATACAAAAAGAAAAAAAAACATAGAGAAAAACAAACATAAAAAAACAAATAAAGACAAAGATAGAGATAG

Max score in the matrix is: 20

Alignment of sequence 01 and sequence 02: (traceback_array)

GGATGCTGAACATCTTGAATAGGAGA
xxxxxxxxx|x|xxxxx|xxx xxxx
TACAAAAAGAAAAAAAAACAT_AGAG

**************************************************

**************************************************

Sequence 01 and Sequence 02:

AGGGTTCAGGAAAGAGTTTGGAAGGATGCTGAACATCTTGAATAGGAGAC
GGGCGAGTGAGCGAGCGTGCGGGCGTGAGTGGGTGCGGGAGTGAGAGCGAGAGCGTGTGCGTGGGTGGGG

Max score in the matrix is: 22

Alignment of sequence 01 and sequence 02: (traceback_array)

GGGTTCAGGAAAGAGTTTGGAAGGATGCTG
x||x|xx||x|xxxxxxxx||xx|xxx xx
TGGGTGCGGGAGTGAGAGCGAGAGCGT_GT
```

2. Part 01 – B
    a. For 1k random genomic sequence – Time required: 02 sec (by jobstats monsoon command)

```
[sk2354@ondemand /scratch/sk2354/A04/Part01 ]$ srun --mem=05GB -t 00:60:00 ./homework B ./hw3_dataset.fa ./test_genome.fasta


The number of arguments passed: 4
The first argument is: /scratch/sk2354/A04/Part01/./homework
The second argument is: B
The third argument is: ./hw3_dataset.fa
The fourth argument is: ./test_genome.fasta

Generate 1K, 10K, 100K, and 1M (million) completely random genomic sequences (50nt) to use as targets for alignment and use SA
 and record time to completion (in seconds / minutes).:


Total 70-character fragments: 427

**************************************************

Sequence 01 and Sequence 02:

CAGAAACTCGAACGGTGCGCTTCTTGAGTCGACAACACACACGCTACCCT
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA

Max score in the matrix is: 20

Alignment of sequence 01 and sequence 02: (traceback_array)

AGAAAC_TCGAACGGTGCGCT_TCTTG_AG_TC
x |||| xxxx|x x|xxxxx xxx|x xx xx
C_AAACCAACCAA_CTTTCGATCTCTTGTAGAT

**************************************************

**************************************************

Sequence 01 and Sequence 02:

CAGAAACTCGAACGGTGCGCTTCTTGAGTCGACAACACACACGCTACCCT
ACAAATAGACATATAAAGATAGACAAACATACAAACAGACAAAGATAAATAAAAATATAAAAATAAAAAC

Max score in the matrix is: 24

Alignment of sequence 01 and sequence 02: (traceback_array)

_GGTGCGCTTCTTGAG_TCGAC_AACACACA
 xxxxx|xxxxx|x|x xxxxx ||xxxxxx
CAAATAGACATATAAAGATAGACAAACATAC
```

b. For 10k random genomic sequence – Time required: 01 sec (by jobstats monsoon command)

```
==================
[sk2354@ondemand /scratch/sk2354/A04/Part01 ]$ make
g++ -c homework.cpp homework.h homework_BLAST.h
g++ -o homework homework.o
[sk2354@ondemand /scratch/sk2354/A04/Part01 ]$ srun --mem=10GB -t 00:60:00 ./homework B ./hw3_dataset.fa ./test_genome.fasta


The number of arguments passed: 4
The first argument is: /scratch/sk2354/A04/Part01/./homework
The second argument is: B
The third argument is: ./hw3_dataset.fa
The fourth argument is: ./test_genome.fasta

Generate 1K, 10K, 100K, and 1M (million) completely random genomic sequences (50nt) to use as targets for alignment and use SARS-
 and record time to completion (in seconds / minutes).:


Total 70-character fragments: 427

**************************************************

Sequence 01 and Sequence 02:

ACCACTTGCGGTAATCTTTTACGATACGTGTTTATATGTCACACCTGATC
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA

Max score in the matrix is: 31

Alignment of sequence 01 and sequence 02: (traceback_array)

CC_ACTTGCGGTAATCTTTTACGATACGTGTT_TATA
x|  |xx|xxx|x xxxx||xx xxx x xxx|  xxxx
ACCAACTTTCGA_TCTCTTGT_AGA_T_CTGTTCTCT

**************************************************

**************************************************

Sequence 01 and Sequence 02:

ACCACTTGCGGTAATCTTTTACGATACGTGTTTATATGTCACACCTGATC
ACAAATAGACATATAAAGATAGACAAACATACAAACAGACAAAGATAAATAAAAATATAAAAATAAAAAC

Max score in the matrix is: 16

Alignment of sequence 01 and sequence 02: (traceback_array)

ACGATACGTGTTTATAT_GTCACACCT
|xxxxxxxxxx|x|xxx xxxxx|x|x
```

c. For 100k random genomic sequence – Time required: 01 sec (by jobstats monsoon command)

```
[sk2354@ondemand /scratch/sk2354/A04/Part01 ]$ make
g++ -c homework.cpp homework.h homework_BLAST.h
g++ -o homework homework.o
[sk2354@ondemand /scratch/sk2354/A04/Part01 ]$ srun --mem=10GB -t 00:60:00 ./homework B ./hw3_dataset.fa ./test_genome.fasta


The number of arguments passed: 4
The first argument is: /scratch/sk2354/A04/Part01/./homework
The second argument is: B
The third argument is: ./hw3_dataset.fa
The fourth argument is: ./test_genome.fasta

Generate 1K, 10K, 100K, and 1M (million) completely random genomic sequences (50nt) to use as targets for alignment and use SARS-C
 and record time to completion (in seconds / minutes).:


Total 70-character fragments: 427

**************************************************

Sequence 01 and Sequence 02:

CTACGTGTGGTGCTCTCTGGCTATTATTTATAAAGGTACCGCTAACTCAC
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA

Max score in the matrix is: 25

Alignment of sequence 01 and sequence 02: (traceback_array)

TAAAGGT
|x||x|x
TTAAAGG

**************************************************

**************************************************

Sequence 01 and Sequence 02:

CTACGTGTGGTGCTCTCTGGCTATTATTTATAAAGGTACCGCTAACTCAC
ACAAATAGACATATAAAGATAGACAAACATACAAACAGACAAAGATAAATAAAAATATAAAAATAAAAAC

Max score in the matrix is: 19

Alignment of sequence 01 and sequence 02: (traceback_array)

TATTATTTATAAAGGTACCGCTAAC
xx xxxxxxxx||x|xx x|xx||x
```

d. For 1M random genomic sequence – Time required: 03 sec (by jobstats monsoon command)

```
[sk2354@ondemand /scratch/sk2354/A04/Part01 ]$ make
g++ -c homework.cpp homework.h homework_BLAST.h
g++ -o homework homework.o
[sk2354@ondemand /scratch/sk2354/A04/Part01 ]$ srun --mem=10GB -t 00:60:00 ./homework B ./hw3_dataset.fa ./test_genome.fasta


The number of arguments passed: 4
The first argument is: /scratch/sk2354/A04/Part01/./homework
The second argument is: B
The third argument is: ./hw3_dataset.fa
The fourth argument is: ./test_genome.fasta

Generate 1K, 10K, 100K, and 1M (million) completely random genomic sequences (50nt) to use as targets for alignment and use SAR
 and record time to completion (in seconds / minutes).:


Total 70-character fragments: 427

**************************************************

Sequence 01 and Sequence 02:

TATCCTCGAAAAAGGAGTTTCATACGTGAATACTCGGTAGTAGAGCAAAA
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA

Max score in the matrix is: 25

Alignment of sequence 01 and sequence 02: (traceback_array)

_GGAGTTTCATACGTGAATACTCGGTAGTAGAGCAA
 xx|xxx|xxxxxxx    |x|x|x|xxxxxx|x|x|
TTAAAGGTTTATACC___TTCCCAGGTAACAAACCA

**************************************************

**************************************************

Sequence 01 and Sequence 02:

TATCCTCGAAAAAGGAGTTTCATACGTGAATACTCGGTAGTAGAGCAAAA
ACAAATAGACATATAAAGATAGACAAACATACAAACAGACAAAGATAAATAAAAATATAAAAATAAAAAC

Max score in the matrix is: 22

Alignment of sequence 01 and sequence 02: (traceback_array)

TAGTAGAGCAAA
x|xxxxx xx|x
```

3. Part 02: A

```
[sk2354@ondemand /scratch/sk2354/A04/Part01 ]$ srun --mem=05GB -t 00:60:00 ./homework C ./hw3_dataset.fa ./test_genome.fasta


The number of arguments passed: 4
The first argument is: /scratch/sk2354/A04/Part01/./homework
The second argument is: C
The third argument is: ./hw3_dataset.fa
The fourth argument is: ./test_genome.fasta

:Having a BLAST - Implement a seed-based Smith Waterman:

Number of lines in read query dataset: 30
Number of lines in read genome dataset: 429
Size of HASH table :4194271
Collisions :236
Number of UNIQUE sequences: 25390
Sequence 01 and Sequence 02:

ACAAGTGTGCC
ACAAGTGTGCCCAAGTGTGCCTACGTGCTAGCGCTAACATAGGTTGTAAC

Max score in the matrix is: 20

Alignment of sequence 01 and sequence 02: (traceback_array)

AAGTGTGC
x|xxxxxx
CAAGTGTG

**************************************************
Sequence 01 and Sequence 02:

CAACACAACAA
CAACACAACAAAACACAACAAAGGGAGGTATGCACTGTTATCCGATTTAC

Max score in the matrix is: 20

Alignment of sequence 01 and sequence 02: (traceback_array)

ACACAACA
|xxxx|xx
AACACAAC

**************************************************
Sequence 01 and Sequence 02:
```

```
**************************************************
Sequence 01 and Sequence 02:

CAACAGAATCT
CAACAGAATCTAACGTTAGATTTCCTAATATTACAAACTTGTGCCCTTTT

Max score in the matrix is: 20

Alignment of sequence 01 and sequence 02: (traceback_array)

ACAGAATC
|xxxx|xx
AACAGAAT

**************************************************
Sequence 01 and Sequence 02:

AACATGGCAAG
AACATGGCAAGACATGGCAAGGAAGACCTTAAATTCCCTCGAGGACAATT

Max score in the matrix is: 20

Alignment of sequence 01 and sequence 02: (traceback_array)

CATGGCAA
xxxx|xx|
ACATGGCA

**************************************************
Sequence 01 and Sequence 02:

ACATGGCAAGG
ACATGGCAAGGCATGGCAAGGAAGACCTTAAATTCCCTCGAGGACAATTA

Max score in the matrix is: 20

Alignment of sequence 01 and sequence 02: (traceback_array)

ATGGCAAG
xxx|xx|x
CATGGCAA

**************************************************
Sequence 01 and Sequence 02:

CATGGCAAGGA
CATGGCAAGGAATGGCAAGGAAGACCTTAAATTCCCTCGAGGACAATTAA

Max score in the matrix is: 20
```

4. Part 02: B
    a. For 1k random 50mers sequence – Time required: 02 sec (by jobstats monsoon command)

```
[sk2354@ondemand /scratch/sk2354/A04/Part01 ]$ make
g++ -c homework.cpp homework.h homework_BLAST.h
g++ -o homework homework.o
[sk2354@ondemand /scratch/sk2354/A04/Part01 ]$ srun --mem=10GB -t 00:60:00 ./homework D ./hw3_dataset.fa ./test_genome.fasta


The number of arguments passed: 4
The first argument is: /scratch/sk2354/A04/Part01/./homework
The second argument is: D
The third argument is: ./hw3_dataset.fa
The fourth argument is: ./test_genome.fasta

Test your code on a set of 1K, 10K, 100K, and 1M (million) completely random 50-mers, aligning them to SARS-COV2 genome. How l

Number of lines in read genome dataset: 429
Size of HASH table :4194271
Collisions :236
Number of UNIQUE sequences: 25390
Sequence 01 and Sequence 02:

TGGAGGAGATTCTATGAGATTCAGGGGAGTTCCCCAACGAACCACCGCTG
ATTAAAGGTTTTTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAAC

Max score in the matrix is: 23

Alignment of sequence 01 and sequence 02: (traceback_array)

AGATTCTATGAGATT_CAGGGGAGTTCCCCAACGAAC
|xxx| |xx xxxx| |xxx||xxxxxxx|x|xxx|x
AAGGT_TTA_TACCTTCCCAGGTAACAAACCAACCAA

**************************************************
Sequence 01 and Sequence 02:

TGGAGGAGATTCTATGAGATTCAGGGGAGTTCCCCAACGAACCACCGCTG
TTTCGATCTCTTTCGATCTCTTGTAGATCTCGAACTTTAAAATCTGTGTG

Max score in the matrix is: 22

Alignment of sequence 01 and sequence 02: (traceback_array)

GAT_TCTATGAGAT_TCAGGGGAGTTCCCCAAC
xxx xxx |xxxxx xx xxx|xxxxx|x|x|x
CGATCTC_TTTCGATCT_CTTGTAGATCTCGAA

**************************************************
```

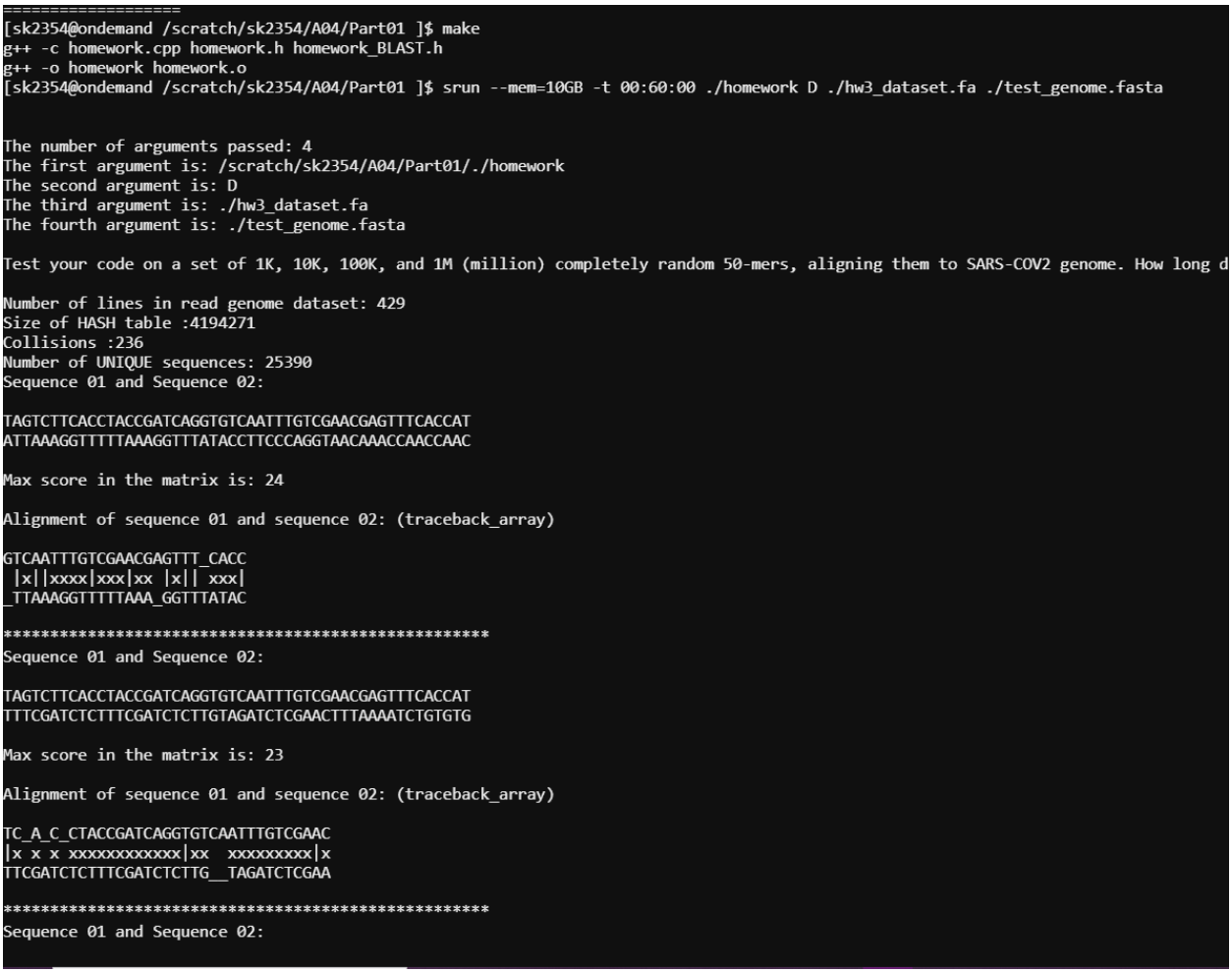
b. For 10k random 50mers sequence – Time required: 01 sec (by jobstats monsoon command)

```
[sk2354@ondemand /scratch/sk2354/A04/Part01 ]$ make
g++ -c homework.cpp homework.h homework_BLAST.h
g++ -o homework homework.o
[sk2354@ondemand /scratch/sk2354/A04/Part01 ]$ srun --mem=10GB -t 00:60:00 ./homework D ./hw3_dataset.fa ./test_genome.fasta


The number of arguments passed: 4
The first argument is: /scratch/sk2354/A04/Part01/./homework
The second argument is: D
The third argument is: ./hw3_dataset.fa
The fourth argument is: ./test_genome.fasta

Test your code on a set of 1K, 10K, 100K, and 1M (million) completely random 50-mers, aligning them to SARS-COV2 genome. How lo

Number of lines in read genome dataset: 429
Size of HASH table :4194271
Collisions :236
Number of UNIQUE sequences: 25390
Sequence 01 and Sequence 02:

GGGAACTATCCTGCACTTATAGAATATCAAGTGATTCGTATATGGTTCTA
ATTAAAGGTTTTTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAAC

Max score in the matrix is: 24

Alignment of sequence 01 and sequence 02: (traceback_array)

ATATGGTTCT
x||xx|x|x|
TTAAAGGTTT

**************************************************
Sequence 01 and Sequence 02:

GGGAACTATCCTGCACTTATAGAATATCAAGTGATTCGTATATGGTTCTA
TTTCGATCTCTTTCGATCTCTTGTAGATCTCGAACTTTAAAATCTGTGTG

Max score in the matrix is: 29

Alignment of sequence 01 and sequence 02: (traceback_array)

TCGTATATGGTTC
|xx xxxxxx||x
TTC_GATCTCTTT

**************************************************
Sequence 01 and Sequence 02:

GGGAACTATCCTGCACTTATAGAATATCAAGTGATTCGTATATGGTTCTA
```

c. For 100k random 50mers sequence – Time required: 01 sec (by jobstats monsoon command)

```
[sk2354@ondemand /scratch/sk2354/A04/Part01 ]$ make
g++ -c homework.cpp homework.h homework_BLAST.h
g++ -o homework homework.o
[sk2354@ondemand /scratch/sk2354/A04/Part01 ]$ srun --mem=10GB -t 00:60:00 ./homework D ./hw3_dataset.fa ./test_genome.fasta


The number of arguments passed: 4
The first argument is: /scratch/sk2354/A04/Part01/./homework
The second argument is: D
The third argument is: ./hw3_dataset.fa
The fourth argument is: ./test_genome.fasta

Test your code on a set of 1K, 10K, 100K, and 1M (million) completely random 50-mers, aligning them to SARS-COV2 genome. How lo

Number of lines in read genome dataset: 429
Size of HASH table :4194271
Collisions :236
Number of UNIQUE sequences: 25390
Sequence 01 and Sequence 02:

AAACGTGCTCGACAGCTATAAGTGTATGCGCCTGGCCATCCCCGGTTCAG
ATTAAAGGTTTTTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAAC

Max score in the matrix is: 19

Alignment of sequence 01 and sequence 02: (traceback_array)

TATAAG__TGTATGCGCCT_GGCCATCCCCGGT
|xx|||  |x|xxxx|xx| xxx|x|x|||x|x
TTAAAGGTTTTTAAAGGTTTATACCTTCCCAGG

**************************************************
Sequence 01 and Sequence 02:

AAACGTGCTCGACAGCTATAAGTGTATGCGCCTGGCCATCCCCGGTTCAG
TTTCGATCTCTTTCGATCTCTTGTAGATCTCGAACTTTAAAATCTGTGTG

Max score in the matrix is: 25

Alignment of sequence 01 and sequence 02: (traceback_array)

GCCATCCC
xx|xxx|x
TTCGATCT

**************************************************
Sequence 01 and Sequence 02:
```

d. For 1M random 50mers sequence – Time required: 03 sec (by jobstats monsoon command)



```
==================
[sk2354@ondemand /scratch/sk2354/A04/Part01 ]$ make
g++ -c homework.cpp homework.h homework_BLAST.h
g++ -o homework homework.o
[sk2354@ondemand /scratch/sk2354/A04/Part01 ]$ srun --mem=10GB -t 00:60:00 ./homework D ./hw3_dataset.fa ./test_genome.fasta


The number of arguments passed: 4
The first argument is: /scratch/sk2354/A04/Part01/./homework
The second argument is: D
The third argument is: ./hw3_dataset.fa
The fourth argument is: ./test_genome.fasta

Test your code on a set of 1K, 10K, 100K, and 1M (million) completely random 50-mers, aligning them to SARS-COV2 genome. How long d

Number of lines in read genome dataset: 429
Size of HASH table :4194271
Collisions :236
Number of UNIQUE sequences: 25390
Sequence 01 and Sequence 02:

TAGTCTTCACCTACCGATCAGGTGTCAATTTGTCGAACGAGTTTCACCAT
ATTAAAGGTTTTTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAAC

Max score in the matrix is: 24

Alignment of sequence 01 and sequence 02: (traceback_array)

GTCAATTTGTCGAACGAGTTT_CACC
 |x||xxxx|xxx|xx |x|| xxx|
_TTAAAGGTTTTTAAA_GGTTTATAC

**************************************************
Sequence 01 and Sequence 02:

TAGTCTTCACCTACCGATCAGGTGTCAATTTGTCGAACGAGTTTCACCAT
TTTCGATCTCTTTCGATCTCTTGTAGATCTCGAACTTTAAAATCTGTGTG

Max score in the matrix is: 23

Alignment of sequence 01 and sequence 02: (traceback_array)

TC_A_C_CTACCGATCAGGTGTCAATTTGTCGAAC
|x x x xxxxxxxxxxx|xx  xxxxxxxxx|x
TTCGATCTCTTTCGATCTCTTG__TAGATCTCGAA

**************************************************
Sequence 01 and Sequence 02:
```

Comparison: Both results from Part 01 B and Part 02 B are same in timings as mentioned above individually.