Bioinformatics Assignment 2: Displaying Data from Files in Python

By Noah Segal-Gould

Background: In bioinformatics, programmatically reading, parsing, and analyzing sequence data from inside FASTA and PDB files give scientists easier access to useful statistics regarding proteins and nucleotides.

Methods: The majority of what my program does is written inside the `main` function, which takes the names of the three input files as inputs and returns no outputs. Through use of the `input` function it takes user input from inside a command line interface window and performs useful calculations which are printed for the user to view. Initially, it loads the FASTA protein data, the FASTA nucleotide data, and the PDB file data as strings, then a regular expression is used to find the name of the protein inside the FASTA protein file and it is printed in the command line interface window. Through use of a `while` loop the function performs different tasks on the three datasets depending on if the user inputs `1`, `2`, or `3` into the command line interface window. In the first case, regular expressions are used to count the total number of each nucleotide present in the FASTA nucleotide file data, and then those statistics as well as the percent frequency of each nucleotide in that data are printed to the window. In the second case, the user is prompted to enter a motif for the FASTA protein data and regular expressions are used to find the index positions of that sequence, which are then printed out. Finally, In the third case, the centroid finding program from class and Moodle 2 was adapted to print centroid data from the PDB file. If the user fails to select one of these three options, the program halts.

Results: The program takes user input and performs tasks based on it. In this example, the user selects option `1`, then option `2`, then within option `2` inputs the motif "`LAV`" and finally inputs option `3`. After that, the user inputs something other than those three options and the program halts. This appears in the command line interface window as follows:

```
Protein Name: cytochrome b

Please enter one of the following options:

1: Nucleotide data:
      This option will output the number of each nucleotide in the sequence and
the frequency of each nucleotide.
2: Protein data:
      This option will allow you to input a specific motif to find, and display
the location of each motif in the data.
3: Centroid data:
      This option will output the centroid location of the protein.

==================== WARNING ====================
If you fail to input one of these options, the program will halt.

Please enter either 1, 2, or 3: 1
A: 1839 (20%), T: 3250 (35%), C: 2050 (22%), G: 2122 (23%)
Total number of unidentified nucleotides: 0

Please enter either 1, 2, or 3: 2
Please enter a motif: LAV
The following locations were found for that motif ("LAV"):
      573, 576, 1563, 2126, 2318, 2525

Please enter either 1, 2, or 3: 3
Centroid size: 37804
Centroid position: 23.065753, -0.07208, 22.827757

Please enter either 1, 2, or 3: exit
Invalid input; this program will now halt.
```

Conclusion: I completed tasks 1 through 5 as assigned in the Lab Report 2 Assignment file. My program's output is as I would expect, and it makes use of regular expressions where the directions specify they should be used.

Acknowledgements: I worked alone on this assignment and made use of code shown in class as well as provided on the course Moodle 2.