Statistical Methods for Omics Data Integration and Analysis 2014

Scatterplot clustering for the integrative analysis of expression and methylation data

M. Carme Ruiz de Villa, Francesc Carmona, Diego Arango del Corro, Sarah Bazzoco, Alex Sánchez

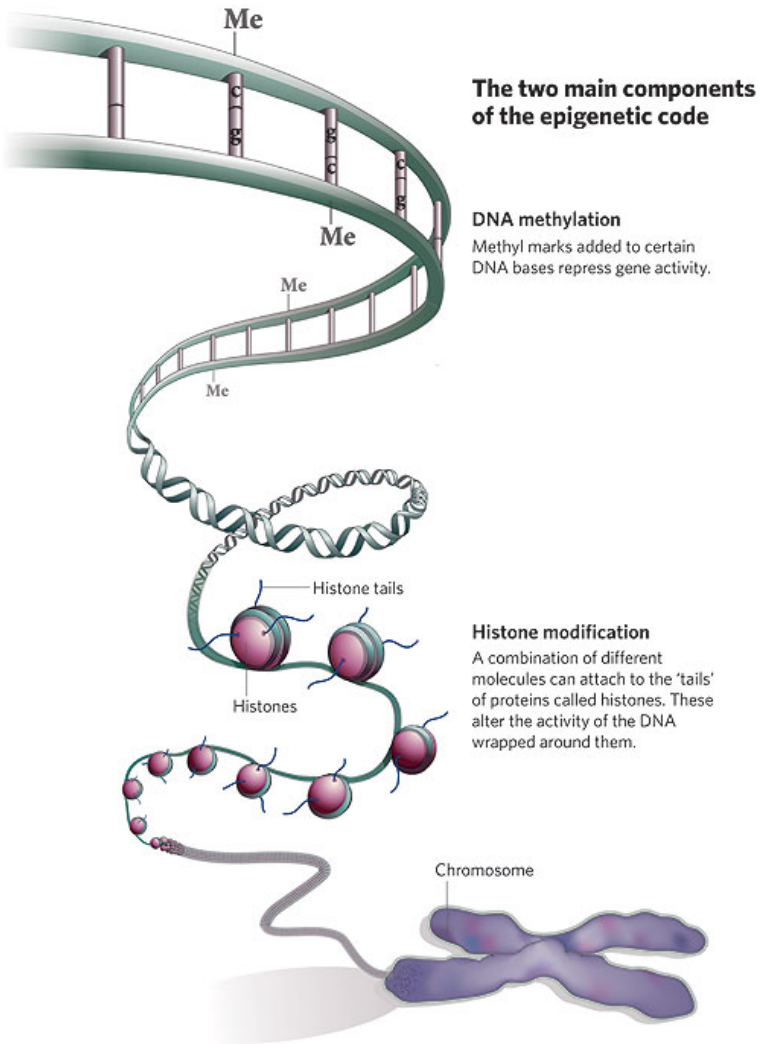
Nov 10-12, 2014

Statistics Department
Facultad de Biología
Molecular Oncology-CIBBIM
Vall Hebron Institut de Recerca

Table of Contents

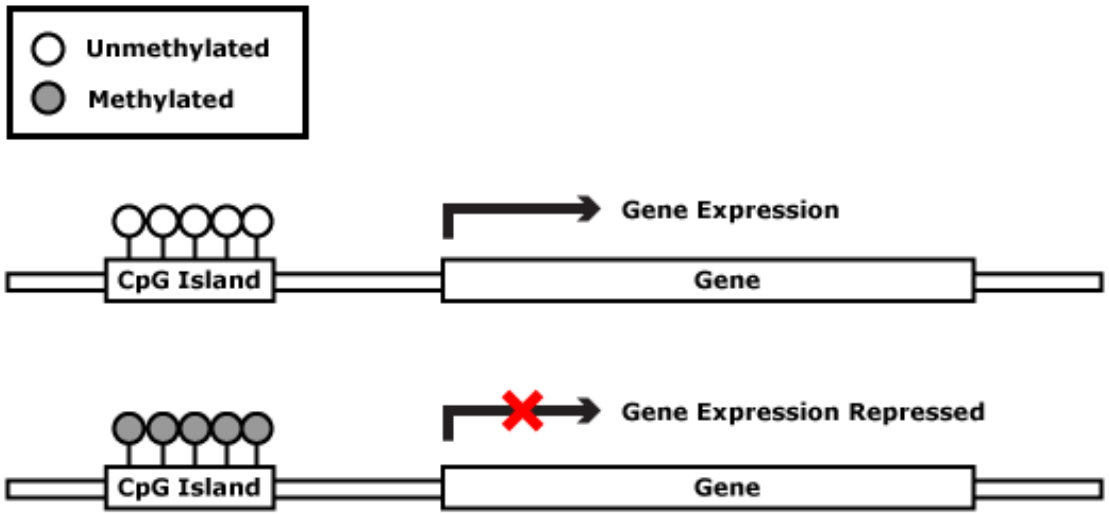
Epigenetics and epigenomics

- Epigenetics, *the study of environmental factors on gene expression in DNA*, shows a renewed impetus:
  - NGS allows in-deep analysis of regulatory mechanisms such as *methylation* or *histone modifications*.
  - There is increasing evidence that many differentiation processes are triggered and maintained through epigenetic mechanisms.



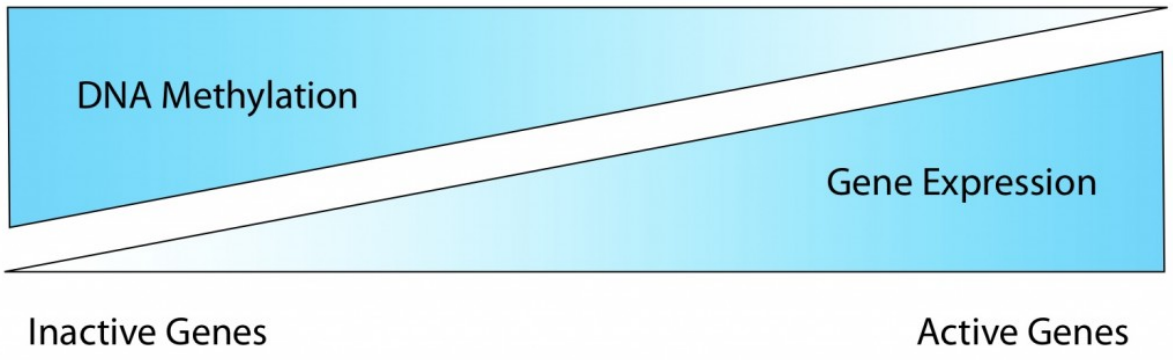
Methylation

- Methylation of CpG dinucleotides in the promoter of genes involved in the oncogenic process has been shown to be a key process contributing to tumor initiation and/or progression.
- Essentially methylation acts by inhibiting gene expression that is, the more methylated is a gene the more repressed is its expression



Methylation and gene expression

- Although the relation between methylation and gene expression is probably continuous ("*the more...the less...*"),



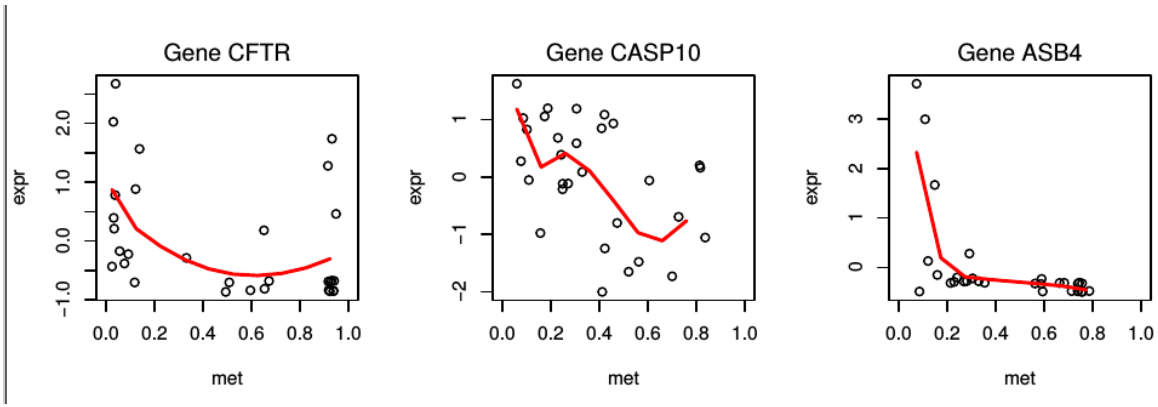
- methylation is, in practice, seen as a dual phenomenon
  - A methylated gene is "off"
  - An unmethylated gene is "on"
- Practical problem: **at which methylation level a gene is seen as "methylated" (is it "turned off")?**

Genome-wide analysis of colorectal cancer

- Methylation of CpG dinucleotides in the promoter of genes involved in the oncogenic process has been shown to be a key process contributing to tumor initiation and/or progression.
- This study originates in a work *aiming at the identification of genes regulated by methylation* as a previous step to obtaining biomarkers for chemotherapy sensitivity
- 30–45 cell lines characterized by increasing sensitivity to several chemotherapy drugs were analyzed using high-throughput methods: *transcriptomics*, *methylation*, *miRNAs*, *SNPs*, and *proteomics*.

Patterns of (negative) association

- Considering the relation between methylation and expression in cancer (the higher methylation the lower the expression...)
- leads to expecting that scatterplots depicting the relation between methylation and expression show a negative correlation.
- This is so and indeed genes known to be regulated by methylation use to show an L-shape pattern in these plots.



Selecting genes by mining scatterplots

- Assuming the relation described above is true...
- Finding genes regulated by methylation is equivalent to finding genes whose methylation–expression scatterplot has an L–shape.
- There is a scatterplot *per* gene and thousands of genes: An automatic method for selecting interesting genes through their scatterplots is needed.

Objectives

The main objectives of this work are:

1. To compare available methods for scatterplot clustering, or to derive new ones if needed.
2. To apply the selected methods on a specific CRC dataset and validate the findings based on their biological relevance.
3. As a secondary objective deriving a binarization point enabling to call a gene “methylated” or “unmethylated” would also be desired.

Overview of approaches

- We have investigated three approaches for selecting L–shaped patterns in scatterplots
  1. Use Conditional Mutual Information to detect threshold point and select genes.
  2. Clustering scatterplots based on the results of Splines Regression and select L-shape clusters.
  3. *Apply Functional Data Analysis techniques to estimate shapes and cluster to extract L–paterns.*
- Only the first two have provided interesting results so the third is omitted.

Selection based on conditional mutual information I

This method was originally proposed by Liu (2012) to study a huge (hundreds of multi–cancer samples) TCGA dataset.
Assume

- That the genes we want to select show an L–shape pattern.
- That methylation is *truly binary*

This has two implications:

- The reflection point of the L-shape is an appropriate choice to binarize methylation data and
- Conditioning on the binarized on-off methylation status, the continuous valued methylation data and expression data should be independent

Selection based on conditional mutual information II

- A relevant issue is how the continuous methylation data are binarized.
- Liu (2012) suggested to use different thresholds, and select the threshold that best separated the two regions.
- The "best" criteria is based on computing mutual information.

Mutual Information and Conditional Mutual Information I

- *Mutual Information* between two random variables X and Y measures the information that these variables share.
- For discrete variables it is defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\, p(y)} \right),$$

- Knowledge of a third variable, Z, can increase or decrease the mutal information between X and Y.
- *Conditional Mutual Information* is the expected value of $I(X;Y)$ once Z is known.

$$I(X;Y|Z) = \mathbb{E}_Z \left[I(X;Y)|Z \right]$$

Mutual Information and Conditional Mutual Information II

The key idea

To determine whether methylation and expression of a gene exhibit an L-shape, one can compute the conditional Mutual Information (MI) for different choices of threshold to binarize the methylation data.

If we consider the continuous valued methylation and expression data as two random variables X and Y, and denote a nominal threshold as t, the conditional MI can be written as a weighted sum of MIs on the two sides of the threshold.

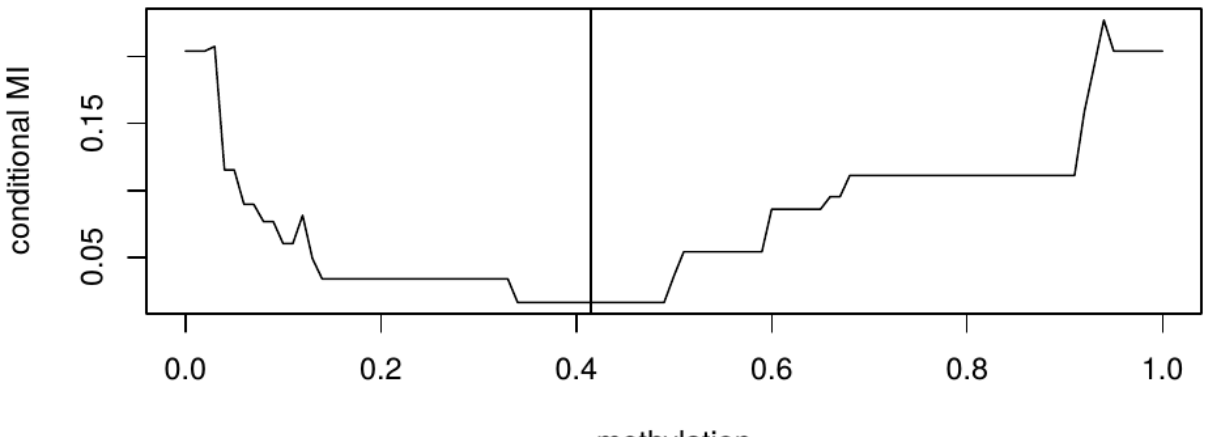
$$cMI(t) = I(X, Y|X > t)P(X > t) + I(X, Y|X \leq t)P(X \leq t)$$

Optimal threshold for binarizing methylation data

- When t is 0 or 1, cMI equals to the mutual information derived from all data points, so:
- for an L-shape gene, as t moves from 0 to 1, $cMI(t)$ first decreases and then increases, and its value approaches zero when t coincides with the reflection point. Therefore,

Optimal threshold

$t^* = \text{argmin}\{cMI(t)\}$ is the optimal threshold for dichotomizing the methylation data of this gene.



Joint distribution estimator

To estimate the MI terms we use a kernel-based estimator, which constructs a joint probability distribution by applying a Gaussian kernel to each data point, and estimates the MI based on the joint distribution. The estimator is as follows:

$$I(X,Y) = \frac{1}{M}\sum_{i=1}^{M} \log \frac{M \sum_{j=1}^{M} e^{-\frac{1}{2h^2}((x_i-x_j)^2+(y_i-y_j)^2)}}{\sum_{j=1}^{M} e^{-\frac{1}{2h^2}(x_i-x_j)^2} \sum_{j=1}^{M} e^{-\frac{1}{2h^2}(y_i-y_j)^2}}$$

where h is a tuning parameter for the kernel width and empirically set $h = 0.3$.

Results (1) Conditional Mutual Information

- Data: Expression and Methylation values from 30 cell lines: two 30×11746 arrays.
- No previous filtering of the genes was needed/performed
- We filtered for L-shapes using a combination of three criteria:
  - Genes with "small" ratio $r < 0.25$
  - Minimum value of unconditioned MI $cMI(0) > 0.1$
  - Median expression on the left side of the optimal threshold t^* must be higher than median expression on the right side.
- Liu(2012) suggests using a random permutation test to select parameter values. We are considering cross-validation
- **A total of 641 genes are selected to be L-shape genes**.

Spline Regression

- Regression based on splines is a form of non-parametric regression that automatically models non-linearities and interactions between variables.
- This is done using *Splines*, continuous functions formed by connecting linear segments. The points where the segments connect are called the *knots* of the spline.
- A particularly efficient form of splines regresion is B-splines.
  - $\varsigma = \{t_1 < \cdots < t_N\}$ non decreasing knot sequence
  - $[t_m, t_{m+1})$ half open interval
  - B_{mp} p-th order polynomial (degree $p-1$) with finite support over the interval and 0 everywhere else so that $\sum_{m=1}^{N-p} B_{mp}(x) = 1$
  - then $s(x) = \sum_{m=1}^{N-p} B_{mp}(x) c_m$

Clustering using Spline regression

To represent the curve we set:

$$y_{ij} = s(x_{ij})$$

So

$$\mathbf{y}_i = \mathbf{B}_i \mathbf{c}$$

with

- $\mathbf{B}_i = [B_{1p}\mathbf{x}_i, B_{2p}\mathbf{x}_i, \ldots, B_{Lp}\mathbf{x}_i]$ the spline basis matrix
- $\mathbf{c}$ the vector of spline coefficients.

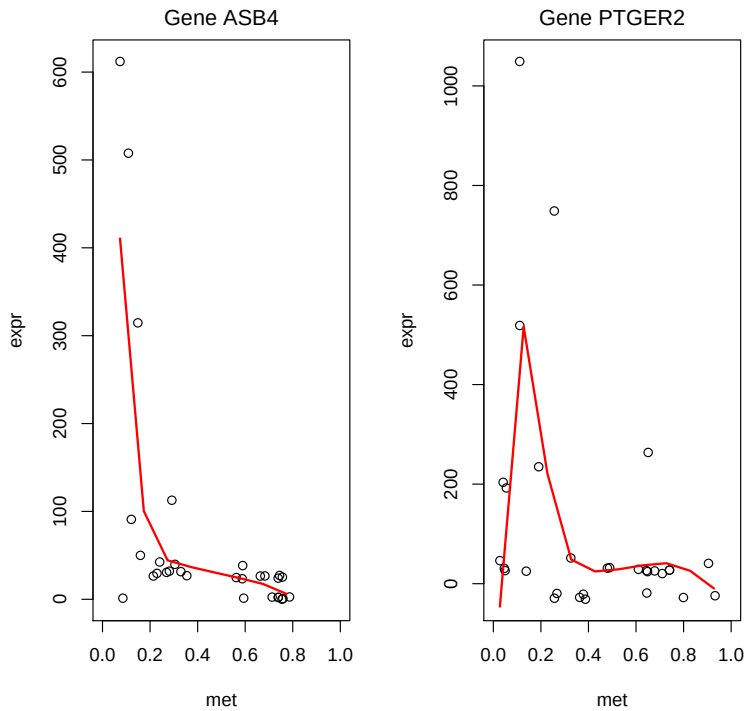
Clustering using Spline regression (3)

Algorithm

1. Selection of the genes with a negative significant correlation
2. Fit cubic regression splines
3. Data to cluster: splines coefficients
4. Calculation of a distance matrix between genes as $1 - \rho$
5. Hierarchical clustering

Results (2) Splines–based regression

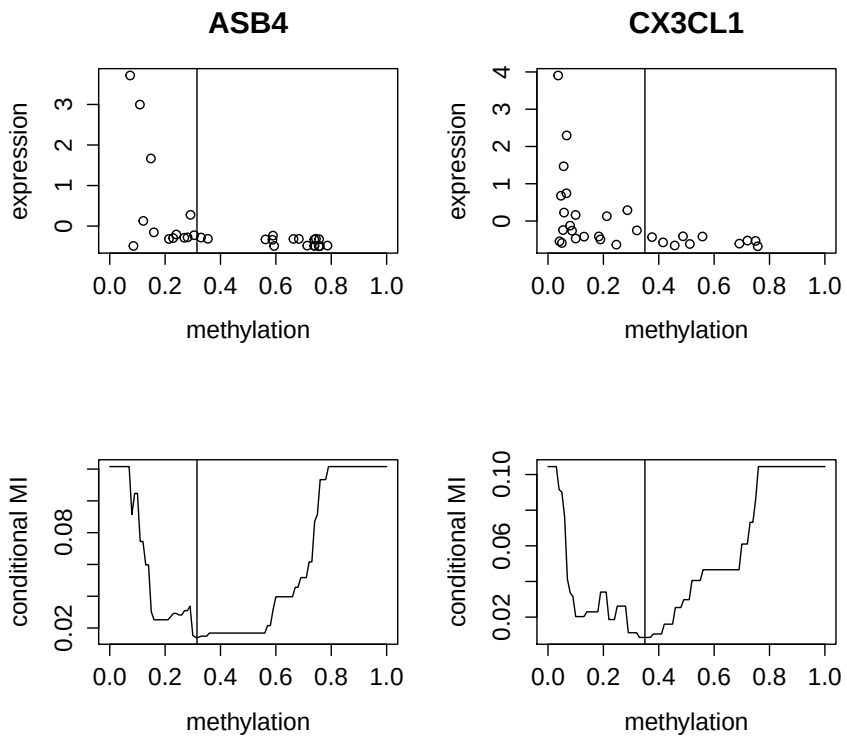
- After the previous selection of genes we worked with 191 genes
- We decided to choose 5 clusters
- The 2 first clusters included the genes with an L-shape



Results (3)

The results of both methods that can be summarized in the following table:

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



Conclusions

- We have found similar results between both methods.
- Biological interpretation is still being done by biological researchers although results are consistent with the hypothesis (we have found genes regulated by methylation).
- 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)

Acknowledgments

- Statistical and Bioinformatics Research Group (*EstBioinfo*), UB.
- Statistical and Bioinformatics Unit (*UEB*), VHIR.
- GRup de Recerca Consolidat Bioestadística i Bioinformàtica (*GRBIO*)