Integrative Analysis to Select Genes Regulated by Methylation in a Cancer Colon Study

Alex Sánchez-Pla[1,2], M. Carme Ruíz de Villa[3], Francesc Carmona[4], Sarah Bazzoco[5], and Diego Arango del Corro[6]

[1]Statistics and Bioinformatics Unit, Vall d'Hebron Research Institute (VHIR)

[2]Statistics Department. Universitat de Barcelona.
asanchez@ub.edu

[3] Statistics Department. Universitat de Barcelona.
mruiz_de_villa@ub.edu

[4] Statistics Department. Universitat de Barcelona.
fcarmona@ub.edu

[5]CIBBIM-Nanomedicine, Vall d'Hebron Research Institute (VHIR).
sarah.bazzocco@vhir.org

[6]CIBBIM-Nanomedicine, Vall d'Hebron Research Institute (VHIR).
diego.arango@vhir.org

Abstract: Methylation is a regulatory mechanism known to be associated with tumour initiation and progression. Finding genes regulated by methylation is a first step to develop therapies that target these genes, for instance to inhibit tumor development. This study addresses this problem by comparing two methods one based on mutual information and a new one based on clustering the coefficients of fitted curves. The methods are tested on a Cancer Colon study and the biological analysis of the resulting lists suggests that at least some of the genes selected are indeed genes regulated by methylation, opening the door to an automatic mining method.

1 Introduction and Objectives

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[**5**]. Finding genes regulated by methylation can lead to a better understanding of such processes and also be a guide to finding new drug targets.

This study originates in a work aiming at the identification of biomarkers for chemotherapy sensitivity in colorectal cancer (CRC). A panel of 50 cell lines derived from colorectal tumors characterized by increasing sensitivity to several chemotherapy drugs was analyzed using different high-throughput data generation methods. Finding genes regulated by methylation was one of the approaches adopted in the search of candidate genes for new therapies.

In cancer–related genes it is relatively common to observe a decrease in gene expression associated with hypermethylation. Indeed methylation is often described as a binary on-off signal ([**3**]) that is, when methylation is "off" the gene can express normally and its expression will be intermediate or high, whereas when methylation is "on", the expression of the gene will be *repressed* and its values will tend to be low.

Research Perspectives CRM Barcelona, Fall 2015, vol. 7, in Trends in Mathematics
Springer-Birkhuser, Basel

As a consequence of this *high-methylation/low-expression* and *low-methylation/high-expression* relation plots depicting methylation and expression will show L–shape patterns so the strategy adopted will be to mine such plots and select those that have such a shape.

The main objectives of this work are: (i) to select an appropriate method for scatterplot clustering that can be used to mine a multiple high-throughput dataset formed by expression and methylation data and extract the desired patterns, (ii) to test the methods selected on a colon cancer dataset formed by a panel of cell lines derived from colorectal tumors.

2 Methods for L-pattern selection

There have been published several methods to relate methylation and expression values. These range from simple correlation analyis [**7**] to more sophisticated approaches such as the one proposed by Liu and Qiu [**4**]. However, in spite of a certain agreement that the two magnitudes are negatively correlated, there is no generally accepted approach to select genes regulated by methylation. This work intends to be one more step in this direction.

2.1 Gene selection based on Conditional Mutual Information

When studying methylation we are faced with two main questions: (i) Which genes exhibit an L-shape, and (ii) What is the optimal threshold for binarizing methylation data for each L-shape gene.

Liu and Qiu [**4**] suggest to determine whether methylation and expression of a gene exhibit an L-shape by computing the conditional Mutual Information (MI) for different choices of the threshold adopted 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.

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

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.

The ratio $r = \frac{\min\{cMI(t)\}}{cMI(0)}$ for an L-shape gene is small, and $t^* = \text{argmin}\{cMI(t)\}$ is the **optimal threshold** for dichotomizing the methylation data of this gene.

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:

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

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

2.2 Gene selection based on Spline Regression

The above approach is appealing but previous studies suggest that it works best when the number of samples is very big -perhaps hundreds or even thousand samples. This is a common sample size when working for example with TCGA samples [**6**] but not for

individual experiments. As an alternative we suggest to fit a curve to each scatterplot, that is to the relation between expression and methylation for each gene, and then cluster these lines and keep those clusters that can be associated with an L-pattern.

The relation between expression and methylation is weak and non-linear, so a reasonable option for modelling this type of data is *splines regression* a form of non-parametric regression that automatically models non-linearities [**2**]. *Splines* are 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 [**2**] where the splines are B_{mp} p-th order polynomial of degree $p-1$ with finite support over the interval and 0 everywhere else:

With this representation we have applied the following algorithm to select genes regulated by methylation

(1) Prefilter genes to be fitted, for instance select those having a significantly negative Spearman correlation coefficient.
(2) Fit a cubic regression spline to each gene and extract the spline coefficients.
(3) Use coefficients to compute a distance matrix based on a "1-correlation" distance.
(4) Perform hierarchichal clustering on this distance matrix.
(5) Select clusters that visually adapt to an L–shape.

3 Results and Application: Selecting L–shaped genes from a Genome-wide analysis of colorectal cancer

We have applied the methods described above to the experimental data obtained from an ongoing CRC study [**1**]. The data analyzed consisted of expression and methylation values obtained respectively from Affymetrix (hgu133plus2 expression microarrays) and Illumina (256K methylation arrays). Expression and methylation data do not have a one to one correspondence so they were preprocessed separatedly using standard approaches for thes types of data and then aggregated on a gene basis so they could be matched. This process yield two 30 (samples) $\times$ 11746 (genes) arrays.

3.1 Results using the Conditional Mutual Information approach

The data were processed using the algorithm for finding the optimal binarization threshold described above and genes with L-shape were selected using a combination of three criteria:

(1) Genes had "small" ratio between conditional mutual and overall mutual information. This was set to $r = cMI/MI < 0.25$.
(2) The minimum value of overall mutual information was at least 0.1, that is, $cMI(0) > 0.1$.
(3) The median expression on the left side of the optimal threshold t^* had to be higher than median expression on the right side.

Applying the above criteria yield a total of 641 genes that could be considered to have a L-shape.

3.2 Results using Splines Regression to select genes

Splines regression cannot be applied to all the genes so a prefiltering step was used, and only genes showing a significant negative Spearman correlation were modelled to avoid an excess of noise that would negatively afect clustering later. A heuristic filter was also applied to guarantee non L-shape removal. Overall this led to keep 191 genes for which splines were fitted and clustered into 5 clusters. The 2 first, majoritary, clusters included 162 genes that could be considered to have a L-shape.

There were a total of 98 genes in common selected by the two methods.

4 Discussion and Conclusions

This study can still be considered preliminary but a certain number of consistent results can be highlighted:

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. While 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 that is, genes known to be regulated by methylation have been found with both methods.

Acknowledgements

The first and third author wish to acknowledge the *Grup de Recerca en Bioestadística i Bioinformàtica* (GRBIO, Grup Consolidat 2014 SGR 464 Generalitat de Catalunya), for funding part of this work.

References

[1] Sarah Bazzocco, Hafid Alazzouzi, M. C. Ruiz de Villa, Alex Sanchez-Pla, John M. Mariadason, and Diego Arango. Genome-Wide Analysis of DNA Methylation in Colorectal Cancer. *Submitted*, 2016.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, second edition edition, July 2009.

[3] Yihua Liu, Yuan Ji, and Peng Qiu. Identification of thresholds for dichotomizing DNA methylation data. *EURASIP Journal on Bioinformatics and Systems Biology*, 2013(1):8, 2013.

[4] Yihua Liu and Peng Qiu. Integrative analysis of methylation and gene expression data in TCGA. In *Genomic Signal Processing and Statistics, (GENSIPS), 2012 IEEE International Workshop on*, pages 1–4, December 2012.

[5] B Sadikovic, K Al-Romaih, J.A Squire, and M Zielenska. Cause and Consequences of Genetic and Epigenetic Alterations in Human Cancer. *Current Genomics*, 9(6):394–408, September 2008.

[6] The Cancer Genome Atlas Research Network, John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, October 2013.

[7] James R. Wagner, Stephan Busche, Bing Ge, Tony Kwan, Tomi Pastinen, and Mathieu Blanchette. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biology*, 15(2):R37, February 2014.