Introduction to Python for Bioinformatics
BINF 5007
Final Project

Instructions:

1. ALL DUE TIMES ARE IN EST
2. Upload required files to the BB Assignment.
3. All answers must be in your own words, and copy-and-paste answers will receive no credit.
4. You must submit **.py files, grangers_analysis.csv** and **images of the plots requested**
5. You are limited to 2 submissions
6. There is no extension for the final project.

Q1. Dr. Granger is interested in studying the relationship between the length of house-elves' ears and aspects of their DNA. This research is part of a larger project attempting to understand why house-elves possess such powerful magic. She has obtained DNA samples and ear measurements from a small group of house-elves to conduct a preliminary analysis (prior to submitting a grant application to the Ministry of Magic) and she would like you to conduct the analysis for her (she might know everything there is to know about magic, but she sure doesn't know much about computers). Please download the elves_data file.

Write a Python script that:

1. Imports the data into a data structure of your choice
2. Loops over the rows in the dataset
3. For each row in the dataset checks to see if the ear length is large (>10 cm) or small (<=10 cm) and determines the GC-content of the DNA sequence (i.e., the percentage of bases that are either G or C)
4. Stores this information in a table where the first column has the ID for the individual, the second column contains the string 'large' or the string 'small' depending on the size of the individuals ears, and the third column contains the GC content of the DNA sequence.
5. Prints the average GC-content for both large-eared elves and small-eared elves to the screen.
6. Exports the table of individual level GC values to a CSV (comma delimited text) file titled `grangers_analysis.csv`.

Q2. It makes sense that larger organisms have larger offspring, but what the mathematical form of this relationship should be is unclear. Let's look at the problem empirically for mammals.

Import the data from the file "Mammal_lifehistories_v2" into a Pandas data frame. There are some extra blank lines at the end of this file, so get rid of them by using the optional read_csv() argument `skip_footer=7`.

Missing data in this file is specified by -999 and -999.00. Tell Pandas that these are null values using the optional read_csv() argument `na_values=['-999', '-999.00']`. This will stop them from being plotted.

1. Graph adult mass vs. newborn mass. Label the axes with clearer labels than the column names.
2. Graph the log (base 10) of adult mass vs. the log (base 10) of newborn mass. Label the axes.
3. For data where order is Rodentia, graph the log (base 10) of adult mass vs. the log (base 10) of newborn mass. Label the axes.
4. This looks like a pretty regular pattern, so you wonder if it varies among different groups. Graph adult mass vs. newborn mass, with both axes scaled logarithmically, and the data points colored by order. Label the axes.
5. Coloring the points was useful, but there are a lot of points and it's kind of hard to see what's going on with all of the orders. Create a subplot for each order.

MetaData: Ecological Archives E084-093-metadata