BBS741 Project #1

Deadline: October 15, 2018 @ 11:59 pm
Email to BBS741.2018@gmail.com
with LastName_FirstName_Project1 as subject line

Part 0 (0 points)

Download and install anaconda then run the script: Test-Package-Load.py. If all packages are installed correctly should not see any warnings (aside from Matplotlib building font cache).

Introductory presentation

Part I - Clustering genes using K-means clustering (25 points)

In Enhancer-Brain-Matrix.txt we have H3K27ac signal z-scores for over 17k enhancers across eight embryonic time points in forebrain.

Building off the script Kmeans.py, use K-means clustering to divide the enhancers into four groups. The script should output a matrix with a column for the enhancer ID, cluster ID, and H3K27ac signal for each time point (see below)

Enhancer	*cluster*	*e10.5*	*e11.5*	*e12.5*	*…..*
…..	1	….	….	….	….

For each of the four clusters, make a heatmap of the H3K27ac signal across the eight timepoints (similar to heatmap on slide 31 of presentation) You may use Python or R for this task and may beautify with inkscape/illustrator, but you must include your code for plotting.

What does the pattern of enhancer activity suggest about each cluster?

Part II - Clustering biosamples using Hierarchical clustering (25 points)

In Enhancer-Mouse-Matrix.txt we have H3K27ac signal z-scores for 100k enhancers across eight embryonic time points in 12 tissues.

Building off the script HClustering.py, use hierarchical clustering to cluster the biosamples into four groups. The script will output a png file "dendrogram.png".

Which biosamples cluster together? Are there any outliers?

Part III- Clustering biosamples using PCA (50 points)

Once again we will use Enhancer-Mouse-Matrix.txt which has H3K27ac signal z-scores for 100k enhancers across eight embryonic time points in 12 tissues.

Building off the script PCA.py, use PCA for dimensionality reduction across gene expression of the 72 biosample-timepoints. The script should take in Enhancer-Mouse-Matrix.txt as input and should output two results:

1. A file with the variance explained by the five components

   PC1 ….
   PC2 ….
   PC3 ….
   PC4 ….
   PC5 ….

2. Matrix for each biosample with the five principal components

	PC1	PC2	PC3	PC4	PC5
Biosample-1	…	…	…	…	...
Biosample-2	…	…	…	…	...
Biosample-3	…	…	…	…	...

   …..

Then using either Python or R, create the following two scatterplots:

1) PC1 vs PC2 for all 72 tissues with dots colored by tissue type (forebrain, stomach etc)
2) PC1 vs PC2 for all 72 tissues with dots colored by time point (e.g. e10.5, e11.5 etc)

Which biosamples cluster together? Do the clusters tend to separate primarily by time point or tissue?