Introduction to Biomedical Informatics by Dr. Junhee Seok
Term Paper Assignment #3

Gene Expression Inference Analysis

This assignment is to analyze the gene expression profiles of colon cancer patients. You can find the expression data from the Gene Expression Omnibus database (GSE39582). The data was generated for the research described in Marisa *et al*, “Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value”, PLoS Medicine, 2013.

You can download the “series matrix file” to easily obtain the pre-calculated gene expression indices and clinical information. In the series matrix file, “!Sample_characteristics_ch1” fields contain the clinical information. The gene expression was measured using Affymetrix U133+2 arrays. The series matrix file contains the expression indices for probe sets of the array. To map the probe sets to genes, the array annotation should be used.

If you use Python, you may want to use `scipy.stats` module for various statistical tests. If you use R, you can easily find such functions. Please google them.

For this data, please answer the following questions.

1. (Data description) Describe about the sample size, the number of probe sets, the number of clinical variables. How many unique genes are there? Why are certain genes measured by multiple probe sets?

2. (Significant genes for sex) What are the sizes of men and women samples? By applying two-class t-tests, please find the significantly associated genes with sex for p-value < 0.001. How many probe sets and unique genes can you find? What are the top 10 unique genes? Which chromosomes do the top 10 significant genes locate? For the top 200 genes in the order of statistical significance, plot a heatmap. Please cluster genes and samples using hierarchal clustering.

3. (Significant genes for KRAS mutation) Please find the significantly associated probe sets (as well as unique genes) with KRAS mutation using two-class t-tests. Please use the p-value criteria 0.001. Report the top 10 unique genes. Please repeat the analysis with permutation tests for the same p-value criteria. How much are two analysis results different or similar? (Hint: NA means “not available”. Such samples should be excluded)

Introduction to Biomedical Informatics by Dr. Junhee Seok
Term Paper Assignment #3

4. (Regression for tumor stage) Please find the significantly associated probe sets (as well as unique genes) with TMN tumor stage using linear regression. Consider samples with stage 0 as samples with stage 1. How many significant genes can you find? Report the top 10 unique genes.

Instead of continuous TMN tumor stages, we will consider two stages. Stage 0, 1, and 2 are “benign” stage, and stage 3 and 4 are “malignant” stage. Please find significant probe sets (as well as unique genes) between benign and malignant stages, using t-tests with p-value < 0.001. Report the top 10 unique genes.

By comparing the results of the regression and two-class comparison analyses, how much are they different or similar to each other?

5. (Biological contexts for tumor stage) We will use the unique genes from the results of the two-class comparison analysis in Problem 4. Let’s call these genes as tumor genes. In this problem, we will investigate the biological functions enriched by the tumor genes. The biological context can be described as gene sets. We will use Gene Ontology (GO) gene sets, you can download from MSigDB (C5). If you need to login, please register you email address or use jseok@stanford.edu. The gene sets are in a GMT file, which is a simple text format and contains the list of genes for a gene set per each line. Among GO gene sets, please find the significantly enriched by the tumor genes by Fisher’s exact tests (or chi-square tests). How many are significant gene sets (p-value < 0.001)? What are the top 10 gene sets? Can you interpret any of the significant gene sets in a biological sense?

If you have any question, do not hesitate to ask me!!!