Analysis of mRNA data of patients with atherosclerosis[1]

Taghi Aliyev

May 13, 2014

Abstract

Keywords: Weighted Gene Correlation Network Analysis, Regression, mRNA, Atherosclerosis

1 Introduction

Atherosclerosis is a heart disease in which an artery wall thickens as a result of accumulation of calcium and fatty materials such as cholesterol and triglyceride. Arteries are blood vessels that carry oxygen-rich blood to your heart and other parts of the body.

2 Weighted Gene Co-Expression Network Analysis

Weighted gene co-expression network analysis(WGCNA) is a systems biology method for describing the correlation patterns among genes across microarray samples. WGCNA is a reverse engineering method for reconstruction of gene correlation network from expression profiles of genes in given patient data. It can be used for finding clusters (modules) of highly correlated genes, for summarizing such clusters using the module eigengene or an intramodular hub gene, for relating modules to one another and to external sample traits (using eigengene network methodology), and for calculating module membership measures [2].

2.1 Network construction

Construction of WGCNA is essential in detection of modules and analysis of interconnectivity among the genes represented in patient data. In these networks, nodes are the genes and strength of connection depends on the pairwise connection of expression profiles of genes.

2.2 Selection of soft-thresholding power

As already mentioned, in order to reconstruct the network architecture, one has to select the soft-threshold power, β, that is used in computation of the network values[3].

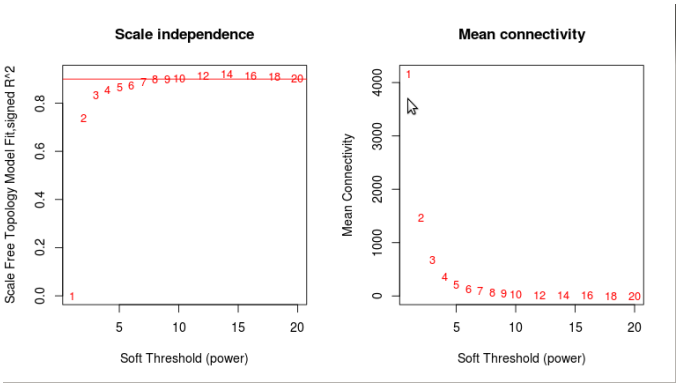

Figure 2.1: Soft-thresholding power selection

2.3 Interconnectivity analysis

2.4 Module Detection

Modules represent set of genes that are highly correlated and have similar connection strength and correlation with all the other genes. Method to compute modules uses Topological Overlap-based Dissimilarity Mapping as its dissimilarity measure and hierarchical clustering. First, we denote Topological overlap between genes i and j as being following:

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{min(k_i, k_j) + l - a_{ij}} \tag{2.1}$$

Where, $l_{ij} = \sum_{u \neq i,j} a_{i,u} a_{u,j}$ represents the amount of genes that both i and j are connected. The topological overlap matrix (TOM) is then generated as being $\Omega = [\omega_{ij}]$. ω_{ij} is a number between 0 and 1 and is symmetric, meaning $\omega_{ij} = \omega_{ji}$. As, module detection methods use dissimilarity measure for their computations, we define Topological Overlap-based Dissimilarity Measure as follows: $d^{\omega}_{ij} = 1 - \omega_{ij}$.

2.5 Transcription Factor Analysis

2.6 Gene Ontology Enrichment Analysis

Using BiNGO plugin on Cytoscape software.

[1] This thesis was prepared in partial fulfillment of the requirements for the Degree of Bachelor of Science in Knowledge Engineering, University of Maastricht, supervisors: Prof. Dr. Ir. Eric Biessens, Dr. Joël Karel, Dr. Evgueni Smirnov, Dr. Marco Manca and Dr. Zita Soons

3 Survival Analysis

Survival analysis examines and models the time it takes for events to occur and it typically examines the relationship of the survival distribution to covariates [5]. One application of survival analysis that will be focused on is Cox Model and Network-Cox model which was proposed by Zhang [8].

3.1 Cox Model

Cox model, also known as Cox proportional hazards model, is an example to survival models that are used in survival analysis of patient data.

3.2 Network-Cox model

4 Regression

This section describes regression models used for further analysis of mRNA data of patients. Regression models were used in order to compute how essential are given set of attributes (i.e., if patient is smoking or not, blood pressure level, level of diabetes etc.). One obvious problem with given data set was dimensionality of the dataset (524 patients and 20842 genes), which naturally leads to overfitting problem. In statistics, overfitting refers to the cases where number of parameters(in this case, genes) is relatively bigger than number of observations(in this case, patients). One way of dealing with this problem is to reduce size of the data by selecting only those parameters that together can explain or represent the data and continue analysis only on that set of parameters.

4.1 Bootstrapping and ensemble feature selection

4.2 Logistic regression

Logistic regression is a type of probabilistic statistical classification model.

5 Experiments

5.1 Setup of experiments

5.2 Module detection

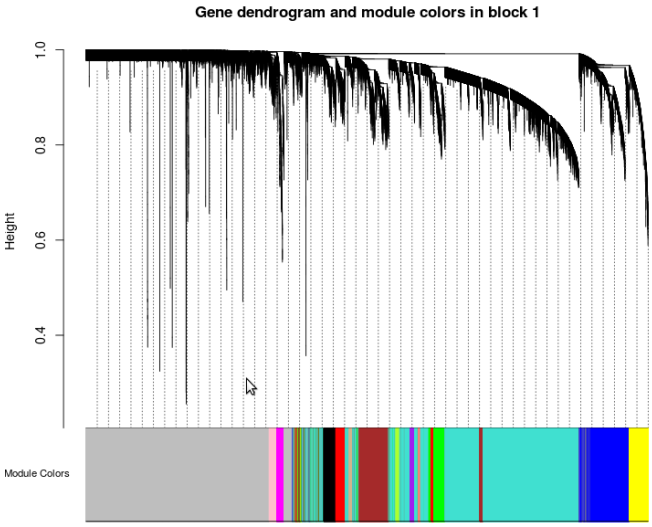

Figure 5.1: Dendrogram representing result of module detection using clustering methods

As it can be seen from figure 5.1, 13 modules were found using R WGCNA Package.

5.3 Results

6 Conclusion

7 Future Work

References

[1] Lingxue Zhang, Seyoung Kim, "Learning Gene Networks under SNP Perturbations Using eQTL Datasets" *http://dx.doi.org/10.1371%2Fjournal.pcbi.1003420*

[2] Langfelder P, Horvath S , "WGCNA: an R package for weighted correlation network analysis". *BMC Bioinformatics 2008, 9:559*

[3] Bin Zhang, Steve Horvath, " A general framework for Weighted Gene Co-Expression Network Analysis". *Statistical Applications in Genetics and Molecular Biology. Volume 4, Issue 1, ISSN (Online) 1544-6115*

[4] Storey JD and Tibshirani R. , "Statistical significance for genome-wide experiments". *Proceedings of the National Academy of Sciences", 100: 9440-9445.*

[5] John Fox, "Cox Proportional-Hazards Regression for Survival Data", *Appendix to An R and S-PLUS Companion to Applied Regression*

[6] Brian S. Everitt and Torsten Hothorn, "CHAPTER 6: Logistic Regression and Generalised Linear Models: Blood Screening, Womens Role in Society, and Colonic Polyps",*"A Handbook of Statistical Analyses Using R"*

[7] William B. King, Coastal Carolina University, "Logistic Regression", *R Tutorials*

[8] Wei Zhang, Takayo Ota, Viji Shridhar, Jeremy R Chien, Baolin Wu, Rui Kuang, "Network-based Survival Analysis Reveals Subnetwork Signatures for Predicting Outcomes of Ovarian Cancer Treatment", *PLoS Comput Biol 9(3):e1002975,2013*

[9] Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods", *PubMed ID: 19942583*

[10] S. Horvath *et al.*, "Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target", *Proc Natl Acad Sci USA 103, 17402-17407 (2006)*

[11] Jia Xue *et al.*, "Transcriptome-Based Network Analysis Reveals a Spectrum Model of Human Macrophage Activation", *Immunity, Volume 40, Issue 2, 274-288*