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 and Alex Sánchez

Nov 10-12, 2014

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

Universitat de Barcelona

Vall d'Hebron
Institut de Recerca

Table of Contents

1. Motivation
   - Genome-wide analysis of colorectal cancer
   - Objectives
2. Methods for selecting L–shaped patterns
   - Selection based on conditional mutual information
   - Selection based on Spline regression
3. Results
   - Results from using cMI to select genes
   - Results from using Splines regression to select genes
4. Discussion and Conclusions

Genome-wide analysis of colorectal cancer

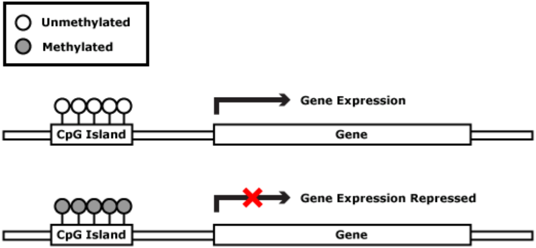
- This study originates in a work aiming at the identification of biomarkers for chemotherapy sensitivity in colorectal cancer (CRC) where the number of available therapies is smaller than in other cancer types.
- The study analyzed a panel of 30–45 cell lines derived from colorectal tumors characterized by increasing sensitivity to several chemotherapy drugs such as
    - Irinotecan,
    - Cetuximab,
    - Oxaliplatin.

Data for the study

- Different high-throughput data were generated:
  - gene expression from Affymetrix (HGU133p2) microarrays,
  - microRNAs from Affymetrix miRNA array,
  - methylation, from Illumina Beadchips and
  - Copy Number Variation from Affymetrix Chip.
- In this work we focus on one of the branches of the work: **the search of genes regulated by methylation**.

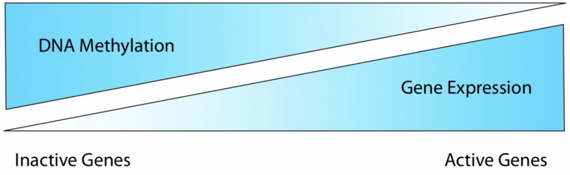
Methylation

- 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 (and especially in cancer) methylation acts by inhibiting gene expression that is, *the more methylated is a gene the more repressed is its expression*



Methylation and gene expression

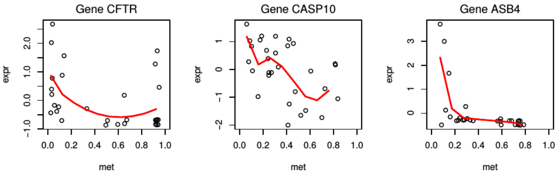
- Although the relation between methylation and gene expression is probably continuous ("*the more...the less...*"),



- 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”)?**

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 required**.

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

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

Data binarization

- 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

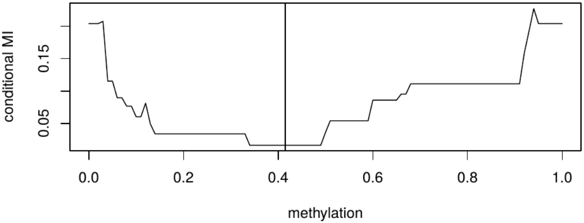
To determine whether methylation and expression of a gene exhibit an L-shape, one can compute the conditional Mutual Information (MI) for different choices of threshold to binarize the methylation data.

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.

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

cMI for L–shaped genes

- When t is 0 or 1, cMI equals to the mutual information derived from all data points, so, for an L–shaped gene i is verified that:
- 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.



Optimal threshold for binarizing methylation data

The behavior of cMI(t) for an L-shape gene, suggests the following criteria to select the optimal binarization point

Optimal threshold

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

L-shape genes will be selected as those verifying reasonable conditions such as:

1. $r = \frac{\min\{cMI(t)\}}{cMI(0)}$ is "small enough".
2. Minimum value of unconditioned MI $cMI(0)$ is "big enough".
3. Left side of the graph reaches higher values than right side

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 regression is B-splines where the splines are B_{mp} p-th order polynomial of degree $p-1$ with finite support over the interval and 0 everywhere.

Clustering using Spline regression

Algorithm

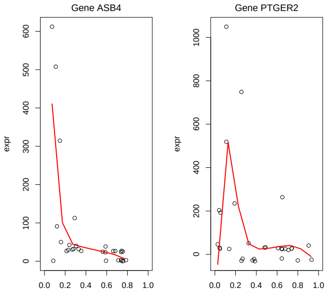
1. Select of the genes with a negative significant correlation. Eventually apply heuristic additional filters that removes genes that are clearly non-L-shaped.
2. Fit a cubic regression splines curve to each pair expression–methylation.
3. Cluster the resulting splines coefficients
4. Select genes in clusters that correspond to L–shapes

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
- Tune L-shape selection using a combination of three criteria:
  - Genes with "small" ratio $r = cMI/MI < 0.25$
  - 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.
- Tuning values selected using cross-validation
- **A total of 641 genes are selected to be L-shape genes**.

Results (2) Splines–based regression

- Using the same dataset we
  1. selected genes with significant negative spearman correlation
  2. applied additional filters to guarantee non L-shape removal.
- After the previous selection of genes we worked with 191 genes
- A hierarchichal clustering yields 5 clearly defined clusters
- The 2 first clusters included the genes with an L-shape

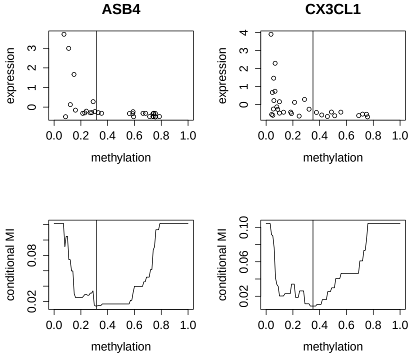


Results (3)

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



Discussion and Conclusions

- cMI based gene selection provides an intuitive approach for selecting L–shaped patterns although it can yield a certain number of "false positives".
  - The method, however works well with big (hundreds) samples which makes it less reliable for normal-size (dozens) datasets.
- Clustering based on the results of Splines regression is also useful in detecting L–shaped patterns.
  - It selects a smaller number of genes than cMI,
  - It is not so dependent from sample size
- Biological interpretation is still ongoing but the results are consistent with the hypothesis (genes known to be regulated by methylation have been found with both methods).

Acknowledgments

GRUP DE RECERCA
EN ESTADÍSTICA I
BIOINFORMÀTICA

UNIVERSITAT DE BARCELONA

UNITAT
D'ESTADÍSTICA I
BIOINFORMÀTICA

Vall d'Hebron
Institut de Recerca

Thanks for your attention!