Biological Data Analysis and Visualization

Introduction

The purpose of this project is to analyze human transcriptome variation across healthy tissues and diseased (cancer) tissues. We use a microarray dataset consisting of 4,454 samples (different individuals/tissues), each with gene expression measurements for 2,561 metabolic genes (PMCID PMC4702927). Note that the analysis would not fundamentally change when using RNAseq instead of microarray data to characterize transcriptome variation.

We use this dataset to investigate the variation of metabolism related genes, and in particular to investigate their potential role in cancer. An intrinsic feature resulting from genomic perturbations in cancer is the deregulation of metabolism in tumor cells. For example, certain metabolic genes need to serve the atypical metabolic needs of tumor cells and are vital to cancer progression. During the project you will analyse expression levels of metabolic genes in healthy tissue and in tumor cells. This analysis can give insight into differences in expression patterns between healthy and diseased tissues, which may be of biological interest and can be relevant to e.g. obtain biomarkers for disease.

Data

We provide a subset of samples from http://merav.wi.mit.edu/. In total, this resource contains microarray data from 4,454 different arrays. From these, we use data from healthy tissues (688x) and from primary tumors (1,444x). We either use all of these data together, or analyze separate subsets for specific tissues (both healthy and tumor tissue). A list of these datasets is provided in Table 1.

Filename	Class 1: healthy	Class 2: tumor
get_normal_vs_tumor_RAW.out	All (688)	All (1,444)
get_normal_vs_tumor2_RAW_Breast.out	Breast (142)	Breast (361)
get_normal_vs_tumor2_RAW_Skin.out	Skin (29)	Skin (43)

Table 1. Datasets. Numbers in brackets indicate number of different samples. For each dataset, two files are provided. These differ in whether quantile normalization is applied or not.

Expected output

For each block, you should hand in a report as a single, self-contained RMarkDown (.rmd) file and the resulting PDF (or Word) file, that:

- clearly describes methodology used;
- clearly presents resulting models;
- interprets the meaning of findings;
- explains your answers to questions asked in the project description.

Submit this report before 23.59 of day 3, through Blackboard.

Block 1: Introduction and data exploration

Today, you will start by inspecting (part of) the data. Download and unzip the project files from Blackboard. Each file contains one header line, followed by N_{sample} rows (here N_{sample} is the sum of the number of healthy and tumor samples provided in Table 1). Column 1 to column 2561 contain data for each of the 2,561 genes. In addition, the last column contains the tissue label for each of the samples, i.e. from which tissue the sample was obtained (“tumor” or “normal”).

First, load the skin tissue data into R. Note that for some visualizations, it is useful to create an additional data frame containing genes as columns and samples as rows, i.e. transpose the original data. This is not trivial, as the data contains both numeric data (the gene expressions) and a factor (sample type). You can use this code (check what **`paste0()`** does):

```
# Transpose just the gene expression
> tdata <- data.frame(t(data[,-2562]))
# Add sample type as column name
> colnames(tdata)<-paste0(data$tissue,1:72)
```

1. Report on the range of expression, generate plots of the expression distribution of both a few selected individual tumor and normal samples, of the average over all tumor samples and the average over all normal samples. Do you notice anything?

2. Make a scatterplot of the average expression level over all tumor samples vs. that over all normal samples. Explain what you see.

3. Make a histogram of the difference between the average expression level over all tumor samples and that over all normal samples. What does this tell you about the number of over- and underexpressed genes? Confirm this by calculating these numbers.

4. Create boxplots of the expression of the genes AMY2B, CLC and NAT1 in tumor and normal samples. What do you see?

Block 2: Visualization and normalization

Start with visualizing the data and looking at normalization. Address the following issues using the skin tissue samples, in the transposed format (**`tdata`**):

1. Generate a heatmap using **`pheatmap()`**, for both log-transformed data and the original data. Also visualize the sample correlation matrix. What do you observe? Are values missing?

2. Apply PCA. What do you notice about the samples and the genes?

3. Try different normalization methods and discuss which is best. Compare the normalized data with the original data. Inspect the correlation between normalized and original values across samples and (some) genes. Visualize the change induced by the normalization.

4. Normalize all three datasets in Table 1, and save the results for using later on in the project.

Block 3: Clustering

Apply hierarchical clustering, and apply *K*-means clustering to the dataset containing Skin tissue (normal and tumor data). Address the following questions:

1. Compare the different clustering results. Do different methods give very different or very similar results?
2. Compare the clustering results with the known disease/tissue labels for the different samples.
3. Can you say something on how many clusters there are in this dataset?

Block 4: Regression

Analyse how gene expression levels in in specific metabolic pathways are related to each other. To do so, make use of linear regression models where you regress gene expression levels with each other. Make a model using one gene as dependent variables, using other genes as predictors/independent variables. Generate three regression models for a given gene: (i) using expression data from all tissues, (ii) using data from normal tissue only and (iii) using data from cancer tissue only. Do so either using data from genes from the pathway only, or using data from all genes in the dataset. (In total this gives six models for one gene.)

Do so with genes from the Oxidative Phosphorylation pathway (PFKFB2, PSPH, PKLR,SDHC) and with genes from the krebs cycle (EHHADH,PCCB,OXCT1,MCCC1,GCDH); pick one gene from each pathway as dependent variable. In total this gives twelve regression models.

Look at least at the following aspects:

1. Create a training set and a test set, and analyse and compare predictive performance on these sets.
2. How many variables are significant? How do you decide on significance?
3. Give some examples of parameter values (sign/size) and interpret these. Why do you need to be careful here?
4. Consider whether it would make sense to log-transform the data before performing the regression. How do your results change when you apply this transformation? Compare amongst others the studentized residual plots, and compare which predictor variables are found as significant when applying this transformation vs. when not applying it.

Note: to get a subset of the data from one specific tissue, you can use:

```
tissue1=”sometissue”
dat=data.frame(dat)
mydat=dat[which(dat$tissue==tissue1),]
mydat=droplevels(mydat)
```

To get a subet of genes, you can subsequently use:

```
genelist= c("gene1”,”gene2”,”gene3”)
mydatnew=mydat[,which(colnames(mydat) %in% genelist)]
```

Block 5: Tests

Apply a statistical test to find out whether the genes from the Oxidative Phosphorylation pathway (PFKFB2, PSPH, PKLR,SDHC) and the genes from the Krebs cycle (EHHADH,PCCB,OXCT1,MCCC1,GCDH) have significantly different average expression in tumor tissue vs. in normal tissue. Comment on your observations. Also consider the use of multiple testing correction.

Block 6: Classification

Build classifiers to discriminate normal from tumor tissue, based on metabolic gene expression patterns. Compare the predictive performance on training and test sets for a number of models. Include at least the following tissue (sub)sets (note that each of these will lead to a different classifier):

- Skin, healthy vs. diseased
- Breast, healthy vs. diseased
- All tissues, healthy vs. diseased

Note that, in particular for “all tissues”, it is wise to pre-select a subset of informative genes (e.g. 100 or 250), as otherwise computation time may become prohibitive.

Address the following questions:

1. What is the performance that can be obtained using the *k*-nearest neighbour classifier? What setting of *k* is optimal?
2. What is the performance that can be obtained using a decision tree classifier? What pruning level is optimal?
3. Can you use a model trained on one of these datasets to predict for any of the others?
4. Can you find informative features (a set of specific genes) that discriminate between normal and tumor tissues?