Dr Ben Galili
Alon Oring
Dr Leon Anavy
Prof Zohar Yakhini

WIS/IDC - Project in statistics and data analysis.

Differential Gene Expression in Acute Myocardial Infraction and in one more dataset

1. Introduction

Gene expression describes the process in which genes that are coded in the DNA of living organisms are transcribed into mRNA. This is part of the bigger process in which genes are being copied (transcribed), processed, translated and modified into the final product, usually a protein. Gene expression profiling measures the levels at which mRNA molecules pertaining to the genes profiled are observed in a sample.

In this exercise, we will perform guided analysis, comparing expression profiles of circulating endothelial cells (CECs) in patients who experienced acute myocardial infraction to CECs in healthy controls. A comparison of two sample classes. You will then select one more gene expression dataset and perform your own analysis there.

2. The Data Set

The data set was taken from:

1) Dataset record in NCBI: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66360
2) Published paper: Muse et al, Sci Rep 2017 https://www.nature.com/articles/s41598-017-12166-0

We extracted the data matrix and provide it as a separate csv attachment (link to download). The csv file needs to be pre-processed before moving to the main analysis steps. Some information should be removed but make sure that you keep all information that is important for the analysis. Specifically, all expression values should be kept and the label of each sample (H – Healthy, M - Myocardial Infraction).

The paper describes a study that seeks to develop an expression-based signature that can detect AMI in patients in a non-invasive manner, by profiling CECs.

3. Analysis

a. High level description of the data and some pre-processing

1) How many genes profiled?
2) How many samples (subjects/patients) in total?
3) How many samples in each class?
4) If there are missing values, then remove the entire row (gene) from the data matrix.
   How many rows left now?
5) Pick 20 genes at random. Draw histograms comparing expression levels of each of these genes in the two classes M and H.

b. WRS for differential expression (DE)

1) Consider some gene, g. Under the null model (which assumes that for g there is no M vs H DE), what is the expected sum of ranks of g's expression levels measured for samples labeled M?
2) Denote this sum of ranks by RS(g). What is the minimal value, m, that RS(g) can take?
3) Under the null model, what is the probability of RS(g) = m? (provide a formula for this and explain it)
4) Under the null model, what is the probability of RS(g) = m+1? what is the probability of RS(g) = m+2?
   (provide formulas and explain them)
5) Draw a histogram of the values of RS(g) in the dataset. Here g ranges over all genes in the data (after the clean-up)

c. Differential Expression

The purpose is to determine the statistical significance of differential expression (DE) observed for each gene in H vs M. Evaluate the DE in both one-sided directions for every gene, using both Student t-test and WRS test.

Report the number of genes overexpressed in M vs H at a p-value better (≤) than 0.05 and separately genes underexpressed in M vs H at a p-value of 0.05. For both directions use both a Student t-test and a WRS test.

d. Correlations

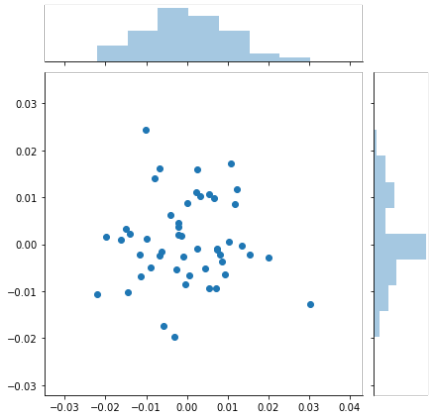
Select the 60 most significant genes from each one of the one-sided WRS DE lists you computed in 3c. Generate a set of 120 genes, D, which is the union of the above two sets.

1) Compute Spearman rho correlations in all pairs within D (120 choose 2 numbers). Represent the correlation matrix as a 120x120 heatmap.

2) What can you report about co-expression of genes in D (co-expression is inferred from the correlation of the expression levels of genes, across a set of samples)? Do we observe any significant co-expression? If so how many pairs, etc.

3) What would have been advantages and disadvantages of computing co-expression for all genes in the study rather than only for genes in D?

4) Provide example datapoints matching the following descriptions. Each example should be constructed over n=50 datapoints. Provide a table description of the example data as well as a jointplot (see figure).



   a) Data with Pearson(x,y) > Spearman(x,y) + 1.2

   b) Data with negative Kendall(x,y) and positive Spearman(x,y) (or vice versa)

5) Can you find 2 pairs of genes that fulfill the conditions in section 4.a and 4.b?

e. Plots and Conclusions of the DE and correlation analysis

1) Construct the DE overabundance plots (blue and green lines as shown in class) for M vs H overexpression (higher expression levels in M) using WRS and t-test using the results you had computed in Section 3c.
   State, for each comparison, the number of genes, k, at which we observe:

   a) FDR = 0.1

   b) FDR = 0.05

   c) FDR = 0.001

   If these events are not observed at any k, then make that statement.

2) What can you say about the difference in results obtained in WRS vs those obtained by Student t-test?

3) Select any 3 differentially expressed genes, from D (which was defined in 3d), and produce a graphical representation of their expression patterns that demonstrates the observed DE.

4) Heatmap
   Draw a heatmap representation of the expression values of the genes in D (from 3d), across the entire cohort (all samples). Order the genes and the samples to produce the maximal visual effect.

f. ML classifiers

1) Split the dataset into a 80/20 train/test random split.

2) Select the 6 most significant DE genes **from the training set** according to WRS – best 3 overexpressed and best 3 underexpressed genes.

3) Perform Naïve Bayes classification to predict the classes M and H and report your results. Use the sklearn library.

4) Build a Decision Tree to predict the the classes M and H and report your results. Use the sklearn library. What would you expect to get had you used all 50K features?

5) Compare the results of the two classifiers.

PART II

New dataset analysis

Report selected analysis steps on a new dataset. Download another dataset from GEO (link) or use any other data, including data from your labs. The selected dataset should have at least 1000 feature rows (possibly genes). It should also have 100 samples and at least two classes of samples. You can choose to work with non-human data if you like.

* Note: your class presentation should include all elements but focus (and spend most time) on PART II. Explain your decisions and choices and present conclusions in a broad context whenever possible.