Homework Exercise 2

1)

Use the file **orf_exons_chr17.txt** that you worked with in the lab exercise.

Recover the coding DNA and protein sequences of the genes.

Find the genes on chromosome 17 with the DNA motif *GCGCGCGCGC* in their coding regions. Find those with the protein motif *RKRKRK*.

2)

A)

Write a program that generates random RNA transcripts, assuming that all 4 RNA nucleotides are of equal probability. Let each sequence begin with a start codon, and end only when a stop codon is introduced (make sure to take the reading frame into account).

B)

Run the simulation 1,000 times. What would be the average transcript length if RNA sequences were created at random?

C)

What is the mean transcript length of random RNA sequences as a function of (mean) GC content? Calculate for the following GC-content values: 10%, 20%, 30%, …, 90%.

Explain the results (how and why does the length of transcripts depend on GC content?).

Bonus Questions

1)

Review the following Python tools and show how to install them:

- Virtualenv
- Pip
- Anaconda

Show how these tools assist in installing Python libraries/packages. Demonstrate how to install the Biopython library (which we will later learn in class) using these tools. Demonstrate how one can install Python libraries without admin permissions using a personal virtual environment.

2)

Explain the difference between mutable and immutable data types in Python. List common data types of each category.

Mutable types are not allowed as dictionary keys. Why do you think it is the case?

Suppose you have a list of strings that you want to concatenate. Which of the following pieces of code does it more efficiently? Why?

Option 1:

```
concatenated_string = ''.join(strings)
```

Option 2:

```
concatenated_string = ''
for string in strings:
    concatenated_string += string
```