CMPSC 300
Introduction to Bioinformatics
Fall 2017

Lab 2: Laboratory Assignment Two: DNA and Python Basics.
Save this lab assignment to: `labs/lab2`

DNA and Python Basics

Bioinformatics requires the use of computational techniques to solve biologically-related problems. Therefore, in this course it is essential for you to have some basic understanding of both biological and computer science concepts related to Bioinformatics before we approach Bioinformatics problems. As we discuss the foundational background material from Biology and computer programming in class in the next two weeks, during the next two lab sessions you will be invited to strengthen your biological understanding and have more hands-on experience with basic programming in Python.

Python has been named the top programming language by IEEE Spectrum magazine in 2017 based on user input and language use-cases. Python is especially useful and popular in the area of Bioinformatics due to its relative easiness to learn, good readability and a large community of developers and users. Specifically, *Biopython*, a project of the *Open Bioinformatics Foundation*, which we will utilize in this class, provides a set of free (open source) tools for biological computation written in Python by a community of developers.

In this lab assignment you are invited to review your understanding of DNA and to explore the magic of Python programming.

Objectives

To strengthen the understanding of DNA structure and DNA replication. To learn and enhance Python programming skills, including how to create variables, write assignment statements, manipulate lists, create repetitive and conditional statements. To utilize basic Python programming skills to write a program that processes and manipulates the DNA sequence.

Reading Assignment

In addition to following the specified sections of the Python tutorial outlined below, please read Chapters 1 and 2 in the "ThinkPython" book. You should also review class slides and videos on DNA structure and replication.

PART 1: DNA Structure and Replication Worksheet

Navigate to the shared course repository ('cs300f2017-share') and from the terminal window type:
`git pull`
to download the lab 2 materials. Once you navigate to 'lab02' directory, you will find an open office word processing file named 'Lab02Part1_DNAReview'. Now, in your own course repository

(cs300f2017-bbill), create 'lab02' directory, and place a copy of the 'Lab02Part1_DNAReview' file there. Rename the file by adding your username to the end of the file name. Now you can open the document, add your name and answer the questions related to DNA structure and replication specified in the document.

PART 2 Task 1: Getting to Know Python

Before we can start writing Python programs for Bioinformatics solutions, you need to get comfortable with the structure and the syntax of this programming language. Just like with any natural or programming language, you need to learn various syntactical and semantic rules when writing in Python.

To get started with Python, please complete Sections 3 and 4.1-4.5 in the Python Tutorial, which can be found in: `https://docs.python.org/2.7/tutorial/`. NOTE: if you have a previous experience in using Python, please browse through the tutorial and then consider lending your expertise to your fellow classmate.

Now that you have been exposed to some basic rules of programming in Python, you are asked to practice those skills by writing programs to accomplish the following small tasks.

PART 2 Task 2: DNA String Counting

DNA (deoxyribonucleic acid) is the building block of every organism. It contains information about hair color, skin tone, allergies, and more. DNA is composed of four bases - adenine, thymine, cytosine, guanine. A string is simply an ordered collection of symbols selected from some alphabet and formed into a word with the length of a string being the number of symbols that it contains. An example of a length 21 DNA string (whose alphabet contains the symbols 'A', 'C', 'G', and 'T') is "ATGCTTCAGAAAGGTCTTACG."

You are to write a Python program that:

Given: A DNA string of length at most 1000 nt.
Return: Four integers (separated by spaces) counting the respective number of times that the bases,'A', 'C', 'G', and 'T' occur in the sequence.

Sample Input:
AAAACCCGGT
Sample Output:
4 3 2 1
Note: your source file is to be called: "baseCounter.py"

PART 2 Task 3: DNA Complement

DNA's bases are paired as follows: A-T and G-C.

You are to write a Python program that inputs a DNA sequence as a String input and returns the complementary strand.

Given: A DNA string s of length at most 1000 bp.
Return: The reverse complement of s.

Sample Input:

AAAACCCGGT
Sample Output:
ACCGGGTTTT
Note: your source file is to be called: "revComp.py"

Required Deliverables

All of the deliverables specified below should be placed into a new folder named 'lab02' in your Bitbucket repository (`cs300f2017-bbill`) and shared with the instructor by correctly using appropriate Git commands, such as `git add`, `git commit -m ''your message''` and `git push` to send your documents to the Bitbucket's server. When you have finished, please ensure that you have sent your files correctly to the Bitbucket Web site by checking the `source` files. This will show you your recently pushed files on their web site. Please ask questions, if necessary.

- 'Lab02Part1_DNAReview' document with your answers to the outlined questions.
- Two Python programs "baseCounter.py" and "revComp.py" for PART 2 tasks 2 and 3. Your source code should be able to run without any effort on behalf of the instructor.

You should see the instructor if you have questions about assignment submission.