Frederik Gantzel 5909161 04/01/2021

INF-503 – Homework 4

Problem 1

A)

```
The number of arguments passed: 3
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW4/./homework
The second argument is: Problem1A
The third argument is: 1000

Alignment bewteen read 1 and COVID genome sequence:
A _ _ _ _ T A C A _ G A A A G A C G T T C T T G _ A G T G T A A T G _ T G A A A A C T A C _ C G A A G T T G T A G G A G A C
|         |   | |   | | | | | |   | | |   | x |   | |   |     | | |   | | | | x |   |   |   x | | |         | | | | | | | |
A G G G T T _ C A G G A A A G A _ G T T _ T G G A A G _ G _ _ A T G C T G A A C A _ T _ C T T G A A _ _ _ _ T A G G A G A C
alignment score: 73

Alignment bewteen read 2 and COVID genome sequence:
_ A C T G C C A C T A A _ _ A G C A T A C A A _ T G T A A C A C A A G C T T _ T C _ G G C A G A C G T G _ G T C C A G A A
  |     | x | | |   | |     |   | x | | | | |   | | |   x |   |       | | |   | |   | | | | | |   x | |   |   | | | | | |
A A _ _ G A C A C _ A A T C A _ C G T A C A A G T G T _ G C _ C _ _ _ C T T C T C A G G C A G A _ A T G A G _ C C A G A A
alignment score: 72

Alignment bewteen read 3 and COVID genome sequence:
A A C G G T C T T T G G C T T G A T G A C G T A G T T T A C _ T G T C C A _ A G A C A T G T G _ A T C _ T G C A C C T C T G A A G A C A T
| |   | |         x | | |   | |     | |       |         | |   | | | | | |   | | |   | |   | |   | x |   | | x | x | |   | |   | | | | | |
A A _ G G _ _ _ _ G G G C _ T G _ _ G A _ _ _ A _ _ _ _ A C A T G T C C A G A G A _ A T _ T G A A A C T T G G A T C T _ T G _ A G A C A T
alignment score: 69

Alignment bewteen read 4 and COVID genome sequence:
A T G A T G C _ _ A _ C T C A A _ C A A C A T T A T C A A C A A T G C A A G A G A T G G T T G _ T G T T C C C T T G A A C A T A A
|   | | x | |     |   | |   | |   | | | |       |   | | | | | |     |   | | |   | |   x | x |   | |     | | |       | | | | x | |
A _ G A G G C T A A G C T _ A A C C A A C _ _ _ A _ C A A C A A _ _ C _ A G A _ A T _ C T C G C T G _ _ C C C _ _ _ A A C A C A A
alignment score: 66
```

```
Alignment bewteen read 5 and COVID genome sequence:
A A T G T T A A C A A A G G T G A A G A C A T _ T C A A C T T C T T A A G A G T G C T T A T G A A A A T T T T A A T C A G C A C G A _ A G T T C
| | | |       | | | x | | |   | | |   | | | |   x | | |         x x | |   |   |   |           | | | |       | | |   | | x | | |   |   | | | | |
A A T G _ _ _ A C A C A G G _ G A A _ A C A T G G C A A _ _ _ _ G G A A _ A _ T _ C _ _ _ _ _ A A A A _ _ _ T A A _ C A C C A C _ A G A G T T C
alignment score: 66

Alignment bewteen read 6 and COVID genome sequence:
A T G T T A C A A A A G A A A A T G A C T C T A A A G A G G G T T T T T T C A C T T A C A T T T G T
| | | | | | | x | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
A T G T T A C G A A A G A A A A T G A C T C T A A A G A G G G T T T T T T C A C T T A C A T T T G T
alignment score: 144

Alignment bewteen read 7 and COVID genome sequence:
A G C T C T T G G A G G T T C C G T G G C T A T A A A G A T A A C A G A A C A T T C T T G G A A T G
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | x | | | | | | | | | | | | | | | |
A G C T C T T G G A G G T T C C G T G G C T A T A A A G A T A A C G G A A C A T T C T T G G A A T G
alignment score: 144

Alignment bewteen read 8 and COVID genome sequence:
G G A C A C T T C G C A T G G T G G A C A G C C T T T G T T A C T A A T G T G A A T G C G T C A T C
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
G G A C A C T T C G C A T G G T G G A C A G C C T T T G T T A C T A A T G T G A A T G C G T C A T C
alignment score: 150

Alignment bewteen read 9 and COVID genome sequence:
G A T G T A A T T A T C T T G G C A A A C C A C G C G A A C A A A T A G A T G G T T A T G T C A T G
| | | | | | | | | | | | | | | | | x | | | | | | | | | | | | | | | | x | | | | | | | | | | | x | | |
G A T G T A A T T A T C T T G G C T A A C C A C G C G A A C A A A T T G A T G G T T A T G T T A T G
alignment score: 132
```

```
Alignment bewteen read 10 and COVID genome sequence:
G A G G A A T A C A A A T C C A A T T C A G T T G T C T T C C T A T T C T T T A T T T G A C A T G A
| | | | | | | | | x | | | | | | | | | | | x | | | | | | x | | | | | | | | | | x | | | | | | | | | |
G A G G A A T A C C A A T C C A A T T C A C T T G T C T C C C T A T T C T T T C T T T G A C A T G A
alignment score: 126

Alignment bewteen read 11 and COVID genome sequence:
A G G G G T A C T G C T G T T A T G T C T T T A A A A G A A G G T C A A A T C A A T G A T A T G A T
| | | | | | | | | | | | | | | | | | | | | | |       | | | | | | | | | | | | | | | | | | | | | | | |
A G G G G T A C T G C T G T T A T G T C T T T _ _ _ A G A A G G T C A A A T C A A T G A T A T G A T
alignment score: 135

Alignment bewteen read 12 and COVID genome sequence:
G T A G A C T T A T A A T T A G A G A A A A C A A C A G A G T T G T T A T T T C T A G T G A T G T T
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
G T A G A C T T A T A A T T A G A G A A A A C A A C A G A G T T G T T A T T T C T A G T G A T G T T
alignment score: 150

Alignment bewteen read 13 and COVID genome sequence:
C A A T G T T T G T T T T T C T T G _ _ T T T T A T T G C C A C T A G T C T C T A G T C A G T G T G
| | | | | | | | | | | | | | | | | |     | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
C A A T G T T T G T T T T T C T T G C C T T T T A T T G C C A C T A G T C T C T A G T C A G T G T G
alignment score: 140

Alignment bewteen read 14 and COVID genome sequence:
A T T A C C C C C T G C A T A C A C T A _ _ _ A T T C T T T C A C A C G T G G T G T T T A T T A C C
| | | | | | | | | | | | | | | | | | | |       | | | | | | | | | | | | | | | | | | | | | | | | | | |
A T T A C C C C C T G C A T A C A C T A G G G A T T C T T T C A C A C G T G G T G T T T A T T A C C
alignment score: 135
```

```
Alignment bewteen read 15 and COVID genome sequence:
G T T T T A C A T T C A A C T C A G G A C T T G T T C T T A C C T T T C T T T T C C A A T G T T A C
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
G T T T T A C A T T C A A C T C A G G A C T T G T T C T T A C C T T T C T T T T C C A A T G T T A C
alignment score: 150
```

As can be seen, the Smith-Waterman algorithm was successfully implemented, and the program is outputting the correct allignments.

B)

```
The number of arguments passed: 3
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW4/./homework
The second argument is: Problem1B
The third argument is: 100

Total time to create 100 random sequences: 0 (s)
Total time to perform alignment on all 100 random sequences: 3 (s)
```

```
The number of arguments passed: 3
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW4/./homework
The second argument is: Problem1B
The third argument is: 1000

Total time to create 1000 random sequences: 0 (s)
Total time to perform alignment on all 1000 random sequences: 35 (s)
```

```
The number of arguments passed: 3
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW4/./homework
The second argument is: Problem1B
The third argument is: 10000

Total time to create 10000 random sequences: 0 (s)
Total time to perform alignment on all 10000 random sequences: 352 (s)
```

As can be seen, it takes a long time to perform alignment on all random sequences. The program seems to scale in a roughly linear fashion, and thus, I estimate that it should take about 3,500 seconds (about an hour) to align 100,000 random sequences, and about 35,000 seconds (about 10 hours) to align 1,000,000 random sequences

Problem 2

A)

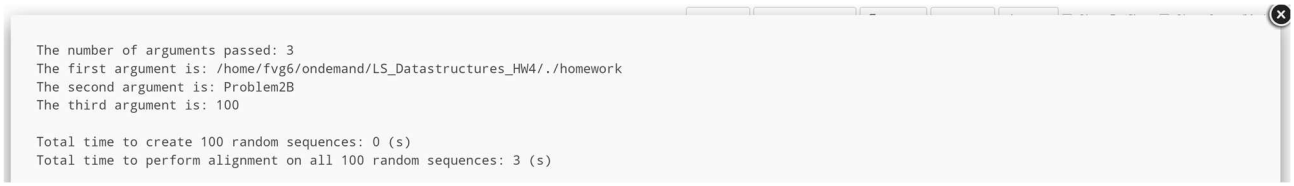
```
The number of arguments passed: 3
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW4/./homework
The second argument is: Problem2A
The third argument is: 1000

Alignment bewteen read 1 and COVID genome sequence:
could not find an 11 character long match between read 1 and the genome sequence

Alignment bewteen read 2 and COVID genome sequence:
_ _ G _ C _ C _ A T _ A _ _ _ A C A A G T G T G _ C C T A T _ T _ G G G T T C C A C G _ _ T G C T A G C G C T A _ A C A
    |   |   |   | |   |       | | | | | | | | |   | | |   |   |   x | |       | |   |     | |     | | |   |   |   |   |
A A G A C A C A A T C A C G T A C A A G T G T G C C C T _ T C T C A G G _ _ _ C A _ G A A T G _ _ A G C _ C _ A G A _ A
alignment score: 53

Alignment bewteen read 3 and COVID genome sequence:
could not find an 11 character long match between read 3 and the genome sequence

Alignment bewteen read 4 and COVID genome sequence:
A _ A T G C G T T A G C T T A C T A C A A C A C A A C A A _ A G G G A G G T A G G T _ T T G T A C T T G C A C
|   | x | |   | x | | | | x | |     | | | | | | | | | | |   |     | |       |     |   | x |     |   | | |   |
A G A G G C _ T A A G C T A A C _ _ C A A C A C A A C A A C A _ _ G A _ _ _ A _ _ T C T C G _ _ C _ T G C _ C
alignment score: 56

Alignment bewteen read 5 and COVID genome sequence:
_ A T G G T A _ A _ A _ T C A A A _ A T G T G _ A A _ G A A T C A T C T G C A A A A T _ _ C A G C
  | |   |   |   |   |   x x | | |   | | |   |   | |   | | |     |     |   | | | | | |     | |   |
A A T _ G _ A C A C A G G G A A A C A T G _ G C A A G G A A _ _ A _ _ T _ C A A A A T A A C A _ C
alignment score: 47
```

```
Alignment bewteen read 6 and COVID genome sequence:
A T G T T A C A A A A G A A A A T G A C T C T A A A G A G G G T T T T T T C A C T T A C A T T T G T
| | | | | | | x | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
A T G T T A C G A A A G A A A A T G A C T C T A A A G A G G G T T T T T T C A C T T A C A T T T G T
alignment score: 144

Alignment bewteen read 7 and COVID genome sequence:
A G C T C T T G G A G G T T C C G T G G C T A T A A A G A T A A C A G A A C A T T C T T G G A A T G
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | x | | | | | | | | | | | | | | | |
A G C T C T T G G A G G T T C C G T G G C T A T A A A G A T A A C G G A A C A T T C T T G G A A T G
alignment score: 144

Alignment bewteen read 8 and COVID genome sequence:
_ _ A G A C T G T _ G T T A T G T A T G C A T C A G _ C T G T A G T G T T A C T A A T C C T _ T A T G A C A G _ C A
    | x | | |   |   |   x | | |   x | | x |   | | |   | |   |     | | | | | | | | | |   x |   x | | |   |   |   | |
G G A C A C T _ T C G _ C A T G _ G T G G A _ C A G C C T _ T _ _ T G T T A C T A A T _ G T G A A T G _ C _ G T C A
alignment score: 62

Alignment bewteen read 9 and COVID genome sequence:
G A T G _ _ _ T T _ T C T T _ T T T T A G C A _ _ C _ A T A T T C _ A G T G G A T G G T T A T G T T C A
| | | |       | |   | | | |   x x | x | x | |     |   |   |     |   | x | x | | | | | | | | | | | |   |
G A T G T A A T T A T C T T G G C T A A C C A C G C G A _ A _ _ C A A A T T G A T G G T T A T G T T _ A
alignment score: 55

Alignment bewteen read 10 and COVID genome sequence:
G A T G G T G A A G T T A T C _ A _ C C _ T T T G A C A A T C T T A A G _ _ A C A C _ _ T T C T T T C T T T G A G A _ G A
| |     |   | | |     | | x |   |   | |   x | |     |   |   | | |     |     x | x |     | | | | | | | | | | | | x |   | |
G A _ _ G _ G A A _ _ T A C C A A T C C A A T T _ _ C _ A _ C T T _ _ G T C T C C C T A T T C T T T C T T T G A C A T G A
alignment score: 58
```

```
Alignment bewteen read 11 and COVID genome sequence:
A G A G T C G A A T G T A C A A C T A T T G T T A A T G G T _ G T T A G A A G G T C C T T T T A T G T C T A T G C T A A T G G A
| |   |    |      | |      | | | x x | | | |   | |   | |   x | | | | | | | | |   |        | x x | | x | | | x |   | |   | |
A G _ G _ _ G _ _ _ G T _ _ _ A C T G C T G T T _ A T _ G T C T T T A G A A G G T _ C _ _ _ _ A A A T C A A T G A T _ A T _ G A
alignment score: 55

Alignment bewteen read 12 and COVID genome sequence:
G T A A A C C T T C A G T T G A A _ C A G A G A A A A A C A A G A T G A _ T A A G _ A A A A T C A A A G C T T G _ T G T T
| | | x |   | | |   |    |   | |   x | | | | | | | | | | | x |   | |   |   x |   x | x x | |   x | |    | |   | | | |
G T A G A _ C T T _ A _ _ T _ A A T T A G A G A A A A A C A A C A _ G A G T _ T G T T A T T T C _ T A G _ _ T G A T G T T
alignment score: 62

Alignment bewteen read 13 and COVID genome sequence:
C _ _ T G T T T G T T T T T G T T A C C T T C T C T T G C C A C T G T A G _ C T _ T A _ T
|     | | | | | | | | | | | x | | x | | | | x | x | | | | | |     | | |   | |   | |   |
C A A T G T T T G T T T T T C T T G C C T T T T T A T T G C C A C _ _ T A G T C T C T A G T
alignment score: 77

Alignment bewteen read 14 and COVID genome sequence:
A T T C A C T G T A C T C T G T T T A A C A C C A _ _ G T T T A C T C A T T C T T A C _ C _ T G G T G T T T A T T _ C
| | |   | |       x | x | | | x x |   | | | | x |     | x | |   | |     | | |     | |   |   | | | | | | | | | | | |   |
A T T _ A C _ _ _ C C C C T G C A T _ A C A C T A G G G A T T _ C T _ _ T T C _ _ A C A C G T G G T G T T T A T T A C
alignment score: 66

Alignment bewteen read 15 and COVID genome sequence:
G _ _ _ _ A C A T T C A A C T T C T T A A G A _ _ _ G T G C T T A _ _ T G A A A A T T _ T T A A T C A G C A C G A A G T
|         | | | | | | | | | |   | |     | x | |       | | x | | | |     |           | |   | |   x | |     | |     | x | |
G T T T T A C A T T C A A C _ T C _ _ A G G A C T T G T T C T T A C C T _ _ _ _ _ T T C T T _ T T C _ _ C A _ _ A T G T
alignment score: 46
```

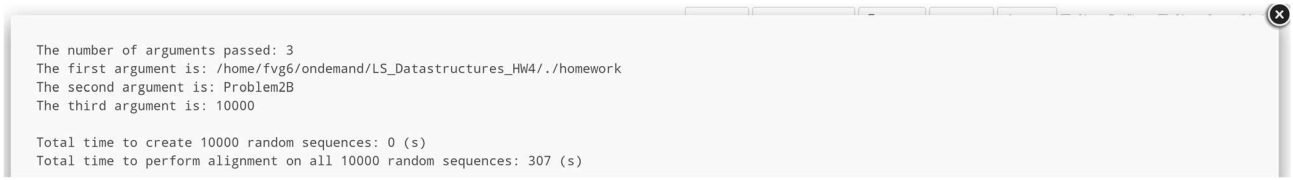
As can be seen, the seed-based Smith-Waterman was implemented. The program choses the first matching seed, and performs alignment on a segment of the COVID sequence around that seed. The first matching seed might not be the best choice however, as can be seen when comparing the results for this problem, and the results for problem 1A: Some of the reads have lower alignment scores.

B)



```
The number of arguments passed: 3
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW4/./homework
The second argument is: Problem2B
The third argument is: 100

Total time to create 100 random sequences: 0 (s)
Total time to perform alignment on all 100 random sequences: 3 (s)
```

```
The number of arguments passed: 3
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW4/./homework
The second argument is: Problem2B
The third argument is: 1000

Total time to create 1000 random sequences: 0 (s)
Total time to perform alignment on all 1000 random sequences: 31 (s)
```



```
The number of arguments passed: 3
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW4/./homework
The second argument is: Problem2B
The third argument is: 10000

Total time to create 10000 random sequences: 0 (s)
Total time to perform alignment on all 10000 random sequences: 307 (s)
```

As can be seen, the seed-based algorithm is faster than the non-seed version, although only slightly. It seems to scale linearly, and thus I estimate that it would take about 3,000 seconds (around 50 minutes) to align 100,000 random sequences, and about 30,000 seconds (around 8 hours) to align 1,000,000 random sequences.

I suspect that I might have done this problem wrong. Right now I am taking each random sequence and splitting it up into k-mers of size 11. I then put this list of k-mers into a direct access hash table. I then go through the list of size 11 seeds made from the COVID sequence for matches. If I were to put the list of seeds into a direct access hash table instead, and then search that hash table for matches with the list of k-mers, the program would go a lot quicker since the list of k-mers is much smaller than the list of seeds. I couldn't figure out how to cut out the corresponding segment of the covid sequence with this method however, so I gave up on it.

C)

```
The number of arguments passed: 3
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW4/./homework
The second argument is: Problem2C
The third argument is: 100

Total time to search all 100 random fragments: 3 (s)
100 fragments were found

After adding random errors:
Total time to search all 100 random fragments: 3 (s)
100 fragments were found
```

```
The number of arguments passed: 3
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW4/./homework
The second argument is: Problem2C
The third argument is: 1000

Total time to search all 1000 random fragments: 26 (s)
1000 fragments were found

After adding random errors:
Total time to search all 1000 random fragments: 26 (s)
999 fragments were found
```

```
The number of arguments passed: 3
The first argument is: /home/fvg6/ondemand/LS_Datastructures_HW4/./homework
The second argument is: Problem2C
The third argument is: 10000

Total time to search all 10000 random fragments: 227 (s)
10000 fragments were found

After adding random errors:
Total time to search all 10000 random fragments: 227 (s)
9997 fragments were found
```

As can be seen, all the random sequence cutouts were found on the genome. Even after adding random errors, almost all the random sequence cutouts were found again on the genome, showing that this seed-based algorithm is resistant to errors in the genome. Judging from the time it took to run the algorithm for 100, 1000 and 10000 random sequence fragments, I'd estimate that it'd take about 4,000 seconds (around an hour and 5 minutes) to run the algorithm with the requires 100,000 random sequence fragments.