Machine learning network-constrained regression of epigenetic data
An MPhil project proposal

Sivo Daskalov (sd760), Corpus Christi College

Project Supervisor: Dr Pietro Lio'

Abstract

DNA methylation is one of the epigenetic mechanisms that control gene expression. Its disruption is considered to be among causes for the development of breast cancer. Publicly available epigenetic datasets will be analyzed in search of genes and methylation patterns responsible for the development of breast cancer. Identifying those patterns could potentially help predict and prevent the disease. This is the reason behind the growing interest for the development of computational methods in epigenetics. Recently suggested network-based regression will be considered for the analysis of the datasets due to the high dimensionality of epigenetic data.

1. Introduction, approach and outcomes

Epigenetics[1] studies the heritable changes in gene expression that are not due to any alteration in the DNA sequence. Its mechanisms include DNA methylation[2] and histone modifications, which result in the heritable silencing of certain genes. Disruption of the gene expression could lead to major pathologies, including cancer[3]. DNA methylation is considered to be the best-known epigenetic marker. It occurs in a complex chromatin network and is influenced by histone structure modifications, commonly disrupted in cancer cells.

There are various epigenetic datasets that are freely available for download and analysis. One of them is the Encyclopedia Of DNA Elements (ENCODE), which could be used for extraction of health control data with respect to the roadmap reference epigenomes[4]. Genotype changes and epigenetic patterns, which could potentially lead to the development of breast cancer, will be explored. Certain genes are expected to be involved the development of this type of cancer[7]. This expectation will either be confirmed or denied through the analysis of the chosen datasets.

DNA methylation data of healthy and diseased cells will be compared. The goal is to find epigenetic patterns that could be used for the prediction and possibly prevention of breast cancer, and to improve our understanding of the processes leading to its development.

Epigenetic datasets are often characterized with high dimensionality and require complex approaches for their analysis. The recently suggested network-based regression[5][6] will be considered for the processing of the datasets. It is a powerful tool that enables the integration of known correlation structures between genes into the analysis of a dataset. This is especially valuable for epigenetics as gene correlation can be taken into account through its integration in the network. Attempts to analyze and improve the classification capabilities of the aforementioned methodology will be made.

Java and Python will be considered for the implementation of the methodology. R could be used for parts of the statistical computation and analysis. Other platforms and languages may also be utilized.

2. Workplan
   - Weeks 1 and 2 – Researching the programming languages R and Python
   - Weeks 3 and 4 - Literature survey of epigenetics and DNA methylation
   - Weeks 5 and 6 – Literature survey on network regression and various data science approaches with application in epigenetics
   - Weeks 7 and 8 – Choice of datasets
   - Weeks 9 and 10 - Data preprocessing
   - Weeks 11 and 12 – Methodology implementation
   - Weeks 13 and 14 – Methodology implementation (cont.)
   - Weeks 15 and 16 – Data processing
   - Weeks 17 and 18 – Data processing (cont.)
   - Weeks 19 and 20 – Evaluation of results
   - Weeks 21 and 22 – Drawing conclusions from obtained results
   - Weeks 23 and 24 – Writing of the project dissertation
   - Weeks 25 and 26 – Writing of the project dissertation (cont.)
   - Weeks 27 and 28 – Finalization

3. References

[1] Robin Holliday (2006) Epigenetics: A Historical Overview, Epigenetics, 1:2, 76-80.

[2] Jones, P. A., & Takai, D. (2001). The role of DNA methylation in mammalian epigenetics. *Science*, *293*(5532), 1068-1070.

[3] Esteller, M. (2008). Epigenetics in cancer. *New England Journal of Medicine*, *358*(11), 1148-1159.

[4] Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., ... & Amin, V. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317-330.

[5] Iuliano, A., Occhipinti, A., Angelini, C., Feis, I. D., & Lió, P. (2016). Cancer Markers Selection Using Network-Based Cox Regression: A Methodological and Computational Practice. *Frontiers in Physiology Front. Physiol., 7*.

[6] Li, C., & Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics, 24*(9), 1175-1182.

[7] Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61-70.