Homework 8

Akshay Kumar Varanasi (av32826)

This homework is due on April 9, 2019 at 4:00pm. Please submit as a PDF file on Canvas. Before submission, please re-run all cells by clicking "Kernel" and selecting "Restart & Run All."

Problem 1 (5 points): In bioinformatics, k-mers refer to all the possible subsequences (of length k) from a read obtained through DNA sequencing. For example, if the DNA sequencing read is "ATCATCATG", then the 3-mers in that read include "ATC" (which occurs twice), "TCA" (which occurs twice), "CAT" (occurs twice), and "ATG" (occurs once). You can read more about k-mers on Wikipedia (https://en.wikipedia.org/wiki/K-mer).

a) Write a function that takes a string of nucleotides as input and returns a **dictionary** with all 2-mers present in that string, and the number of times that each 2-mer occurs. Then use your function to find the 2-mers in the DNA sequence `my_seq` defined below.

The output of your function should be a dictionary that is structured like this (although it will have several more entries):

```
{"AT": 2, "TC": 2, "CA": 1}
```

where each key is a 2-mer itself (e.g., "AT") and each value is the number of times that 2-mer occurs.

b) Come up with a short DNA sequence and use it to verify manually that your function generates the correct result. Explain your reasoning in 2-3 sentences.

In [9]:
```
# Find all 2-mers in this sequences
my_seq = "CCTCTCCCTTATCGTCAATCTTCTCGAGGATTGGGGACCCTGCGCTGAACATGGAGAACATCACATCAG
G"

# Your code goes here
def twomer_count(seq):
    twomers = {}
    for i in range(len(seq)-1):
        if seq[i:i+2] in twomers.keys():
            twomers[seq[i:i+2]]+=1
        else:
            twomers[seq[i:i+2]]=1
    print(twomers)


twomer_count(my_seq)

# Verfication with simple example
my_seq = "CCATCC"

twomer_count(my_seq)
```

```
{'CC': 5, 'CT': 7, 'TC': 9, 'TT': 3, 'TA': 1, 'AT': 6, 'CG': 3, 'GT': 1, 'CA':
6, 'AA': 3, 'GA': 6, 'AG': 3, 'GG': 6, 'TG': 4, 'AC': 4, 'GC': 2}
{'CC': 2, 'CA': 1, 'AT': 1, 'TC': 1}
```

As we can see, the function has been verified with simple example "CCATCC". We can see that there are following possible 2-mers: "CC","CA","AT" and "TC". And only "CC" appears twice whereas others appear only once.

Problem 2 (5 points): DNA sequences are typically stored in a format called FASTA (pronounced fast-ay). A single FASTA file may contain many different sequences. For example, you may have a FASTA file for a mouse, and each mouse gene sequence is stored as a separate sequence in that FASTA file. All sequences in a FASTA file begin on a new line with a greater-than symbol ">" (without quotes).

Write a function that takes the *name* of a FASTA file as input, opens that file, counts the number of sequences in the file (by counting the number of lines in the file that start with a ">" symbol), and returns the count. Download the file "hepatitis_b_genome.fasta (http://wilkelab.org/classes/SDS348/2019_spring/homeworks/hepatitis_b_genome.fasta)" to your computer and use your function to count the number of sequences in the file.

In [8]:
```
# Your code goes here
def fasta_seq_count(filename):
    with open(filename,"r") as inputfile:
        count=0
        lines = inputfile.readlines()
        for i in range(len(lines)):
            if lines[i].startswith(">"):
                count+=1
        print(count)


fasta_seq_count("hepatitis_b_genome.fasta")
```

```
7
```

There are seven sequences in the file "hepatitis_b_genome.fasta"