Python programming for big data and the scientist

PhD course in Uppsala

13th of October 2016

Guidelines

The answers to these exercises are to be written in Python v2.7. Due to the time limits of the teachers, there will be no hand-ins or correction. The answers however will be heavily detailed and published on Monday, October 17th. We hope you have fun solving the problems !

0.1 Credits

This exercise was originally designed by Aline Dousse and Jia Li while working in Zdobnov's group. It was then adapted by Charles E. Vejnar and further developed by Lucas Sinclair while working at the Bioinformatics Core facility at the EPFL and in Alexander Eiler's group in Uppsala.

1 Simple calculations

You have a fasta file (`seq.fas`) containing protein sequences and their headers. Write a python script that reads this file and calculates the mass of each protein based on the amino-acid mass below. The output is written into a file with the header for each sequence, and a line with the protein mass.

```
G = 57.021464, A = 71.037114, S = 87.032029, P = 97.052764, V = 99.068414,
T = 101.04768, C = 103.00919, L = 113.08406, I = 113.08406, N = 114.04293,
D = 115.02694, Q = 128.05858, K = 128.09496, E = 129.04259, M = 131.04048,
H = 137.05891, F = 147.06841, R = 156.10111, Y = 163.06333, W = 186.07931
```

Note that headers begin with "$>$" in fasta formatted file.

2 Short functions and libraries

Write the following functions for integer lists:

- `coun_odd` to get the number of odd numbers in the given list,
- `get_even` to get the list of all even numbers in the given list,
- `arith_mean` which returns the arithmetic mean of values given in the list,
- `median` that return the median of values given in the list.

Use the above functions to display the number of odd numbers, the list of even numbers, the arithmetic mean, and the median of the numerical list entered by the user.

Bonus: Can you solve the problem with `numpy` functions ?