PREDICTION OF EPIGENETIC BIOMARKERS IN BREAST CANCER CELL LINES BY LOGISTIC REGRESSION OF METHYLATION-EXPRESSION ASSOCIATIONS

Leandro A. Loss*[1], Anguraj Sadanandam[1], Steffen Durinck[1], Shivani Nautiyal[2], Diane Flaucher[2], Victoria E. H. Carlton[2], Martin Moorhead[2], Yontao Lu[2], Joe W. Gray[1], Malek Faham[2], Paul Spellman[1], Bahram Parvin[1]

[1] *Lawrence Berkeley National Laboratory*
Berkeley, CA, USA

[2] *Affymetrix Inc.*
Santa Clara, CA, USA

* *Email: LALoss@lbl.gov*

Epigenetics refers to the study of heritable changes that cannot be explained by changes in the DNA sequence [1, 2]. One mechanism of epigenetic regulation involves DNA methylation of CG dinucleotides, commonly represented as CpG. Patterns of methylation in the CpG islands play an important role in regulating gene expression during both normal cellular development and disease processes. Increased methylation of CpG islands (hypermethylation) in tumor suppressor genes has been observed during tumor progression and metastasis as a result of aberrant methylation patterns [3, 4]. At the same time, aberrations leading to decreased methylation of CpG islands (hypomethylation) of oncogenes are known to occur [5]. Since hyper and hypomethylation of the genome are considered widespread attributes of tumors, predicting the regulation of gene expression through CpG island methylation at an epigenome level will provide a better understanding of the tumor pathobiology and progression, while revealing potentially new biomarkers.

In this work, we propose a computational pipeline to identify epigenetically regulated genes from a panel of cell lines. Associations between regulation data and CPG methylation are formed and ranked through logistic regression, which assesses the data's statistically significant negative correlation and, therefore, the gene's epigenetic regulation. More specifically, the proposed computational pipeline for prediction of epigenetically regulated genes consists of three steps: (i) dimensionality reduction of the methylation profile, which reduces the subsequent computational load significantly and is comprised of two sub-steps: (i.i) clustering of methylation profiles on the basis of proximity, where sub-groups of methylation sites are formed, and (i.ii) clustering within methylation sub-groups on the basis of similarity, where *spectral clustering* is used for identification of patterns within the data; (ii) association of the clustered methylation data with the gene expression data, where methylation sites are used for estimating the site's original genomic position; and (iii) logistic regression and ranking of the methylation-expression associations, where an exponential function is used to assess the data's negative correlation and statistical significance.

Positive aspects of our framework include: (i) high-volume relational data are managed efficiently; (ii) our clustering approach is non-iterative and non-sensitive to the initial condition, leading to stable solutions; (iii) our logistic regression yields flexibility, allowing for the natural distribution and scale of data; and (iv) confidence in the solution is established in terms of statistical significance (i.e., p-value).

Fifty-eight epigenetic biomarkers, including COL1A2, TOP2A, TFF1, and VAV3, were predicted from experiments with a panel of 45 breast cancer cell lines widely used in the Integrative Cancer Biology Program (ICBP)[6] and a recently developed high-throughput technology for measuring genome-wide methylation patterns (mTACL)[7].

References

1. Russo VEA, Martienssen RA and Riggs AD. Epigenetic mechanism of gene regulation. *Cold Spring Harbor Laboratory Press* 1996.
2. Bock C and Lengauer T. Computational epigenetic. *Bioinformatics* 2008; **24**:1–10.
3. Esteller M. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene* 2002; **21**:5427–5440.
4. Jones PA and Baylin SB. The fundamental role of epigenetic events in cancer. *Nature Rev. Genetics* 2002; **3**:415–428.
5. Widschwendter M, Jiang G, Woods C, Mller HM, Fiegl H, Goebel G, Marth C, Mller-Holzner E, Zeimet AG, Laird PW and Ehrlich M. DNA Hypomethylation and Ovarian Cancer Biology. *Cancer Research* 2004; **64**:4472–4480.
6. Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, Clark L, Bayani N, Coppe JP, Tong F, Speed T, Spellman PT, DeVries S, Lapuk A, Wang NJ, Kuo WL, Stilwell JL, Pinkel D, Albertson DG, Waldman FM, McCormick F, Dickson RB, Johnson MD, Lippman M, Ethier S, Gazdar A and Gray JW. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer cell* 2006; **10**:515–527.
7. Nautiyal S, Carlton VEH, Lu Y, Ireland J, Flaucher D, Moorhead M, Gray JW, Spellman PT, Mindrinos M, Berg P and Faham M. A High-Throughput Method for Analyzing Methylation of CpGs in Targeted Genomic Regions. *PNAS in press.*

*Corresponding author.