# CS/BIOL 123A
# Programming Programing Assignment 2:

## INTRODUCTION:

In this programming assignment you will learn how to download Accession # information and FASTA file DNA sequence data from the NCBI DB. You will then analyze the GC content of each sequence and predict the species from which the DNA sequence was extracted.

Different species have different GC count percentages. A GC count percentage is defined as

$$GC\ percentage = \frac{G\ count + C\ count}{G\ count + C\ count + A\ count + T\ count} \times 100\%$$

For example, the table below provides the range of GC percentages for the listed organisms.

| Species | Range of GC Percentage |
|---|---|
| Human | 36% - 60% |
| House Mouse | 42% - 66% |
| Norway Rat | 51% - 58% |
| Dog | 17% - 45% |
| Chimpanzee | 35% - 45% |

From the above table, you will notice that all of the GC% ranges overlap to varying degrees. So how can one predict an organism for a calculated GC%? One way is to take into account where, within the GC% rage for an organism, the calculated GC%

resides. For example, a GC% = 40% falls in the middle/median of the range for chimpanzee, and close to the upper limit of Dog, outside the range for Norway Rat, outside the range for House mouse, and near the lower limit for human. So, one might rank the likelihood of a prediction in descending order of likelihood as

1. Chimpanzee
2. Dog, Human

based on the distance from the median value of each range. The remaining organisms are not listed because the GC 40% value lies outside their respective ranges. So, an interpretation might be that the analyzed DNA sequence is more likely from a Chimpanzee than any other listed organism, and the next likely possibilities include Dog or Human.

## INSTRUCTIONS (CS, MSBI, BI MINORS, and SE MAJORS):

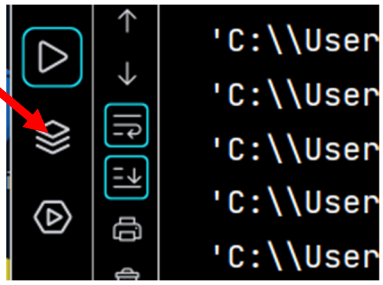
1. Download the files named
    a. cs123a_prog_assignment_2.py
    b. cs123a_list_of_acc_nums.txt
    c. cs123a_list_of_species_gc_count.txt

2. The cs123a_prog_assignment_2.py file provides an example of how to get the FASTA file for a given accession number. A FASTA file contains a DNA, RNA, or protein sequence for an associated accession number. In the example cs123a_prog_assignment_2.py code, line 112 prints the FASTA file DNA sequence for the accession number "NC_000007.14", see line 81.

3. Make sure your Python environment has the following packages installed:
    a. copy
    b. Bio
    c. xmltodict

    Use the `pip3 install <package name>` to install each package if not already installed. If you are using the PyCharm IDE, then simply click the Packages option and enter the package name to install

Select PyCharm package option.




Enter package name and click install.


4. Modify the code to read in the accession numbers in the "cs123a_list_of_acc_nums.txt" file.

5. Modify the code to read in the "cs123a_list_of_species_gc_count.txt" file. Each line in the file contains the species name, low GC% and high GC%. Create a table or dictionary, or, …, whatever you like to hold that information.

6. For each accession number that was read in, download the associated FASTA file sequence. NOTE: Each accession number might have multiple DB ID numbers. (See lines 90 to 100). Each ID is a separate FASTA sequence for a single accession number. <- Do not worry about why that is the case. For now, just realize that for a single accession number, there might be multiple ID numbers, and each ID number has an associated FASTA sequence.

7. Compute the GC% for the retrieved FASTA sequence and associated IDs.

8. From the calculated GC%, use the table/dictionary, …etc. that you created from the "cs123a_list_of_species_gc_count.txt" file to predict and rank the likely organism(s) similar to the example provided in the INTRODUCTION section.

9. Print your findings as follows

ACC: <Acc number>
ID: <ID #>
RANKING: <most likely organisms>,
<next most likely organism>,
… ,
<least likely organism> <- NOTE: *enclose in "( … )" rankings of organisms at the same level as in the "2. Dog, Human" in the example above. One would print the following*

RANKING: Chimpanzee,
(Dog, Human)

10. Submit your code to “Programming Assignment 1” on Canvas by the specified due date.

## DONE FOR PROGRAMMER TYPES

## NON-PROGRAMMER TYPES SEE NEXT PAGE

## INSTRUCTIONS (BIOLOGY, BIOCHEM, CHEMISTRY, and <NON-PROGRAMMER TYPE> MAJORS):

These instructions assume you are familiar with or are comfortable with using Excel spreadsheets and know how to find or insert formulas into cells that perform the desired calculation(s). If not, let the instructor know and links to YouTube videos will be provided.

1. Download the files named
    a. cs123a_list_of_acc_nums.txt
    b. cs123a_list_of_species_gc_count.txt

2. Your task is to use an Excel spreadsheet to help analyze the GC% in the DNA sequences associated with each of the ACC numbers in the cs123a_list_of_acc_nums.txt file. Then you will compare the GC% that you calculate with the GC% ranges listed in the cs123a_list_of_species_gc_count.txt file.

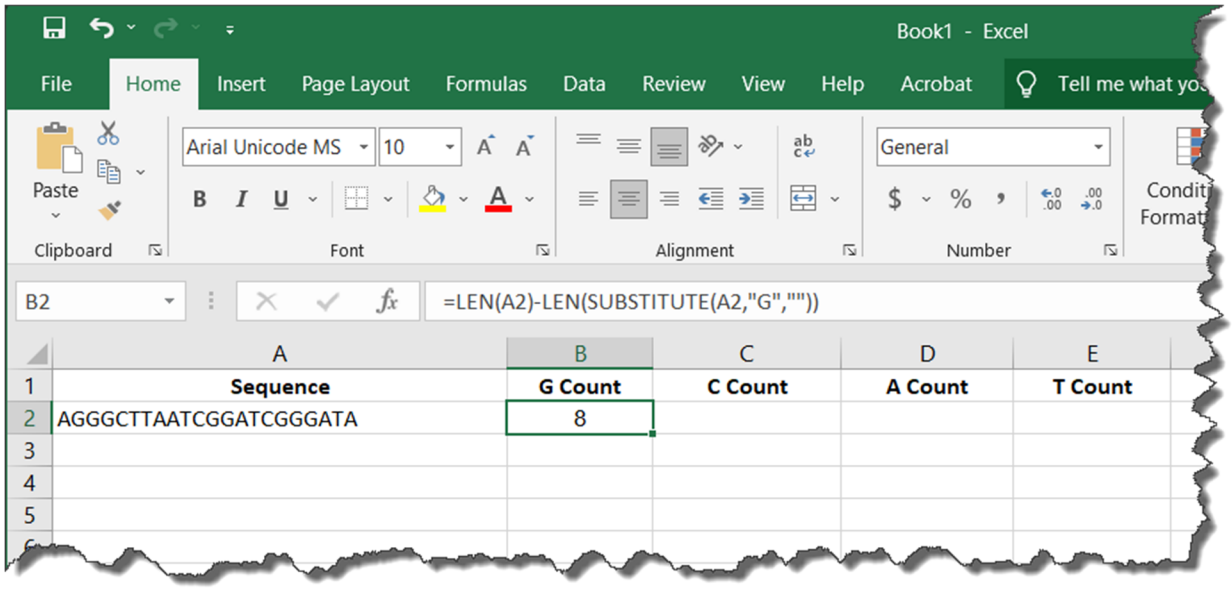
3. Each line in the cs123a_list_of_species_gc_count.txt file contains the species name, low GC% and high GC%.

4. For each NC_ accession number, retrieve and save to an Excel spreadsheet its corresponding nucleotide sequence from the NCBI DB. You may organize the sequences in columns or rows. It is up to you. It will not make a difference with respect to what you will need to do to determine the GC% or how you compute GC%.

5. Select cells in the spreadsheet that will contain the GC% for each sequence as defined below.

$$GC\ percentage = \frac{G\ count + C\ count}{G\ count + C\ count + A\ count + T\ count} \times 100\%$$

6. You will need to find out the Excel formula that you will need to insert a count of a nucleotide, e.g., G or C into a cell. For example, the illustration below shows how to count the number of "G"s in the sequence in cell A2.




7. Compute the GC% for each of the retrieved FASTA sequence and associated IDs.

8. On a separate tab in your spreadsheet, insert the organism and GC% range information that is contained in the file named "cs123a_list_of_species_gc_count.txt" You can organize this info however you like.

9. Now go back to the first tab with the sequences that you retrieved from the NCBI DB. On that tab, select the cells that you will display the ranking information as indicated below.

ACC: <Acc number>
RANKING: <most likely organisms>,
<next most likely organism>,
… ,
<least likely organism> <- NOTE: *enclose in "( … )" rankings of organisms at the same level as in the "2. Dog, Human" in the example above. One would print the following*

RANKING: Chimpanzee,
(Dog, Human)

This will require you to do a bit of "Excel" research to find or construct the formula(s) to determine the rankings. It might also require you to find out how to reference the data in a cell that is located on a different tab.

NOTE: If one were to change the GC% range information located on the second tab that you created, the ranking results should change or be updated as appropriate.

10. Submit your spreadsheet to the Programming Assignment 1 on Canvas.

# DONE FOR NON-PROGRAMMER TYPES