Modeling Cell Fate

Prof. Dr. Volkhard Helms
Ruslan Akulenko, Mohamed Hamed, Christian Spaniol
Summer Semester 2013

Saarland University
Chair for Computational Biology

Exercise Sheet 3
Due: June 11, 2013 10:15

Cluster analysis of DNA methylation data using R with further interpretation of results uding DAVID.

In this assignment you will practically learn how to apply different clustering methods to reveal additional knowledge from DNA methylation data using R. In the first part you will study available source for DNA methylation data TCGA (The Cancer Genome Atlas). The data from this portal will be used for further analysis.

In the second part of the assignment, the following clustering techniques will be covered: hierarchical aglomerative clustering, k-means clustering and Affinity Propagation. These are the most widely used approaches which you will learn with later visualization of your results.

In the final part, you will see how clustering helped to discover additional knowledge about groups of genes and DNA methylation in general by interpreting the results using DAVID.

Submission

(a) You are advised to work in groups of two people. If necessary, we will suggest teammates. Please submit your solutions in English.

(b) Submit your solutions on paper at the beginning of the lecture in the lecture hall or in Room 3.09, both E2 1. Alternatively you may send an email with a single PDF attachment to :

r.akulenko@bioinformatik.uni-saarland.de late submissions will not be considered.

(c) If appropriate, include source code listings into the submitted document, we will not merge and layout your source code. If relevant sources are missing on the exercise sheet, they will not be graded.

(d) Do not forget to mention your names/matriculation numbers. :)

Discussion of this exercise will be on Tuesday, June 18th at 12:45 c.t. in the lecture room (E2 1 007).

Getting Started.

In the previous assignment you were introduced **R**, free software environment for statistcal computing and graphics. If you didn't use IDE for R before, you are welcome to use Rstudio, which you can get here http://www.rstudio.com/ It will definitily simplify the usage of R, visualization of plots, tracking variables and packages.

Exrcise 1. Getting familiar with TCGA portal, downloading data and preprocessing them (20 points).

The Cancer Genome Atlas (TCGA) data portal https://tcga-data.nci.nih.gov/tcga/ provides a platform to search, download, and analyze data sets related to specific cancer type. It stores genome related data like DNA methylation, protein or gene expression, somatic mutations etc. for more than 20 types of cancer. The portal is hosted by the National Cancer Institute and the National Human Genome

Research Institute.

1.1 Describe briefly in your own words what do **data levels, beta value, tumor – matched** and **normal – matched** mean with respect to the data stored in TCGA. Why there might be **several** probe entries for one gene in the level 3 file (for example Infinium HumanMethylation27 chip)?
1.2 Discuss briefly how different DNA methylation (1) and gene expression (2) data can be matched so that we have values (1) and (2) for the same genes from the same samples.
1.3 Download DNA methylation data for Breast invasive carcinoma by
(Download Data) → (Data Matrix) → (disease = **BRCA**, data type = **DNA Methylation**, Batch Number = **All**, data **level 3**, available, center/platform = **JHU_USC (HumanMethylation27)**, Access Tier = All, Tumor/Normal = **Tumor – matched** + **Normal – matched**) → (Select "Methyl" and then "Build Archive"). As a result you should get 345 sample files. In order to save your time, we will use the subset of the data – sort files according to their names (ascending) and take first **100** files. Preprocessing stage starts with cleaning part – remove entries with empty/NA beta values, gene names; entries with undistinct gene names. Next average beta values for the same genes – if there are several records (different probe id-s) with the same gene name, take all beta values from those records, average them and assign obtained value to the gene. In the end you should get only one entry of beta value for every gene in every sample. Now comes the second shrinkage (the same reason as for amount of files) – take only those genes which start with the letters A, B, C.

Exrcise 2. Clustering of the data (70 points).
2.1 Perform Hierarchical Agglomerative clustering (Euclidean distance, method= "ward") of genes and plot the dendrogram. Visually determine clusters and draw red borders around those clusters.
2.2. In order to perform k-means clustering of genes you need to determine the number of clusters first. Use **all** following approaches and **discuss** the results (additionally report the number of clusters):

(a) Rule of thumb.

(b) Choose number of clusters using Silhouette. Create Silhouette plot (one axis contains silhouette width, another genes). Determine average silhouette and the number of genes in each cluster, compute total average silhouette for all clusters.

(c) Sum of squared errors scree plot for a number of cluster solutions. Create the plot and define the elbow – the number of clusters.
Pick up two cluster solutions and validate them with cluster.stats function from fpc package.
One way to visualize k-means results is to plot genes using first two principle components (you can obtain them using prcomp function from stats package) and color them according to the clusters.
2.3. Run Affinity Propagation to cluster genes.
http://cran.r-project.org/web/packages/apcluster/vignettes/apcluster.pdf
Explain briefly the main idea behind this method and the way you can manipulate the number of clusters. Compare all three clustering methods.

Exercise 3. Pathway enrichment analysis using DAVID (10 points).
DAVID is the Database for Annotation, Visualization and Integated Discovery, which provides a comprehensive set of functional annotation tools. http://david.abcc.ncifcrf.gov/
Take one cluster of genes obtained by Affinity Propagation which contains genes CD1A and CD1E and run pathway enrichment analysis. Do not forget, that we use human DNA methylation samples. Discuss the results.

Good Luck!