CSE 4065 – Computational Genomics

Programming Assignment # 1

150115062 - Nurcihane KÖROĞLU

150114022 - Oğuzhan BÖLÜKBAŞ

Description of the Project

In this project, we have worked on a particular region in a genome, so we have only taken some part of one genome as input. This input can be defined as one line of text in a file where the text contains only A, T, G or C and nothing else (even whitespace). The total number of characters may be very high, but in this assignment, we will assume that we only have 500 characters/bases.

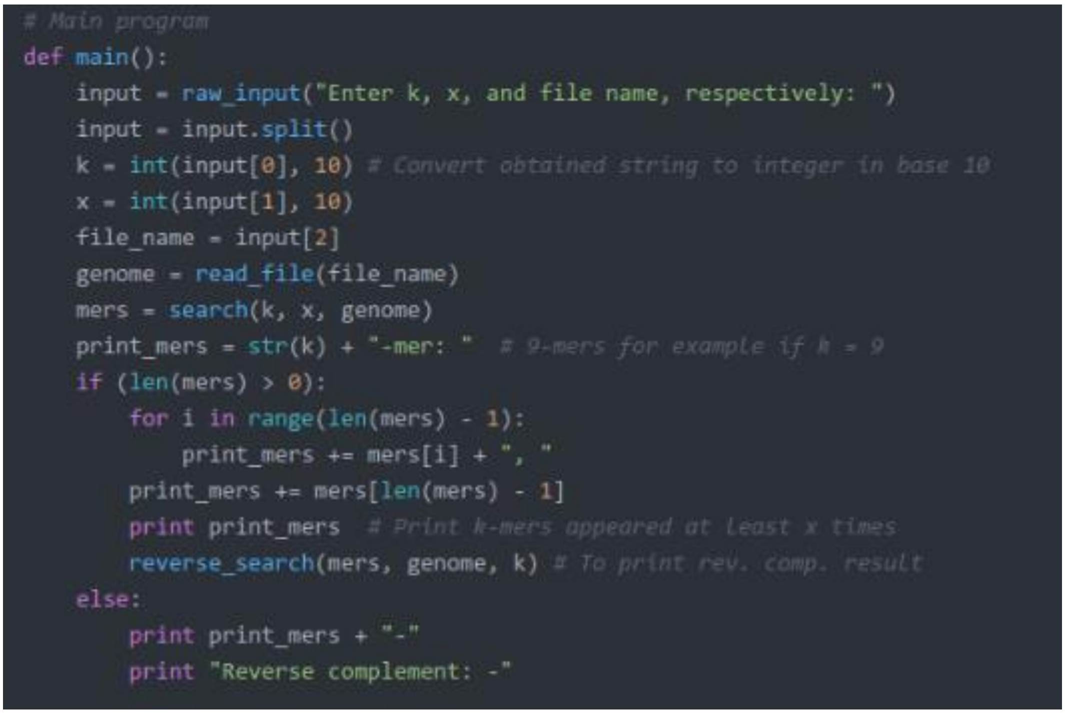
The objective of the project is to find all possible k-mers which is appearing at least x times. Assume the value of k will be at most 9 and x will be at least 2.

After we have found all possible k-mers, we have searched for the reverse complement of each k-mer. Then we give it as output if we find any.

Inputs are: Integer k, integer x, string name of an input file. We ask them to user respectively.

Output: All possible k-mers in the file appearing at least x times. Reverse complement of each k-mer if found any.

Algorithm



```
# Main program
def main():
    input = raw_input("Enter k, x, and file name, respectively: ")
    input = input.split()
    k = int(input[0], 10) # Convert obtained string to integer in base 10
    x = int(input[1], 10)
    file_name = input[2]
    genome = read_file(file_name)
    mers = search(k, x, genome)
    print_mers = str(k) + "-mer: "  # 9-mers for example if k = 9
    if (len(mers) > 0):
        for i in range(len(mers) - 1):
            print_mers += mers[i] + ", "
        print_mers += mers[len(mers) - 1]
        print print_mers  # Print k-mers appeared at least x times
        reverse_search(mers, genome, k) # To print rev. comp. result
    else:
        print print_mers + "-"
        print "Reverse complement: -"
```

Firstly, we take k and x values and file name from the user separated with space char. in one line. Then, we splits the input in order to take inputs correctly. After obtaining file name, we read the file with *read_file()* function and obtain the genome. Then we send the genome to *search()* function in order to find k-mers which appear at least x times. Then we print result. Then we find how many times of reverse complement of the k-mers appear in the genome and print result of it. We find this using *reverse_search()* function. This function calls another function named as *reverse_complement()* to get reverse complement of a pattern. For example, *reverse_search()* sends “ATTGCCGTA” to *reverse_complement()* and obtains “TACGGCAAT”.

Input Files

- input.txt

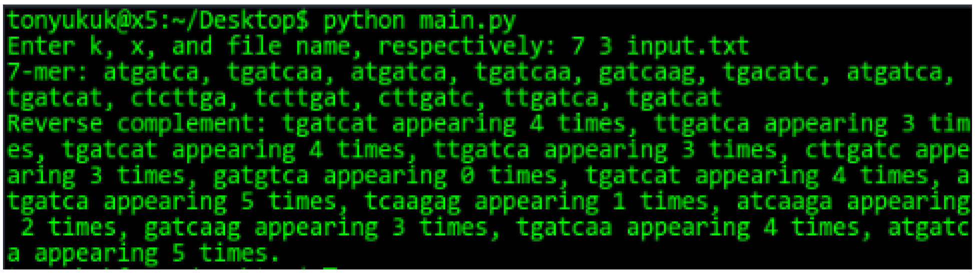
atcaatgatcaacgtaagcttctaagcatgatcaaggtgctcacacagtttatccacaacctgagtggatgacatcaagataggtcgttgt
atctccttcctctcgtactctcatgaccacggaaagatgatcaagagaggatgatttcttggccatatcgcaatgaatacttgtgacttgtgc
ttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggattacgaaagcatgatcatggctgttgttctgtttatct
tgttttgactgagacttgttaggatagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaattgataatga
atttacatgcttccgcgacgatttacctcttgatcatcgatccgattgaagatcttcaattgttaattctcttgcctcgactcatagccatgatg
agctcttgatcatgtttccttaaccctctattttttacggaagaatgatcaagctgctgctcttgatcatcgtttc

- input2.txt

ACCTTTGCAACTCATGACCAAGCTATAATTAAATTACGGACCATGGCTCGTGCCG
GAACTAGGGTTGCGTTAAGCGTAGCCGCAGCGTTGCACCCGCGGTACGAGTTTGG
GTAAAGCCCCTCGAAACAAGTAAGATGTTTATTAACGAGCAAAGAGGGACCGAG
TGTGATCCTCTCGTTACACCGACCGGTTAGTAAGTTTCCAATCCTAGCCTCGTCAG
AACACGCTTTCGCCAGAAACCCTGCTGAATCGCTTCCAGAACAACTAAAGACTCC
TTTAGAACTGTTACCCAAGCGGGTGTAATCTGGCTTTTCCTTGGGAAAAGAAGTC
TCAGCACACCCTAGCTCACATCCCCGGATAGGTACCGGCTTTTAGTGATAATCTT
GAAGGGTGGATTATGCCTGAAGCACCTGCTGGGCCCGATAGCAGGTGAGACCGT
CCACCGGTCCTCTATACCAGTATAGAATAGGCATGTGTGCCCGTCCGTTACAGGC
CCAGGT

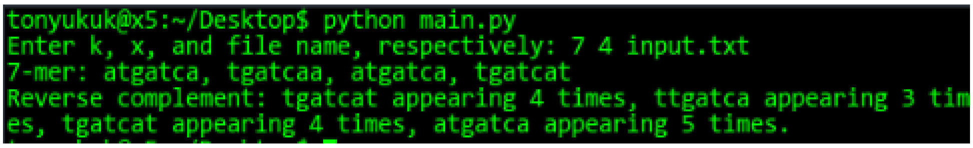
Results

- k = 7, x = 3, file name = input.txt



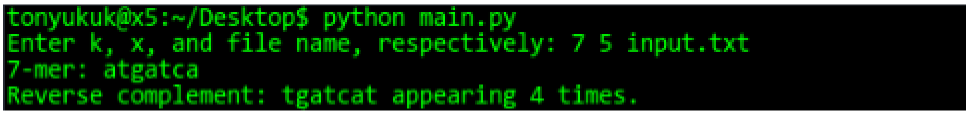
```
tonyukuk@x5:~/Desktop$ python main.py
Enter k, x, and file name, respectively: 7 3 input.txt
7-mer: atgatca, tgatcaa, atgatca, tgatcaa, gatcaag, tgacatc, atgatca,
tgatcat, ctcttga, tcttgat, cttgatc, ttgatca, tgatcat
Reverse complement: tgatcat appearing 4 times, ttgatca appearing 3 tim
es, tgatcat appearing 4 times, ttgatca appearing 3 times, cttgatc appe
aring 3 times, gatgtca appearing 0 times, tgatcat appearing 4 times, a
tgatca appearing 5 times, tcaagag appearing 1 times, atcaaga appearing
 2 times, gatcaag appearing 3 times, tgatcaa appearing 4 times, atgatc
a appearing 5 times.
```

- k = 7, x = 4, file name = input.txt



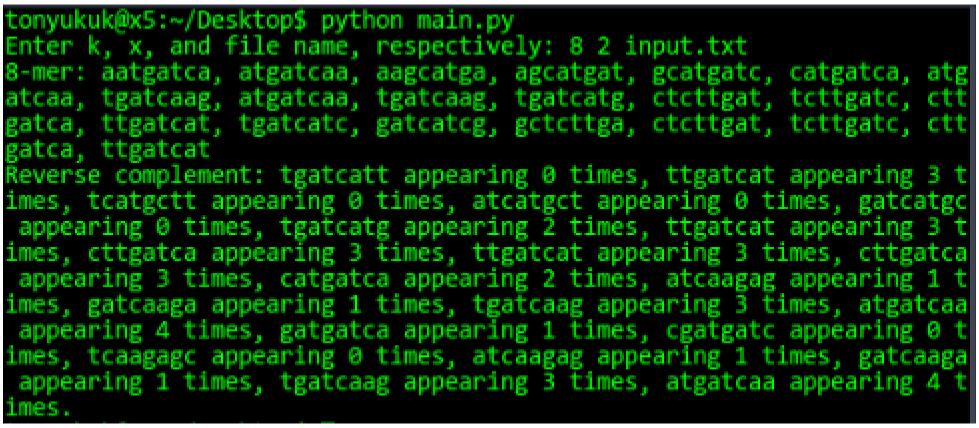
```
tonyukuk@x5:~/Desktop$ python main.py
Enter k, x, and file name, respectively: 7 4 input.txt
7-mer: atgatca, tgatcaa, atgatca, tgatcat
Reverse complement: tgatcat appearing 4 times, ttgatca appearing 3 tim
es, tgatcat appearing 4 times, atgatca appearing 5 times.
```

- k = 7, x = 5, file name = input.txt



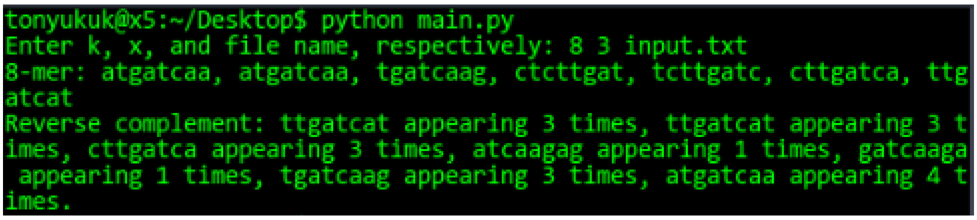
```
tonyukuk@x5:~/Desktop$ python main.py
Enter k, x, and file name, respectively: 7 5 input.txt
7-mer: atgatca
Reverse complement: tgatcat appearing 4 times.
```

- k = 8, x = 2, file name = input.txt



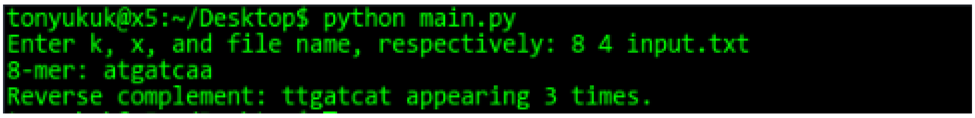
```
tonyukuk@x5:~/Desktop$ python main.py
Enter k, x, and file name, respectively: 8 2 input.txt
8-mer: aatgatca, atgatcaa, aagcatga, agcatgat, gcatgatc, catgatca, atg
atcaa, tgatcaag, atgatcaa, tgatcaag, tgatcatg, ctcttgat, tcttgatc, ctt
gatca, ttgatcat, tgatcatc, gatcatcg, gctcttga, ctcttgat, tcttgatc, ctt
gatca, ttgatcat
Reverse complement: tgatcatt appearing 0 times, ttgatcat appearing 3 t
imes, tcatgctt appearing 0 times, atcatgct appearing 0 times, gatcatgc
 appearing 0 times, tgatcatg appearing 2 times, ttgatcat appearing 3 t
imes, cttgatca appearing 3 times, ttgatcat appearing 3 times, cttgatca
 appearing 3 times, catgatca appearing 2 times, atcaagag appearing 1 t
imes, gatcaaga appearing 1 times, tgatcaag appearing 3 times, atgatcaa
 appearing 4 times, gatgatca appearing 1 times, cgatgatc appearing 0 t
imes, tcaagagc appearing 0 times, atcaagag appearing 1 times, gatcaaga
 appearing 1 times, tgatcaag appearing 3 times, atgatcaa appearing 4 t
imes.
```

- k = 8, x = 3, file name = input.txt



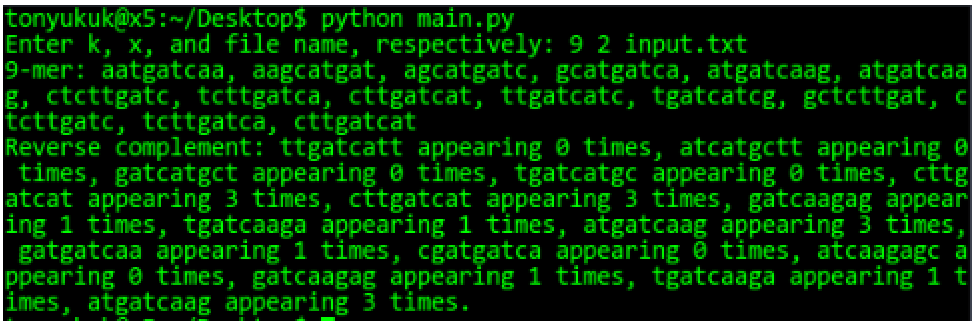
```
tonyukuk@x5:~/Desktop$ python main.py
Enter k, x, and file name, respectively: 8 3 input.txt
8-mer: atgatcaa, atgatcaa, tgatcaag, ctcttgat, tcttgatc, cttgatca, ttg
atcat
Reverse complement: ttgatcat appearing 3 times, ttgatcat appearing 3 t
imes, cttgatca appearing 3 times, atcaagag appearing 1 times, gatcaaga
 appearing 1 times, tgatcaag appearing 3 times, atgatcaa appearing 4 t
imes.
```

- k = 8, x = 4, file name = input.txt



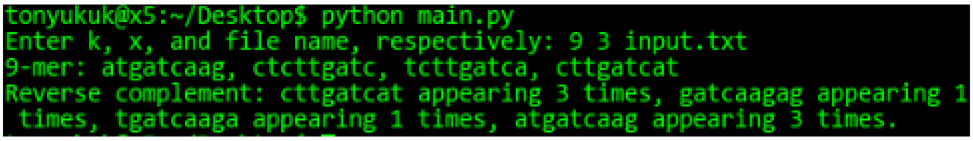
```
tonyukuk@x5:~/Desktop$ python main.py
Enter k, x, and file name, respectively: 8 4 input.txt
8-mer: atgatcaa
Reverse complement: ttgatcat appearing 3 times.
```

- k = 9, x = 2, file name = input.txt



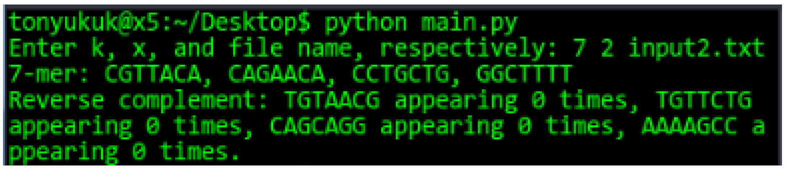
```
tonyukuk@x5:~/Desktop$ python main.py
Enter k, x, and file name, respectively: 9 2 input.txt
9-mer: aatgatcaa, aagcatgat, agcatgatc, gcatgatca, atgatcaag, atgatcaa
g, ctcttgatc, tcttgatca, cttgatcat, ttgatcatc, tgatcatcg, gctcttgat, c
tcttgatc, tcttgatca, cttgatcat
Reverse complement: ttgatcatt appearing 0 times, atcatgctt appearing 0
 times, gatcatgct appearing 0 times, tgatcatgc appearing 0 times, cttg
atcat appearing 3 times, cttgatcat appearing 3 times, gatcaagag appear
ing 1 times, tgatcaaga appearing 1 times, atgatcaag appearing 3 times,
 gatgatcaa appearing 1 times, cgatgatca appearing 0 times, atcaagagc a
ppearing 0 times, gatcaagag appearing 1 times, tgatcaaga appearing 1 t
imes, atgatcaag appearing 3 times.
```

- k = 9, x = 3, file name = input.txt



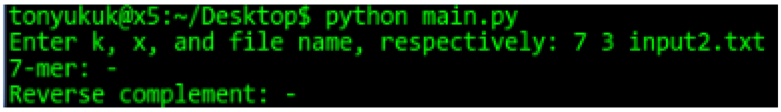
```
tonyukuk@x5:~/Desktop$ python main.py
Enter k, x, and file name, respectively: 9 3 input.txt
9-mer: atgatcaag, ctcttgatc, tcttgatca, cttgatcat
Reverse complement: cttgatcat appearing 3 times, gatcaagag appearing 1
 times, tgatcaaga appearing 1 times, atgatcaag appearing 3 times.
```

- k = 7, x = 2, file name = input2.txt



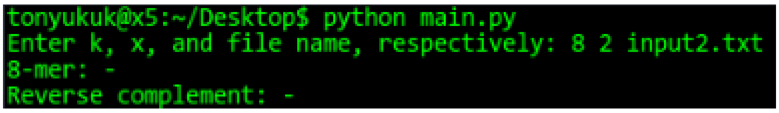
```
tonyukuk@x5:~/Desktop$ python main.py
Enter k, x, and file name, respectively: 7 2 input2.txt
7-mer: CGTTACA, CAGAACA, CCTGCTG, GGCTTTT
Reverse complement: TGTAACG appearing 0 times, TGTTCTG
appearing 0 times, CAGCAGG appearing 0 times, AAAAGCC a
ppearing 0 times.
```

- k = 7, x = 3, file name = input2.txt



```
tonyukuk@x5:~/Desktop$ python main.py
Enter k, x, and file name, respectively: 7 3 input2.txt
7-mer: -
Reverse complement: -
```

- k = 8, x = 2, file name = input2.txt



```
tonyukuk@x5:~/Desktop$ python main.py
Enter k, x, and file name, respectively: 8 2 input2.txt
8-mer: -
Reverse complement: -
```

Unfortunately, we cannot obtain many different results as input.txt file from input2.txt file because input2.txt file has generated randomly. This random generation avoid us to get good results as input.txt. We cannot find results in range from k=7, x=3 to k=9, x=5.