

Homework 2

Practicing Python

You will accept the homework problem invitation through GitHub Classroom. The link is in canvas.

This will create a hw2-python-practice-yourname repository. You will clone this repository to your system or open in visual studio

```
$ git clone git@github.com:biodataprogram/hw2-python-practice-YOURNAME.git
```

I have created starter scripts for you to update and add the necessary code to solve the problem.

0. Write a UNIX script which has the job of setting up the data. It will be called `setup.sh`
 - a. Have it download the **Thermus aquaticus** genome from NCBI, available at this url: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/399/775/GCF_001399775.1_ASM139977v1/GCF_001399775.1_ASM139977v1_genomic.fna.gz
 - b. Have it download the **Bacillus subtilis** genome at this url: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/009/045/GCF_000009045.1_ASM904v1/GCF_000009045.1_ASM904v1_genomic.fna.gz
 - c. uncompress the files
1. Write a python script called 'problem1.py'
 - a. Have your python script calculate the percent GC (number of G and C bases in the sequence file divided by the total length of the sequence) content of the two files downloaded above. Print out the GC percentage for each of the files.

Bonus - make the script more generic to print out filename and GC content for any set of files passed in on the command line

2. Write a script called 'problem2.py' reads in file sequences.txt.
 - a. It has several columns, have it print out a new report to STDOUT with three columns: GeneID, gene length (gDNA column), mRNA length

Bonus: print the data sorted so the longest gene (gDNA) comes first and smallest is last.

3. Write a script called 'problem3.py' reads in file sequences.txt.
 - a. Generate a summary table of statistics about gene length - what is the average length, longest and shortest genes
 - b. have the script create file that contains a summary of gene sizes binned by 500bp. Say there were 4 genes in the file of lengths 1100, 750, 1800, 950. You print out the following to a file called 'lengths_binned.csv'

```
bin_size,count
0,0
500,2
1000,1
1500,1
```