BIO-3027 Python Assignment - 2024

GMS Nimantha Viraj Senavirathne

The assignment folder can be found in:

GitHub - AllMouthyKing/BIO-3027-Python-Assignment: Assignment/Exam project for BIO-3027

Do be warned that delays between requesting information from the Entrez database has been implemented to lessen traffic, hence running the code **can take up 3 minutes** if on a bad day. So do give it time to finish. Can observe which entrez id is being searched up on the console. The end of the pipeline occurs when the results (png/text files) are opened automatically for viewing.

For this assignment, a crude data analysis and visualisation pipeline was created to observe the results of a high throughput RNA-sequence experiment carried out on frozen pancreatic/liver/skeletal tissue samples from Diversity Outbred mice. The experimental results were obtained from the following sources:

Title: RNA-seq expression analysis of liver, adipose, skeletal muscle, heart, and pancreatic islets from DO founder strain mice fed high-fat, high-sugar diet.

Pancreatic: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE266921

Liver: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE266923

Skeletal: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE266922

Made public on May 13, 2024, with contributors, Churchill GA, Baker CN of The Jackson Laboratory.

The data files were in csv format, and two files per each organ type was downloaded; the mrna_annot and vst_mrna.csv files as I deemed them more important than the raw count data.

The annotation file contains the entrez id which was vital for the pipeline while variance stabilizing transformation (vst) count matrix whose data has been transformed so that the variance is roughly constant across different mean values, allowing for more reliable comparisons across genes and samples.

Basic pre-processing is carried out on the data, then heatmaps were produced for 20 random genes, their average expression and for heatmaps for the 10 highest and lowest expressed genes for each organ type.

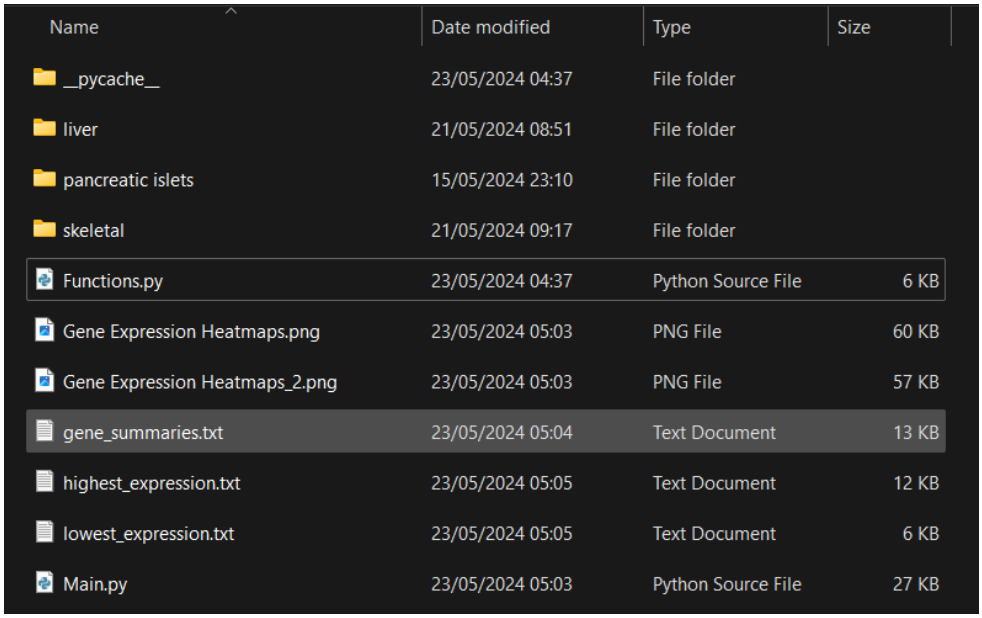
The entrez_id was obtained from the above listed genes then their gene summary was obtained from the entrez database.

The pipline consists of two python files (Main.py/Functions.py) and five results (two heatmap.png and three text files containing gene summary information obtained using the entrez database). To prevent congestion, each time a new organ type is selected from the basic

gui implemented in the code, the result files (png/text) would be overwritten with the new data. Hence it is advised to save a copy of the results in a different location if saving each result is a must.

The main code is present in the Main.py. Running it will open a basic GUI containing a dropdown menu to choose which dataset you want to analyse before clicking the Run Script button.

The file structure consists of the main folder (Python_Assign) containing the five results above, the two .py files, and three folders illustrated as ‘skeletal’, ‘liver’, and ‘pancreatic islet’ containing the raw data. This order should be adhered when downloading to prevent any errors in file reading:



I am unsure if the relative path files using os.path.join(folder, .csv) I used in the Main.py would work, so please do contact me if any issues occur:

nimantha.senavirathne@gmail.com

gmsen0083@uit.no

The versions of the used modules are given below:

Pandas version: 1.5.3

Seaborn version: 0.12.2

Biopython version: 1.83

PIL version: 10.0.1

Biopython version: 1.83