

Homework 2

Instructions: This homework is due at the beginning of lab, Wednesday, September 13th. You can edit this file and copy/paste into it by opening it in Libreoffice on Linux, PDFfiller Google App, Acrobat, MS Word, and others. Bioinformaticians are especially agile at using the internet to gather and share information, so if you do not remember or see the answer in class materials, feel free to use the web, but do not plagiarize it. Also, like anything you can obtain from the internet, information about bioinformatics may be wrong or only partially correct. Be sure to check the source of the information to help judge its quality. This particular homework will require you to write three Python scripts. I would like you to submit a digital copy of these scripts to me, along with your answers to questions 1 and 2 and output files for questions 3 and 4, all compressed into one .zip file. Make sure to run each script to ensure they work. I also encourage you to write and test each script one line at a time and test every variable you define.

1. Write a Python script that calculates the volume of a cylinder given a height of 6.1cm and a radius of 3.2cm. What is the volume?
2. Quick questions about strings and lists:
 - a. What does mutable mean?
 - b. Are strings mutable?
 - c. Are lists mutable?
 - d. Which data types can be stored in a list?
 - e. How does the list sort function sort lists of lists?
3. Write a Python script that imports the fasta file, "hw2seqs.fasta", found in the "/ptmp/bcbio444/hw2" folder in hpc-class, and translates all sequences starting with the first "AUG" in the sequence and stopping at the first STOP codon. Put the translated sequences in a new fasta file, preserving the original names.
4. Write a Python script to import the file "atv10_asm.fa", found in the "/ptmp/bcbio444/hw2" folder in hpc-class, a real fasta file for the Arabidopsis thaliana genome, and calculate the GC content of each chromosome in the file. Output the info into a tab separated 2 column file where the first column is the chromosome and the second is GC content. You must use dynamic inputs and outputs for this script (no hardcoded file names). Also, you must calculate GC content using a function.