Processing of Biological Data

Prof. Dr. Volkhard Helms
Tutor: Trang Do
Winter Semester 2021

Saarland University
Chair for Computational Biology

Exercise Sheet 5

Due: 25.01.2022 10:15 am

Submission

- You are advised to work in groups of two people.
- Submit your solution/questions by email to trangdht.bioinfo@gmail.com as a single PDF with the name format **PBDA5_Lastname1_Lastname2**. The PDF should contain your answers AND the formatted source code. Additionally, submit your source code files packed in a single .zip archive, NOT as individual files.
- Late submissions will not be considered.
- Do not forget to mention your names and matriculation numbers in the PDF file.

Note that this assignment sheet is counted as 100 points just like the four other assignments. But you can actually earn 150 points from the two problems of this assignment sheet. So, 50 points are bonus points. In total, students need to earn at least 250 points from the five assignments to be admitted to the final exam.

Exercise 5.1: Differential Analysis of Multi-omics Data (75 pts)

The NCBI Gene Expression Omnibus (GEO) is an ample source of raw and processed biodata. For this exercise, you will perform differential analysis on gene expression and methylation data collected from normal immortalized keratinocytes from skin affected by human papillomavirus (HPV). Each data type includes three samples of four categories: normal, HPV-16 positive, HPV-18 positive, and HPV16-positive with E7-deficient tissues.
Differential analysis of expression can be performed with ease using many available packages in **R language** in a few simple steps:

(a) Download the *processed* dataset for gene expression from GEO series GSE83259 and for methylation from series GSE83261. Report the type of tissues and accession numbers for each file (sample) in the datasets. Retrieve the gene expression and methylation beta values for differential analysis in later steps. (10 pts)
Hint: Inspect and apply the functions **getGEO(), pData(), featureData(), exprs()** from R package **GEOquery** for this task.

(b) For each dataset, plot the samples w.r.t the first two principle components and inspect the clustering of these samples. (10 pts)
Hint: Function **prcomp()**.

(c) Perform differential analysis:

(i) *Gene expression*: Identify the Diffentially Expressed Genes (DEGs) between all possible pairs of tissue types with R package **limma**. DEGs should be selected with FDR-corrected p-values smaller than 0.05. Report the number of significantly up-/down-regulated probes in all pairwise comparisons and produce a volcano plot summarizing the DEGs. (15 pts)

(ii) *Methylation*: Identify the Diffentially Methylated CpG (DMCs) between all possible pairs of tissue types with R package **limma**. DMCs should be selected with FDR-possible p-values smaller than 0.05 and absolute Log Fold Change of at least 0.2. Report the number of hypo-/hyper-methylated CpGs in all pairwise comparisons and produce a volcano plot summarizing the DMCs. (15 pts)

Hint:
limma tutorial: https://www.bioconductor.org/packages/devel/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html (Part 6)
Conversion from probe IDs/CpGs to gene symbols can be conveniently done by looking up the information from **featureData()** in R.

(d) Identify and report the sets of genes that are (i) *hypomethylated and upregulated* or (ii) *hypermethylated and downregulated* between normal tissue and HPV-16 positive tissue. Perform Gene Ontology Analysis on all the found genes with FDR-corrected p-value smaller than 0.05 and return the list of top 10 GO terms for Biological Process and KEGG pathways using any tool or package (i.e. DAVID, Panther, ShinyGO,...). (15 pts)

(e) Briefly discuss the tool used for differential analysis of expression and methylation data. (10 pts)

Exercise 5.2: Multi-omics data integration by Similarity Network Fusion (75 pts)

In the second part of the assignment, you will implement in **Python** the Similarity Network Fusion (SNF) to integrate the normalized log-base-2 of the somatic copy-number alteration (SCNA) and methylation beta values of cholangiocarcinoma tumors retrieved from The Cancer Genome Atlas (TCGA). These datasets are to be found as SCNA.csv and MET.csv, respectively, in the supplementary.

Part I. Data preparation. For each given dataset:

(a) Distance Matrix (DM): Implement **get_DistanceMatrix()** method in Tools.py to compute the pairwise Euclidean distance across all columns. (10 pts)

(b) Affinity Matrix (AM): Compute an affinity matrix (similarity matrix) from with KNN nearest neighbors using **get_AffinityMatrix()** method in Tools.py. Inspect the provided function and describe how the similarity kernel used for constructing the affinity matrix was formed. What is the importance of using a kernel matrix in this step? (15 pts)

Part II. SNF Implementation

(a) Initial Transition Matrix ($S0$): Implement **get_InitialTransitionMatrix()** method in Tools.py to compute an initial transition matrix from an affinity matrix by retaining only K largest values for each row. All other values should take zero. Finally, return the rowwise normalized matrix. (10 pts)

(b) Implement **perform_SimilarityNetworkFusion()** method in Tools.py by updating the set of Transition Matrix $S = \{S0_1, S0_2, S0_i, ..., S0_n\}$ and the set of Affinity Matrices $A = \{AM_1, AM_2, AM_i, ..., AM_n\}$ for t iterations. (15 pts)

$$\begin{cases} S_i(t+1) = S_i(t) \times \delta_i(t) \times S_i(t)^T, \text{ where } \delta_i(t) = \frac{\sum_{j,j\neq i} \sum_k A_{jk}}{|A|} \\ A_i(t+1) = S_i(t+1) + I_{A_i}, \text{ where } I \text{ is identity matrix of } A_i \end{cases}$$

(c) Implement **get_FusedAffinityMatrix()** method in Tools.py to compute the Fused Affinity Matrix FAM as the average of all updated matrices in A. (10 pts)

$$FAM_A = \frac{\sum_i \sum_j A_{ij}}{|A|}$$

(d) The **SNF()** method in Tools.py summarizes SNF by combining **get_InitialTransitionMatrix()**, **perform_SimilarityNetworkFusion()** and **get_FusedAffinityMatrix()** with default values $K = 5$ and $t = 200$. Meanwhile, **plot_SNF()** performs spectral clustering on any affinity matrix and plot the heatmap representing $n\_clusters = 5$ clusters.
Use **plot_SNF()** to inspect the computed Fused Affinity Matrix from **SNF()** and all other Affinity Matrices in A. Briefly compare and discuss the outcomes. How does the result change when we use different values of t and K? (15 pts)