Analysis of mRNA data of patients with atherosclerosis[1]

Taghi Aliyev

May 26, 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. Recent research concerning atherosclerosis focused on finding new target genes that will lead to better understanding of the disease and generation of new treatments. Although, most of the research done so far focused on systems biology methods [12], standard mapping techniques, random forests, neural networks and some others, lately, also some papers have been published trying to analyse patient data using more complex and statistical methods. Considering successful usage of Weighted Gene Co-expression Network Analysis, Cox regression models in identification of target genes in cancer research[11, 10, 8] lead to the idea of performing experiments and analysis using these techniques in research of cardiovascular diseases.

2 Patient data

Patient data that is used for this study is collected by CTMM and represents mRNA and clinical data of 514 patients. There are 20826 genes in gene data set and each gene is represented with its expression profile over all patients. Clinical data about patients represented time for the event which was used as a variable to predict in case of Cox and Net-Cox regression and also included personal data of patients. Main focus was on gene data set rather than classical approach of using personal patient data as predictors of the cardiovascular event.

[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 Weighted Gene Co-Expression Network

Gene co-expression networks represent the interaction between the nodes and helps to investigate the functionality of genes on system-level. Construction of such a networks is straightforward: Nodes represent the genes and nodes are connected if genes are significantly correlated. This already presents one challenge of picking correct threshold that will define the significance of the connection. One approach proposed by Steve Horvath and his team was the idea of giving weight to the connection between each pair [3]. They achieved this by creating a general framework for "soft" thresholding that gave the weights to pair of genes in weighted network. Parts of network construction and approaches proposed by S. Horvath are explained in next subsections.

3.1 Network construction

Weighted gene co-expression networks are reverse engineering methods for finding the connection and correlation between the genes using genes‘ expression profiles. Their construction is essential in detection of modules and analysis of interconnectivity among the genes represented in patient data. General flowchart for constructing a gene co-expression network is represented in Figure 2.1.

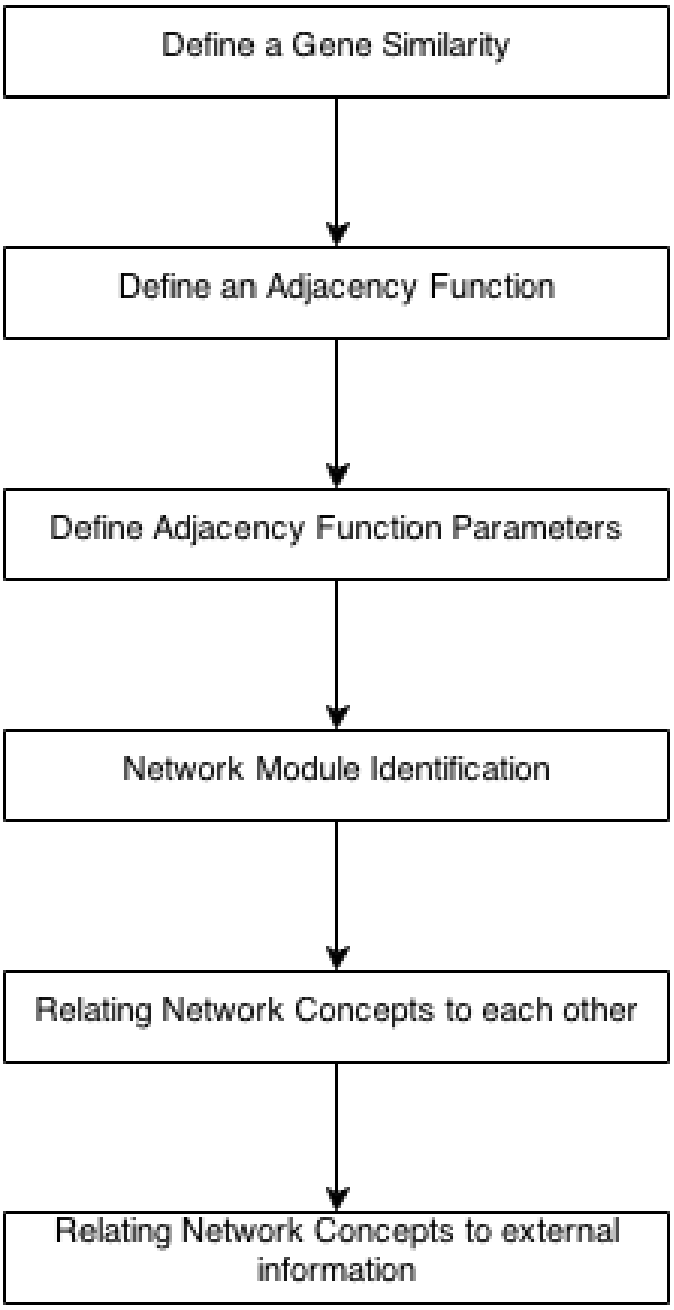

Figure 3.1: Flowchart showing the steps in network construction

As it is pointed out in Figure 2.1, first couple of steps relate to deciding on the measure and the function that will be used to assign weight of correlation between pair of genes in the network. Once the similarity measure and adjacency function is decided upon, next step is to find the best fitting parameters.
It has been shown that in many real networks, as well as random networks, probability that a node is connected with k other nodes decays as a power law $p(k) \sim k^{-\gamma}$ and this statement is the defining property of scale-free networks and topology.

Details of module detection and interconnectivity of genes will be explained in later sections.

3.2 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].

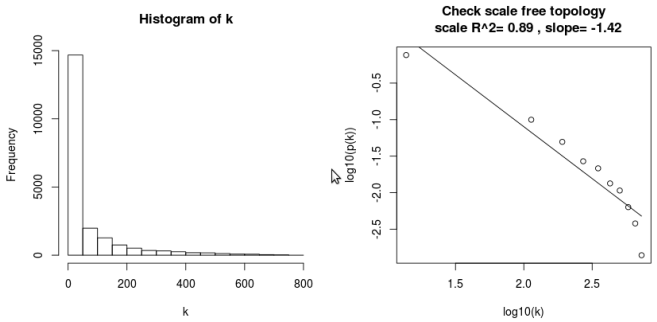

Figure 3.2: Scale-free topology fit plot

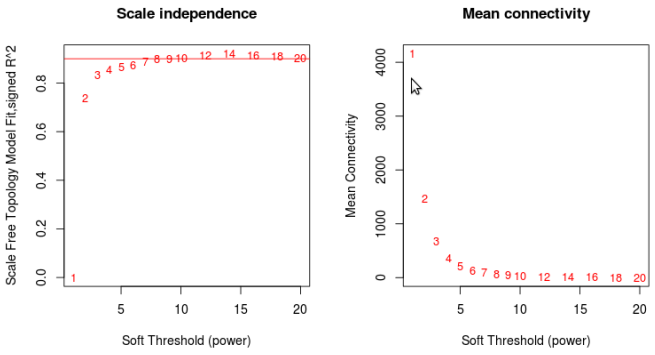

Figure 3.3: Soft-thresholding power selection

3.3 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:

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{min(k_i, k_j) + l - a_{ij}} \tag{3.1}$$

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: $d_{ij}^{\omega} = 1 - \omega_{ij}$.

3.4 Interconnectivity of genes

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].

4 Survival Analysis

4.1 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.

4.2 Bootstrapping and ensemble feature selection

Write about fighting overfitting and big number of genes

4.3 Network-Cox model

5 Experiments

5.1 Setup of experiments

5.2 Gene Ontology Enrichment Analysis

In order to perform Gene Ontology Enrichment Analysis (GOEA) on the gene data set, results of module detection method were used. Each module was analysed separately. In order to perform GOEA, BiNGO plugin on Cytoscape(version 3.1) software was used. Main idea was to get the feeling about modules and processes represented by these modules.

5.3 Transcription Factor Analysis

ToppFun. CTMM data

5.4 MicroRNA Analysis

ToppFun. CTMM data

5.5 Module detection

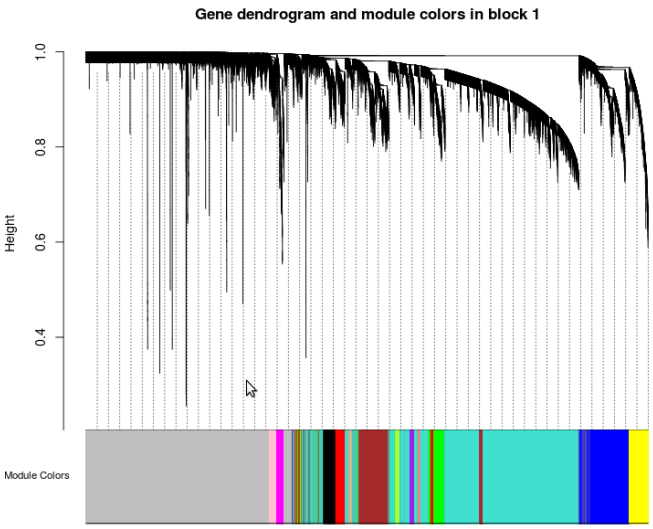

Figure 5.1: Dendrogram representing result of module detection using clustering methods

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

5.6 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*

[12] Stephen A Ramsey, Elizabeth S. Gold, "A systems biology approach to understanding atherosclerosis", *EMBO Mol Med. 2010 March, 2(3) : 79-89*

[13] Albert, R. and Barabasi, A. L., "Emergence of scaling in random networks science", *Science, 286(5439), 509-512*