Vilnius University
Institute of Mathematics and Informatics
L I T H U A N I A

INFORMATICS (09 P)

APPLICATION OF NOVEL DATA MINING METHODS IN HUMAN EPIGENOME RESEARCH AIMED AT EARLY DIAGNOSIS OF COMPLEX NON-MENDELIAN DISEASES

Ieva Merzvinskaite

October 2017

Technical Report MII-DS-09P-17-<report nr.>

VU Institute of Mathematics and Informatics, Akademijos str. 4, Vilnius LT-08663, Lithuania
www.mii.lt

Abstract

The aim of my research is to find out, by applying various data mining and statistical methods, whether common aging related patterns exist in cancer/ control groups and across different tissues in the obtained DNA methylation data.

In the analysis, sixteen different publicly available datasets from TCGA[2] database were investigated to find out how epigenetic aging processes manifest themselves in cancer tissues. The second goal is to identity if normal and cancer tissues are different in terms of epigenetic aging. And if the onset of cancer can be predicted early using aberrant epigenetic aging markers.

I used the publicly available datasets from TCGA database. Each dataset comes from a different case/ control cancer study where DNA modification of samples from different tissues were analyzed using *Illumina Infinium®450k Human DNA methylation Beadchip* microarray technology. The preliminary data analysis was done using statistical and data analysis methods like *Principal component analysis*, *Multiple linear regression*.

The first year literature review results revealed that we are solving an important problem and our approach is unique. Preliminary data analysis identified multiple probes showing differiantial aging trends in various cancers. And ontology analysis revealed that the probes may be related to the some processes independently of cancer type.

Keywords: Cancer, Aging, Bioinformatics

[2]The Cancer Genome Atlas

Contents

1 Introduction

The aim of the research is to find out, by applying various data mining an statistical methods, whether a general aging related pattern exists across different tissues.

DNA methylation data is used for the analysis in this report. DNA methylation is an epigenetic modification when methyl group is added to C-5 position of the cytosine nucleotide in the deoxyribonucleic acid (DNA). Epigenetic modifications is important to cell specialization. DNA methylation plays an important role in gene expression. The increased methylation (*hypermethylation*) in gene promoter area might silence an expression, and a decrease of DNA methylation (*hypomethylation*) might increase the expression of gene. DNA methylation is also important for chromosomal stability, and the abnormalities in it were associated with various diseases, and cancer [BJ11].

Various studies revealed that DNA methylation changes during the time. Age is a major risk factor for cancer. While investigating DNA methylation changes during a lifetime, it was discovered that it is possible to calculate human age using methylation levels in particular genomic positions. A high difference between chronological and predicted age, an accelerated aging, was associated with an increased risk to get a disease [LYU+16, MSM+15, S13].

In this report I search for aging related epigenetic marks which would differ between healthy and cancer samples. Identification of such patterns would reveal how DNA methylation is associated with aging process, and which abnormalities in DNA methylation is related to cancer. It would help to understand disease etiology, diagnose it early, and find the way to stop its progression.

Various statistical and data analysis techniques were applied to analyze different cancer datasets. DNA methylation data analysis is challenging because of extreme variation across different tissues; $p \gg N$ problem [HTF09] when a number of features (various genomic positions) is higher than a number of samples; data quality problem; batch effects; outliers (bad samples which could misrepresent the results). Data pre-processing (*normalization, identification and removal of potencial batch effects, outliers removal using PCA*) and analysis technique present in this report helped to solve most of problems and get preliminary results.

Preliminary results suggest that aging related pattern exist in all cancer datasets. Significant genomic positions (probes) were identified. Probes were mapped to corresponding gene. And gene ontology analysis revealed that most of genes are involved in similar molecular, biological processes, and are related to a same signaling pathway.

2 Methods

Sixteen different publicly available datasets from TCGA [NtNHGRIN17] were used in the analysis. Each dataset consists of healthy and cancer patients samples with their medical records, and methylation values from various genomic positions.

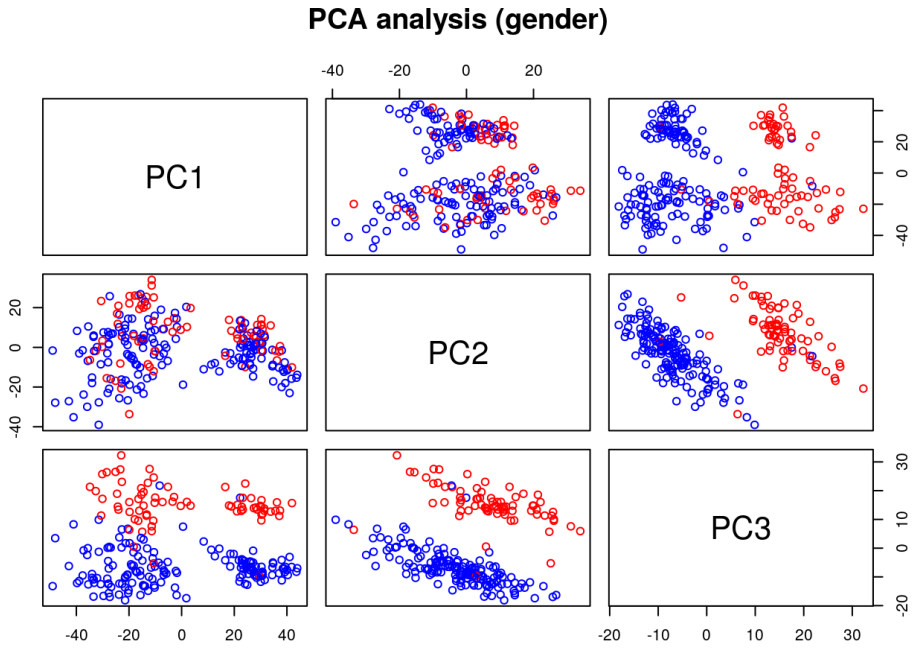

Figure 1: Gender effect shown in the first three principal components (kidney tissue analysis).

2.1 Pre-processing

Raw data was normalized using functional normalization [FLL+14].

A majority of samples come from *caucasian* race patients. The race effect (a potencial batch effect) was removed by leaving only one race (*caucasian*) samples.

Missing values were presented in the data. They were imputed using mean values.

The existing outliers were removed using first three principal components (PCs) after principal component analysis (PCA) [Pea01, Hot33]. The samples, which fall outside 3 standard deviations from mean in at least one of three PCs, were removed.

Principal components, which correlated with samples' medical records (potential batch effects) such as batch number, vital status, gender (figure 1), and etc., were used as confounders in the further analysis.

2.2 Analysis

Significant probes identification was done using linear model. Zero and alternative models were defined and compared. Zero model is the subset of alternative model. Both models are applied to the data, and the residual sum of squares (RSS) is calculated. F statistic is used to test if RSS differs in two models. If calculated p value is below the threshold 0.05 it is assumed that the models differs.

Linear model were chosen because it was noticed that DNA methylation is expressed exponentially at early age, but later it is expressed linearly. And analyzed data is from older patients.

Different approaches were defined to reveal different aging related patterns.

Firstly, I wanted to identify aging related probes. Zero model (1) and alternative model (2) were compared, and the probes which p value was below the threshold were assumed to be significant.

$$methylation\_expression \sim 1 + age + diagnosis + confounders \tag{1}$$

$$methylation\_expression \sim 1 + diagnosis + confounders \tag{2}$$

The second question was if age:disease interactions in the alternative model (3) explains more than a simple zero model (4).

$$methylation\_expression \sim 1 + age : diagnosis + age + diagnosis + confounders \tag{3}$$

$$methylation\_expression \sim 1 + age + diagnosis + confounders \tag{4}$$

The third approach by comparing alternative (5) and zero model (6) helped to identify in which probes the effect of aging differs between cases and controls.

$$methylation\_expression \sim 1 + age + age : diagnosis + diagnosis + confounders \tag{5}$$

$$methylation\_expression \sim 1 + age + confounders \tag{6}$$

The last approach was used to identify differentially aging probes. Zero (8) and alternative models (7) were compared.

$$methylation\_expression \sim 1 + age + age : diagnosis + diagnosis + confounders \tag{7}$$

$$methylation\_expression \sim 1 + confounders \tag{8}$$

Lists of significant probes were got for each tissue using previously described method.

Significant probes were mapped to corresponding genes and ontology analysis was done to identify major molecular, biological processes and signaling pathways in which participate related genes.

3 Results

Preliminary analysis revealed that aging related patterns exist in the data.

Significant probes identified in previous analysis (See Methods) were mapped to corresponding genes. Gene ontology analysis was done using PANTHER tool [PAN17, TKC+03] for each cancer dataset. It revealed that most of genes are involved in similar

molecular, biological processes, and are related to the same signaling pathway.

4 Conclusions and further work

Preliminary results revealed that aging related patterns exist in analyzed cancer data. And gene ontology analysis showed that genes, related to significant probes, might come from similar biological processes. Further investigation is required to get a better understanding of presented biological processes, and how they are related to aging. Understanding of biological processes would help to improve the analysis, and preliminary results.

The following steps will be done to improve the analysis and preliminary results:

1. Review preliminary results and continue investigation into gene ontologies
2. Review current pre-processing and analysis steps
3. Continue investigation into aging related patterns in cancer

References

[BJ11] Stephen B. Baylin and Peter A. Jones. A decade of exploring the cancer epigenome — biological and translational implications. *Nature Reviews Cancer*, 11(10):726–734, 2011.

[FLL+14] Jean-Philippe Fortin, Aurélie Labbe, Mathieu Lemire, Brent W Zanke, Thomas J Hudson, Elana J Fertig, Celia MT Greenwood, and Kasper D Hansen. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biology*, 15(11):503, 2014.

[Hot33] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, (24):417–441,498–520, 1933.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction (Second Edition)*. Springer, 2009.

[LYU+16] Perna L, Zhang Y, Mons U, Holleczek B, Saum K-U, and Brenner H. Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a german case cohort. *Clinical Epigenetics*, 8(64), 2016.

[MSM+15] Riccardo E Marioni, Sonia Shah, Allan F McRae, Brian H Chen, Elena Colicino, Sarah E Harris, Jude Gibson, Anjali K Henders,

Paul Redmond, Simon R Cox, Alison Pattie, Janie Corley, Lee Murphy, Nicholas G Martin, Grant W Montgomery, Andrew P Feinberg, M Daniele Fallin, Michael L Multhaup, Andrew E Jaffe, Roby Joehanes, Joel Schwartz, Allan C Just, Kathryn L Lunetta, Joanne M Murabito, John M Starr, Steve Horvath, Andrea A Baccarelli, Daniel Levy, Peter M Visscher, Naomi R Wray, and Ian J Deary. Dna methylation age of blood predicts all-cause mortality in later life. *Genome Biology*, 16(1):25, 2015.

[NtNHGRIN17] National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). The Cancer Genome Atlas (TCGA). Link: `https://cancergenome.nih.gov/`, 2017.

[PAN17] PANTHER. The Protein ANalysis THrough Evolutionary Relationships (PANTHER). Link: `pantherdb.org`, 2017.

[Pea01] K. Pearson. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 11(2):559–572, 1901.

[S13] Horvath S. Dna methylation age of human tissues and cell types. *Genome Biology*, 14(10):R115, 2013.

[TKC+03] Paul D Thomas, Anish Kejariwal, Michael J Campbell, Huaiyu Mi, Karen Diemer, Nan Guo, Istvan Ladunga, Betty Ulitsky-Lazareva, Anushya Muruganujan, Steven Rabkin, Jody A Vandergriff, and Olivier Doremieux. Panther: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Research*, 31(1):334–341, 2003.