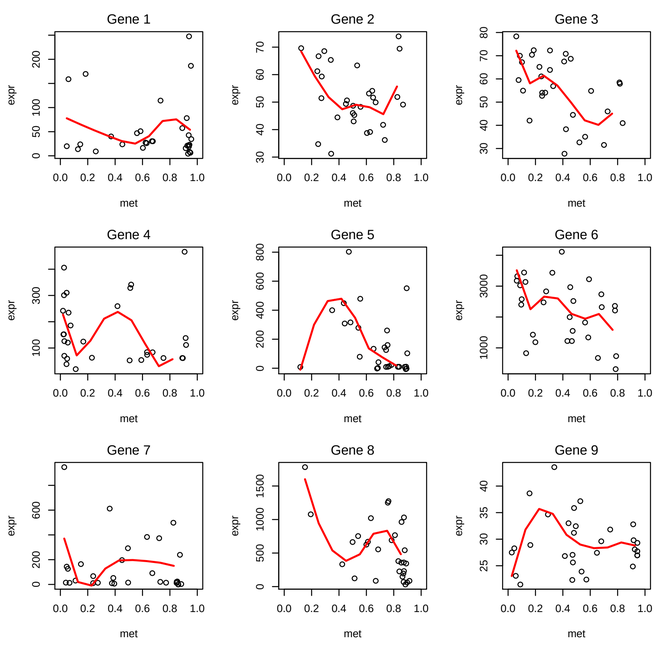
Scatterplot clustering for the integrative analysis of expression and methylation data

M. C. Ruiz de Villa[1], F. Carmona[1], D. Arango[2], J.L. Mosquera[3], A. Sánchez[1,3]

[1] Departamento de Estadística, Universidad de Barcelona, Barcelona
[2] CIBIM, Vall d'Hebron Institut de Recerca (VHIR), Barcelona
[3] Statistic and Bioinformatics Unit. VHIR. Barcelona

1 Introduction

- This study originates in a work searching for colon cancer biomarkers [1] where 30 cell lines, characterized by increasing sensitivity to a drug, were analyzed using several high-throughput methods including expression microarrays and methylation.
- Here we consider the problem of establishing which genes were regulated by methylation.
- For each gene/methylation locus one has 30 points and a scatterplot showing their relation so *we need methods to find patterns of scatterplots.*



2 Objectives

- Study how gene expression is regulated by methylation in a set of colon cancer cell lines.
- Set up a method to detect the level of methylation at which a gene can be considered regulated by methylation (to be "on").
- Compare this method with other that have been developed to
  - detect methylation thresholds and
  - detect patterns in scatterplots.

3 Methods for pattern selection

3.1 Based on Conditional Mutual Information

- When studying methylation we are faced with two main questions:
  1. Which genes exhibit an L-shape, and
  2. What is the optimal threshold for binarizing methylation data for each L-shape gene.
- Following [2] in order to determine whether methylation X and expression Y of a gene exhibit an L–shape, the conditional Mutual Information $cMI(t)$ for different choices of threshold t is computed.

$$cMI(t) = I(X, Y|X > t)P(X > t) + I(X, Y|X \leq t)P(X \leq t)$$

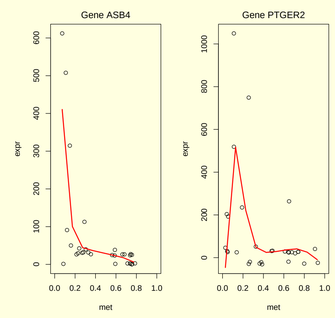
- If the relation between methylation and expression shows an L-shape as t moves from 0 to 1, $cMI(t)$ first decreases and then increases, its value approaching zero when t coincides with the reflection point.
- For an L-shape gene it is verified that:
  - The ratio $r = \frac{\min\{cMI(t)\}}{cMI(0)}$ is small,
  - $t^* = \text{argmin}\{cMI(t)\}$ is the **optimal threshold** for dichotomizing the methylation data of this gene.

3.2 Based on Spline regression

- As an alternative to the previous method we suggest that spline regression [3] can be used for scatter plot clustering.
- In spline regression a curve $y = s(x)$ is represented as $\mathbf{y}_i = \mathbf{B}_i\mathbf{c}$ where
  - $\mathbf{B}_i = [B_{1p}\mathbf{x}_i, B_{2p}\mathbf{x}_i, \ldots, B_{Lp}\mathbf{x}_i]$ the spline basis matrix and
  - $\mathbf{c}$ is the vector of spline coefficients.
- This suggests the following method (and algorithm) for detecting L–shaped genes based on **Clustering Spline Coefficients**:
  1. Select genes with significant correlation.
  2. For each selected gene fit a cubic splines regression model.
  3. Obtain a distance matrix between all genes using the $1 - \rho$ distance computed on spline coefficients.
  4. Perform a hierarchical clustering and
  5. Select genes in the *L-shaped cluster(s)*.

4 Results

- Spline regression: The 2 first clusters included the genes with an L-shape



- Conditional Mutual Information

  We filtered for L-shapes using a combination of three criteria:
  - the ratio $r < 0.25$
  - unconditioned MI $cMI(0) > 0.1$
  - the median expression on the left side of the optimal threshold t^* is higher than the median expression on the right side.
- Comparison between the methods:

Initial selection	191	641
Cluster	Splines	cMI
1	140	102
2	22	16
Total	162	118

- In summary...
  - We have found similar results between both methods.
  - Biological interpretation is in progress but preliminary (unpublished) results are consistent with the hypothesis.
  - Sample size is a limiting factor: cMI works better with hundreds of samples but one may have a very small number (real cases: 30, 12)

[1] Sarah Bazzocco, Hafid Alazzouzi, M. Carme Ruiz de Villa, Alex Sanchez-Pla, John M. Mariadason, Diego Arango (2013) *Genome-Wide Analysis of DNA Methylation in Colorectal Cancer*. Submitted.

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

[3] Jeffrey Racine. (2012) A primer on regression splines. `http://cran.r-project.org/web/packages/crs/vignettes/spline_primer.pdf`