Corresponding author: Alex Sánchez, asanchez@ub.edu

Please, indicate the type of presentation you prefer (oral or poster): oral

The work will be submitted to the scientific committee in order to decide the acceptance and type of presentation of the work.

Please, choose the two main topics of your contribution from the following list.
Primary topic: Bioinformatics
Secondary topic: Multivariate Analysis

Bayesian Statistics
Bioinformatics
Clinical trials
Design of Experiments
Epidemiology
Functional Data Analysis
Longitudinal data analysis
Mixed Effects Models
Multivariate Analysis
Nonparametric Methods
Statistical Genetics
Statistical methods in Agriculture
Statistical methods in Biology
Statistical methods in Ecology
Statistical methods in Medicine
Statistical methods in Psychology
Space and space-time modelling
Survival Analysis
Time series
Others (specify)

Scatterplot clustering for the integrative analysis of expression and methylation data

M. Carme Ruíz de Villa, Francesc Carmona, Diego Arango del Corro,
Josep Lluís Mosquera and Alex Sánchez
[1]*{mruiz_de_villa, fcarmona, asanchez}@ub.edu, Departament d'Estadística,*
Universitat de Barcelona
[2]*diego.arango@vhir.org, CIBBIM. Vall dHebron Institut de Recerca.*
[2]*jl.mosquer@vhir.org, Unitat d'Estadística i Bioinformàtica.*
Vall dHebron Institut de Recerca.

Abstract

Methylation analysis is becoming a common approach to complement transcriptomic studies and for each gene the relation between the percentage of methylation and gene expression can be visualized using a scatterplot, what, given the high number of genes lead to the need for clustering scatterplots to detect similar pattern of gene regulation by methylation. Several methods for doing this are compared and the benefits and problems of each approach are discussed.

Keywords: Methylation analysis, clustering scatterplots, bioinformatics.

1. Introduction

DNA methylation, a process involving the addition of a methyl group to the cytosine or adenine DNA nucleotides is known to have the effect of reducing gene expression and it is considered to be one of the main systems for epigenetic regulation [1]. In many disease processes such as cancer but also alzheimer or parkinson disease, certain regions of the genome may acquire abnormal hypermethylation which results in a decrease or even *silencing* of the expression of certain genes. Finding out which genes are affected by methylation and in which form this occurs becomes crucial for the understanding of the disease process [2].

There are many methods for the analysis of methylated data [3] but, in short, what most methods yield consists of a percentage of methylation per each locus, where “locus” here means the potentially methylable genomic region associated to each gene, usually located some hundreds of bases before the transcription start site of the gene. Although there is not a one-to-one correspondence between locus and genes it can be roughly considered to be so: each gene can be regulated by one -occasionally more- locus, each locus is associated to one, occasionally more genes.

To study the way that methylation acts in a given context one can study the expression of the gene and the methylation of the associated locus and see how they are related, for instance by plotting them in a scatterplot [4]. The shape of the scatterplot is considered to be characteristic of the degree of regulation: if a gene is regulated by methylation it is expected to show negative correlation with gene expression. Although the idea seems clear it has been seen that this “negative correlation” may appear in a variety of patterns which makes it difficult to classify or even to clearly decide if the gene is or not regulated by methylation.

In a previous work [5] the authors developed a heuristic method to classify the scatterplots obtained by plotting the methylation degree and the gene expression of 40 samples of increasingly lethal colon cancer cells. Although the method performed relatively well allowing to distinguish genes regulated by methylation from those that were not, it could not perform in a completely automatic way and had difficulties in defining a threshold to call a gene “regulated by methylation”.

In this work we extend this investigation by applying other methods that have been recently published, either specifically for the analysis of methylated data [6] or more generally to cluster scatterplots [7].

The results of the study go in two directions. By one side it is shown that any of these methods can be a good strategy for the analysis of methylation and that, overall, clustering the methylation scatterplots is a good approach to identify methylation regulated genes and types of regulation.

By the other side it is shown that no method performs definitely better than the other and that this performance depends on a variety of factors that range from the available samples (and of course the sample size) to the disease type being considered.

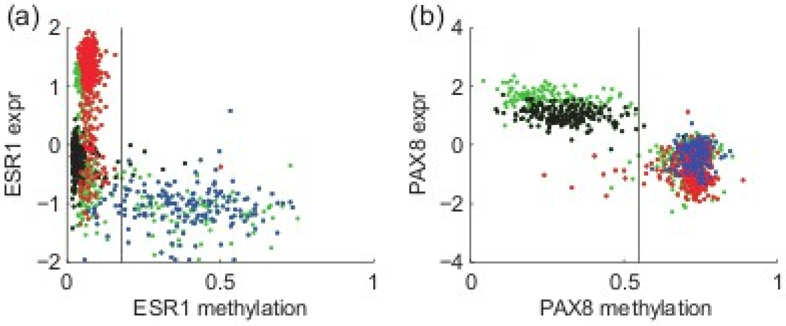

Figure 1: Two example genes whose expression is controlled by methylation [6].

2. Bibliography

[1] Christoph Bock (2012) *Analysing and interpreting DNA methylation data* Nature Reviews Genetics 13, 705-719

[2] Daura-Oller E, Cabre M, Montero MA, Paternain JL, Romeu A (2009). *Specific gene hypomethylation and cancer: New insights into coding region feature trends*. Bioinformation 3 (8): 340–343.

[3] Shen L, Waterland RA. (2007) *Methods of DNA methylation analysis*. Curr Opin Clin Nutr Metab Care. 2007 Sep;10(5):576-81.

[4] Laird, P. (2011) Principles and challenges of genome-wide DNA methylation analysis. Nature Reviews Genetics11, 191–203

[5] Sarah Bazzocco, Hafid Alazzouzi, Ruiz de Villa,MC, Sánchez Pla, A., John M. Mariadason, Diego Arango (2013) *Genome-Wide Analysis Of Dna Methylation In Colorectal Cancer* Submitted

[6] Yihua Liu and Peng Qiu . *Integrative analysis of methylation and gene expression data in TCGA* 2012 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)

[7] Zhanpan Zhang, Xinping Cui, Daniel R Jeske, Xiaoxiao Li, Jonathan Braun and James Borneman4 . *Clustering Scatter Plots Using Data Depth Measures* . J Biomet Biostat 2011, S5

http://dx.doi.org/10.4172/2155-6180.S5-001