Project no. 2 : CpG islands and DNA Methylation analysis

deadline: 24.11 11:59 pm

Input data:

1. CpG islands:
   http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/cpgIslandExt.txt.gz
2. DNA Methylation:
   http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibMethyl450/wgEncodeHaibMethyl450A549Etoh02SitesRep1.bed.gz
3. Chromosomes sizes:
   http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.chrom.sizes

Definition:

Shores: CpG islands borders +- 2kb
Shelves: Shores borders +- 2kb
Seas: Between shelves

Tasks:

1. Prepare files with coordinates of CpG islands, Shores, Shelves and Seas in BED format *(chrN start end)*. Remember to consider chromosome boundaries. Please use only **autosomal chromosomes**
2. Set DNA Methylation coordinates as the middle of its range
3. Find how many DNA Methylations are located in CpG islands, Shores, Shelves and Seas and show results using chart with information about percentage and number of methylations in regions (bar plot, pie... - you can use matplotlib, seaborn, bokeh..)

Methods:

1. Project should be prepared using jupyter-notebook (or similar tool).

Points:

max points for this project: 6p.

- 4p. for task_1
- 2p. for task_3

max. points after deadline: 3p.

my emai: m.wlasnowolski@mini.pw.edu.pl

Please put your script into repository (bitbucket/github etc) and share it with me till deadline.