BIOI 3500, *Advanced Bioinformatics Programming*
Assignment 10

Assignment 10 :: 50 points :: Due: 11:59 p.m. Tuesday, April 20th, 2021

Objectives

This exercise will give you the opportunity to practice parsing and manipulating the content of FASTA files using Biopython.

Description

Download from Canvas the file **`human.1.protein.faa.gz`** available under Files/Assignments/Assignment10.

Using Biopython write a (procedural) Python (version 3) program, **`netID_Assign10.py`** where **`netID`** is your UNO NetID, that computes and displays the frequency distribution of the amino acids from all the sequences in a given file, and that reports the sequence id of the sequence with the lowest and highest amino acid content for each particular amino acid. Note that the Biopython **`SeqIO`** format name for FASTA files is “fasta”. Output should look exactly like the following (except for the actual numbers and ids that are made up), for all 20 amino acids:

```
A    6.86%    NP_001123300.1  NP_001123300.1
C    2.17%    NP_00623300.1   NP_0013340.1
…
W    1.18%    NP_0312550.1    NP_0543300.1
Y    2.54%    NP_0034545.1    NP_0015455.1
```

Here, the first column is the single-letter abbreviation of the amino acid, the second column contains the frequency of the amino acid in the **`human.1.protein.faa`** file, the third column contains the sequence id of the protein that has the lowest content of that amino acid and the last column contains the sequence id of the protein with the highest content of that amino acid. Columns should be tab-separated and amino acid should be listed alphabetically.

Amino acid content per protein is simply computed by counting the frequency of an amino acid in the protein divided by the length of the protein.

For example, a sequence like this:

```
SDFSSAATCW
```

would have an S content equal to 0.3, a D content equal to 0.1, and so on. If there are ties, simply return the sequence that is first alphabetically (for example, if **`NP_00623300.1`** and **`NP_001123300.1`** are a tie, **`NP_001123300.1`** should be reported).

Results should be written to file **`netID_results.txt`** where **`netID`** is your UNO NetID.

You do not have to use command-line options for this assignment—all filenames can be “hard-coded” into your program.