Analysis of mRNA data of patients with atherosclerosis[1]

Taghi Aliyev

June 23, 2014

Abstract

Atherosclerosis is a specific type of arteriosclerosis, in which artery walls thickens as a result of accumulation of cholesterol, toxins, triglyceride and other irritants and is a result of processes taking place in a complex network of cells. Main focus of this paper is on Weighted Gene Co-expression Network Analysis(WGCNA), Feature Selection, Survival Analysis and Logistic Regression. WGCNA is a widely-used technique for getting a better idea of underlying correlation among the genes. Modules that are obtained as a result from these network analysis can be used to find transcription factors or enriched gene ontologies . In order to find possible candidate genes, Survival analysis and Logistic regression methods are exploited. One problem faced by both methods is the overfitting problem. Feature selection and Network-Cox model are used for dealing with the high dimensionality of the input. Above mentioned methods were able to produce set of genes already associated with the atherosclerosis or its symptoms and also introduced possible new targets, making them interesting alternative to include in future research.

Keywords: Weighted Gene Correlation Network Analysis, Regression, mRNA, Atherosclerosis, Feature Selection, Networks

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. When irritants enter the blood, they may cause the endothelial cell dysfunction. This dysfunction leads to cholesterol, toxins and some other irritants to enter the inner layer of the blood vessels. Sudden reaction of the macrophages and smooth muscle cells, try to cover the damage and reduce the amount of irritants, but as a result, they become the part of the inner layer wall of the blood vessel, so leading to the artery walls thickening and losing flexibility.

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 [1], 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[10, 9, 7] lead to the idea of performing experiments and analysis using these techniques in research of cardiovascular diseases. The goal and main focus of this article is to investigate the application of Weighted Gene Co-expression networks and Cox models in research of cardiovascular diseases.

Weighted Gene Co-expression networks are proven to be a powerful tool in analysis of biological networks [10]. Even though, there exists a framework with implementation of Weighted Gene Co-expression networks [3], there are still questions to be answered by a researcher in order to make best use of the framework. Couple of different analysis tools and techniques are introduced in order to make better sense of the results that Weighted Gene Co-expression framework returns.

One of the problems introduced by the application of Cox models is the overfitting problem and their poor performance on high dimensional data sets. Therefore, one aspect that will be explored is the possible reduction of dimensionality of gene data set through the use of ensemble feature selection and Network-Cox models [8, 7].

The structure of this article is as follows. Section 2 will describe the data set used for this research. In Section 3, introduction will be given to weighted gene co-expression networks. Later in that section, specific processes regarding co-expression networks and their construction will be discussed and explained. Next, sur-

[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

vival analysis and Cox models will be discussed in Section 4. Also in Section 4, couple of enhancements of Cox regression model, namely Network-Cox model(4.3) and Ensemble Feature Selection through the use of linear support vector machines and recursive feature elimination (4.2) are presented to the reader. Section 5 will explain a regression approach, namely logistic regression technique, which measures the relationship between the predictors and the outcome class, which, in most cases, is binary. After this, Section 6 shows different experimental setups and discusses their results. Section 7 summarizes the findings and results of the research. At last, Section 8 goes through some additional enhancements and techniques that might help and are interesting to investigate in the research of cardiovascular diseases.

2 Data

Patient data that is used for this study is collected by CTMM and represents mRNA and clinical data of 461 patients. There are 20826 genes in mRNA data set. These genes are represented with their expression profiles over all patients. Clinical data is composed of personal medical history of patients and the entries representing times till cardiovascular event occurred. Main focus was on gene data set rather than classical approach of using personal patient data as predictors of the cardiovascular event.

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 of genes [2, 3]. They achieved this by creating a general framework for "soft" thresholding that gave the weights to pair of genes in weighted network. Section 3.1 will mention the steps needed for construction of such networks. Sections 3.2, 3.3, 3.4 and 3.5 will delve into these steps more in detail. Section 6, Experiments, will show the result of the Weighted Gene Co-expression Network analysis and Section 7 and 8 will mention possible future work that can be done on these results.

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. Each co-expression network corresponds to an adjacency matrix, which in its own term encodes the correlation between each pair of genes. 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 3.1.

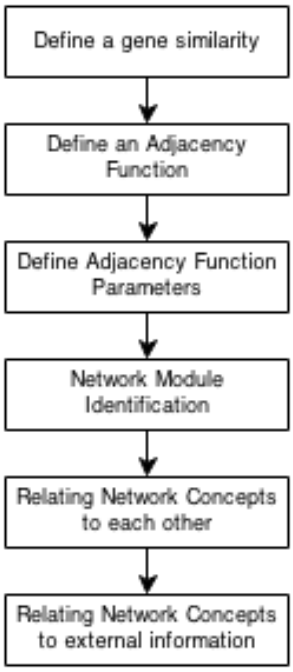

Figure 3.1: Flowchart showing the steps in network construction

As it is pointed out in Figure 3.1, first couple of steps relate to deciding on the measure and the function that will be used to assign weight of correlation between each 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.

3.2 Adjacency function and Similarity measure

First step in gene co-expression network construction is the decision a researches has to make on adjacency function and similarity measure. This similarity measures the level of correlation/agreement between gene expression profiles across all the patients. It is standard to use Pearson correlation as a similarity/co-expression measure between pair of genes in network analysis. Pearson correlation between genes i and j will be denoted as $s_{ij} =\mid \mathrm{cor}(i,j) \mid$. Then, similarity matrix is denoted by $S = [s_{ij}]$.

Next step is to define adjacency matrix using similarity measure defined above. One way of choosing adjacency matrix that will be using "soft" thresholding instead of "hard" thresholding is power adjacency function. The power adjacency function is defined as follows:

$$a_{ij} = \text{power}(s_{ij}, \beta) =\mid s_{ij} \mid^{\beta} \tag{3.1}$$

One important thing to take a look at is that, power adjacency function uses β in order to determine the adjacency value between genes i and j. Choice of β, which is explained in next subsection, is made such that network fits the scale-free topology.

3.3 Selection of soft-thresholding power

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 [11]. It is known that scale-free networks are very heterogeneous and are dominated by a few hub genes that have high connectivity value.

As already mentioned, one has to select the soft-threshold power, β, that is used for the computation of the adjacency function. One way to achieve this goal is to look at different values of β and consider only the ones that lead to a network satisfying scale-free topology. High R^2 value is a indicator of approximate fitting. Figure 3.2 shows the relation between candidate values for β and R^2 values. It can be seen that, lowest power for which scale-free topology criterion is satisfied is 7 (Roughly, $R^2 \geq 0.9$). β value of 7 is chosen as a soft-thresholding power for adjacency function for this data set and is the power used throughout the rest of this article.

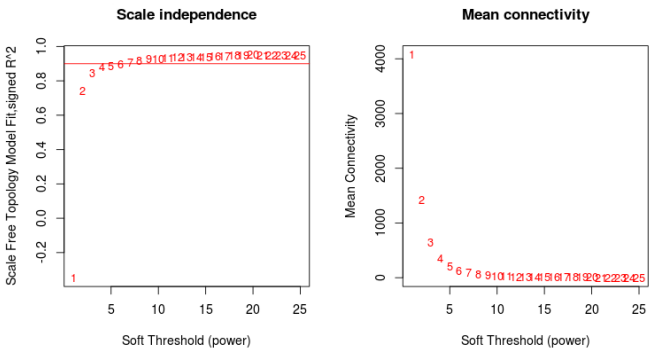

Figure 3.2: Soft-thresholding power selection

One other way of showing that β value of approximately satisfies scale-free topology is to plot $\log_1 0(p(k))$ versus $\log_1 0(k)$. Higher R^2 value and straighter line is indicative of the scale-free topology. As it can be seen from Figure 3.3, our data set actually satisfies scale-free topology when β value is set at 8.

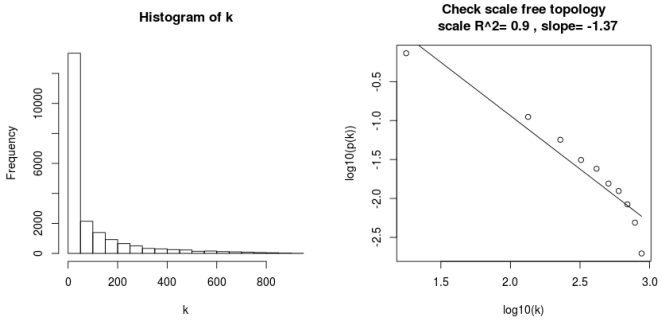

Figure 3.3: Scale-free topology fit plot

3.4 Module Detection

Important point of co-expression networks is to find set of genes that are tightly connected to each other. Modules represent this kind of set of genes. In weighted gene co-expression networks gene dissimilarity measure in conjunction with a clustering method identifies modules. Topological Overlap Mapping based dissimilarity measure is used for module detection in this article as it was found to result in biologically meaningful modules [12]. The topological overlap of two genes reflects their relative interconnectedness.

In order to define Topological Overlap Mapping (TOM), first connectivity of a gene in a weighted network should be defined. Following is the connectivity measure of a gene:

$$k_i = \Sigma_{j=1}^{n} a_{ij} \tag{3.2}$$

Ravasz and colleagues proposed the following topological overlap matrix for unweighted gene co-expression networks, which is then adapted for weighted co-expression networks by Steve Horvath and his team:

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{min(k_i, k_j) + l - a_{ij}} \tag{3.3}$$

where $l_{ij} = \Sigma_u a_{iu} a_{uj}$, and k_i is the gene connectivity defined by Formula 3.2. The topological overlap matrix $\Omega = [\omega_{ij}]$ is a similarity measure as its entries defined by Formula 3.3, are non-negative and symmetric. Module detection in weighted gene co-expression networks makes use of dissimilarity measure and in order to get TOM-based dissimilarity measure, TOM-based similarity measure is subtracted from 1. In other words, the topological overlap based dissimalirty measure is defined by the following formula:

$$d_{ij}^{\omega} = 1 - \omega_{ij} \tag{3.4}$$

In order to group genes together into the modules, one general approach which is also used in this article, is the usage of hierarchical clustering coupled with TOM-based dissimilarity d_{ij}^{ω}.

4 Survival Analysis

In most research projects, one interesting aspect to analyse is the correlation between some covariates/predictors and an event of interest. Survival analysis is the discipline that aims at creating models that is able to inspect the connection between predictors and time-to-event. This type of analysis is mostly applied in health sciences from which the name "survival analysis" is derived, as in most cases event of interest is the time till death. In case of cardiovascular pathology and research, event of interest is time till the cardiovascular event happens. Some examples of cardiovascular events are heart attacks, infarction and cardiac arrest.

Couple of notations should be made clear about survival analysis before proceeding forward. One component that is of interest in survival analysis, as already mentioned, is the distribution of survival times. Commonly used representation of distribution of survival times is the *hazard function*, which assesses the instantaneous risk of demise at given time t, conditional on survival time:

$$h(t) = \lim_{\Delta t \leftarrow 0} \frac{Pr[(t \leq T < t + \Delta t) | T \geq t]}{\Delta t} = \frac{f(t)}{S(t)} \quad (4.1)$$

where T represents survival time, $f(t)$ is the probability density function and $S(t)$ is the complement of the distribution function. One example of this, is when one has a constant hazard, $h(t) = v$, then this implies an exponential distribution with density function $f(t) = ve^{-vt}$.

Common aspect of data sets analysed using survival analysis is called *censoring*. Most common format of censoring represents observations for which trial period might have expired or patient might have left the trial before the event of interest happening, which is also called right-censoring. In more general terms, censored data is continuous and means event of interest did not happen for one or another reason. Censoring complicates the likelihood function and modelling process. Censored data might be either right-censored, left-censored or both. An observation is *Left-censored* if its initial time is unknown. If an observation is both left and right-censored, then it is called *interval-censoring*.

Main focus of this article is to analyse how good of predictors of time till event are some set of genes. One application of survival analysis that will be focused on is Cox Proportional-Hazards model. Different approaches of choosing set of predictor genes exists. Later in the section, enhancement of simple Cox model, namely Network-based Cox model proposed by Zhang [7], and one approach of choosing set of predictor genes, namely Ensemble Feature Selection, are explained.

4.1 Cox Proportional-Hazards Model

As already mentioned, examination of interest is the relation between predictors and time-to-event. Most commonly, this examination entails a linear-like model for the log-hazard. One way of writing a parametric model based on the exponential distribution is as following:

$$\log h_i(t) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \quad (4.2)$$

or, equivalently,

$$h_i(t) = \exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}) \quad (4.3)$$

Equations 4.2 and 4.3 show this relation as linear for log hazard and multiplicative for the hazard itself. In these equations, i represents the observation and $x's$ are the covariates/predictors. In this article, i represents patients and $x's$ are the predictor genes used for creation of the model. The constant α represents log-baseline hazard, since $logh_i(t) = \alpha$ when all the covariates equal to zero.

Cox Model is different from generic case as it does not specify function for baseline hazard $\alpha(t) = h_0(t)$. So, Cox model can be expressed using following formula which looks similar to equation of parametric model on the exponential distribution(Equation 4.2):

$$\log h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \quad (4.4)$$

or, equivalently,

$$h_i(t) = h_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}) \quad (4.5)$$

Cox models are *semi-parametric*, as baseline hazard α might enter the model in any form, whereas covariates enter the model linearly. Additionally, Cox models are *proportional-hazards model*. One way of showing this property is to analyse two different observations, namely i and j, that have different values for their covariates. Let η_i and η_j represent linear predictors for each observation accordingly. In other words,

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \quad (4.6)$$

,and, equivalently,

$$\eta_j = \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_k x_{jk} \quad (4.7)$$

Then, following two equations are representation of their hazard models:

$$h_i(t) = h_0(t) e^{\eta_i} \quad (4.8)$$

$$h_j(t) = h_0(t)e^{\eta_j} \tag{4.9}$$

Using Equations 4.8 and 4.9, the hazard ratio of two observations can be find:

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)e^{\eta_i}}{h_0(t)e^{\eta_j}} = \frac{e^{\eta_i}}{e^{\eta_j}} \tag{4.10}$$

From Equation 4.10, it can be seen that hazard ratio between two different observations, is independent of time t, which implies that Cox models are actually proportional-hazards model.

Cox model offers couple of advantages with it. Main advantage of Cox models is that, baseline hazard should not be specified, which helps to deal with having arbitrary and possibly incorrect assumptions about the form of baseline hazard. As Cox explains in his paper about Cox proportional-hazards model [13], baseline hazard can be estimated using the method of partial likelihood.

Although, Cox model is a popular choice for survival analysis, it has its own disadvantages. One disadvantage of Cox models is their lack of being able to deal with high dimensionality of input data. It leads and is linked with the problem more commonly known as *overfitting*. Two different approaches of dealing with this problem are introduced in next subsections.

4.2 Feature Selection

As already mentioned, Cox models have trouble dealing with input data that has high dimensionality and high number of attributes compared to number of samples. One approach adapted from Abeel et al. [14], is to preprocess list of genes by using different feature selection methods in order to reduce the dimensionality of data set. This approach leads to smaller set of genes selected as being predictors and representatives of whole gene set. Feature selection approach is a common technique used in Machine Learning applications but has drawn attention in field of Bio-Informatics only lately.

Selection of set of genes out of huge number of genes for purposes of regression analysis can be looked at as being an attribute selection problem for classification tasks. Main aim of attribute selection is to find a small set of features (genes) that best explains the given data set.

In order to apply feature selection, one of the first points to take into account is the characteristics needed from feature selection method. In case of analysis of large number of genes compared to samples/patients, it is important that technique to be used scales well to high-dimensional spaces and it is preferable if the space complexity of the parameters to be estimated by the selection method depends on the number of the observations rather than the number of the features. Also, another aspect of an interest is to be able to easily interpret the results of selection or classification. Linear Support Vector Machines(SVM) and Recursive Feature Elimination (RFE) are an interesting choice to exploit, as they contain all the characteristics and the qualities needed from the selection method for purposes of this research.

Support Vector Machines

As already mentioned, a linear SVM is a classification model for which the influence of each gene is explicitly available and number of features to estimate depends mainly on the number of patients, rather than on the number of genes, which is an important feature because of the small patient-to-gene ratio (approximately, 1:40). Furthermore, SVMs are known to scale well to high-dimensional spaces, their performance grow by having more samples to train on and SVMs have shown state-of-the-art performance in computational biology problems [15].

For classification purposes, support vector machines aim at finding a hyperplane that separates the input space with maximum margin. Figure 4.1, show an example of SVM models. In figure 4.1, H1 does not separate classes, H2 does separate classes but SVM model that will be trained in this case is the model corresponding to the H3 as, it both separates the classes completely and does that by achieving the maximum margin between the classes.

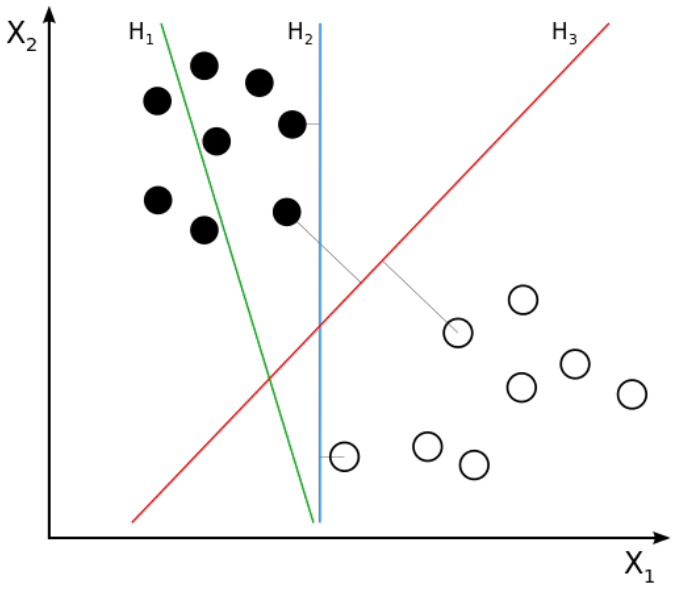

Figure 4.1: Simple example of an SVM model in two-dimensional space

Recursive Feature Elimination

Weights of the dimensions (in the case of this research, genes) show the contribution of that specific dimension to the hyperplane produced by SVM. These weights can be used in order to rank the features from most impor-

tant to least important. Recursive feature elimination exploits this ranking and adopts backward elimination strategy in order to iteratively eliminate least important genes. Starting from full set of genes, a linear SVM is estimated and least important genes are eliminated. During next iteration, a linear SVM is estimated using only remaining set of features and again fraction of genes is removed from set of all genes. This procedure continues until desired number of genes is reached or there are no more genes to eliminate.

One important parameter for RFE is the fraction of the genes, E, to be eliminated every iteration of feature elimination method. This parameter plays essential role in time complexity of the feature elimination. Reducing the value of E increases time complexity of the algorithm. In order to rank all the features at once, E should be set at 100%. For purposes of this research, E is set at 20%, meaning $\frac{1}{5}^{\text{th}}$ of the remaining genes is eliminated at the end of each iteration. Genes are ranked by the absolute value of their weight during the iteration they have been eliminated at. Genes removed at later iterations are put on top of the ranking.

Sensitivity analysis of fraction E has been analysed in biomarker detection paper by Abeel [14] and it has been shown that choice of this parameter does not affect overall stability of feature selection. Thus, sensitivity analysis is disregarded in this article.

Ensemble Feature Selection and Bootstrapping
SVM and RFE are feature selection techniques that can provide good results most of the times. In order to increase the stability of feature selection, power of these two methods are combined into ensemble feature selection technique. Ensemble feature selection makes use of both SVM and RFE in order to give ranking to set of features separately and returns aggregated result of these two methods.

Another enhancements added to feature selection is the bootstrapping of samples. Bootstrapping is the common approach used in Machine Learning for reduction of variance. It creates diverse training sets (with replacement) by including part of the original data set in order to train SVM and RFE models on. Bootstrapping used in this research is simple random sample with replacement, meaning there is a chance for an individual to appear more than once in the subset.

More formally, following notation can be made about feature selectors and results obtained from those: Let FS_i represent i^{th} feature selector. Assumption can be made on that, each FS_i will return set of ranking values for each gene, which in its turn can be noted more formally as $\text{fsr}_{i,k} = (f^1_{i,k}, ..., f^N_{i,k}$, where $f^j_{i,k}$ denotes the rank of feature j at the bootstrap k using feature selector i and N denotes the number of features. In case of this research, aggregation is done on results obtained from RFE, which internally uses SVM. Different RFE classifiers have been applied on different bootstrap samples and once the results are obtained, they are aggregated.

For aggregating the results obtained by bootstrapping the training data, one scheme proposed is the *Complete linear aggregation*. Complete linear aggregation gives equal weight to different bootstrap samples and sums up the rankings of a given feature over all bootstrapped samples in order to get final result. Aggregated rank of all the features obtained from selector FS_i using complete linear aggregation can be computed as follows:

$$f_i = (\Sigma^t_{k=1}\text{fsr}^1_{i,k}, ..., \Sigma^t_{k=1}\text{fsr}^N_{i,k}) \tag{4.11}$$

, where t represents the number of bootstrap samples used for ensemble feature selection.

4.3 Network-Cox model

As already mentioned, Cox regression performs and generalizes poorly when the dimensionality of input data is high. Most common techniques used in prior is the introduction of so-called L1 and L2 - norm penalties to Cox regression models. One novel approach proposed by Wei Zhang [7] is the combination gene co-expression networks and cox models. Network-Cox models exploits the information about modular relations among genes which has been mostly ignored in previous survival analysis research and the fact that groups of genes are co-expressed under certain conditions.

Gene co-expression network used in network-cox model is similar to the one computed in Section 3. It uses Pearson correlation coefficients for computing the correlation level among genes using gene expression dataset.

In order to include network information into cox regression models, following constraint is added to general formulation of Cox model mentioned in Formula 4.3:

$$l_{\text{pen}}(\beta, h_0) = l(\beta, h_0) - \frac{1}{2}\lambda\beta'[(1-\alpha)L + \alpha I]\beta, \tag{4.12}$$

where $\lambda\beta'[(1-\alpha)L + \alpha I]\beta$ represents the Laplacian constraint which is a encoding of a prior knowledge obtained from network. L in equation 4.12, is a positive semidefinite matrix derived from network information, I is an identity matrix, λ is a parameter controlling the weighting between total likelihood and the network constraint. α is another parameter controlling the weighting between network matrix and identity matrix in network constraint. Decision on the values of these variables are subject to Section 5.

The objective function defined by equation 4.12, can be solved by alternating the optimization of β and baseline-hazard function $h_0(t)$. For optimization of β, Newton-Raphson method is used. Before, explaining the

Newton-Raphson method for optimization, first and second derivatives of the objective function with respect to β should be defined. First derivative of objective function is as follows:

$$\frac{\delta l_{\text{pen}}(\beta, h_0)}{\delta \beta} = X'\Delta - \lambda\Gamma\beta, \tag{4.13}$$

where $\Delta = \delta - exp(X'\beta)H_0(t)$

Second derivative of the objective function is defined as follows:

$$\frac{\delta^2 l_{\text{pen}}(\beta, h_0)}{\delta\beta\delta\beta} = -X'DX - \lambda\Gamma, \tag{4.14}$$

where D is the diagonal matrix with $D_{ii} = exp(X_i'\beta)H_0(t_i)$. Now, given this information, Network-cox model can be solved using following algorithm proposed by Zhang [7].

Data: **X - gene expression profiles**, **S - normalized graph weight matrix**
Result: **Returns** β, weights of each feature/gene
1. **Initialization :** $\beta = 0$**; Compute** $L = I - S$
2. **Do** until convergence
(a) **Do** Newton-Raphson iteration
i. Compute the first derivative $\frac{\delta l_{\text{pen}}(\beta, h_0)}{\delta\beta}$
ii. Compute the second derivative $\frac{\delta^2 l_{\text{pen}}(\beta, h_0)}{\delta\beta\delta\beta}$
iii.Update $\beta = \beta - l''_{pen}(\beta, h_0)^{-1} l'_{pen}(\beta, h_0)$
(b) Update $h_0(t_i) = 1/\Sigma_{j \in R(t_i)} exp(X_j'\beta)$
3. **Return** β
Algorithm 1: Solution algorithm for Net-Cox models

As it can be seen, from third phase of Newton-Raphson iteration, in order to optimize β, inverse of second order partial derivatives, also called the Hessian matrix, of the objective function should be computed, which is a time consuming operation. An alternative to this approach is to reduce the covariant space from p to n, where p represents amount of genes and n represents amount of patients. This approach relates the singular value decomposition which in its turn exploits the low rank of the gene expression matrix **X**. Using simple calculus, one can see that from

$$\frac{\delta l_{\text{pen}}(\beta, h_0)}{\delta \beta} = X'\Delta - \lambda\Gamma\beta = 0, \tag{4.15}$$

it can be implied that $\beta = \Gamma^{-1}X'\eta$ for some value of η. So, using this information, objective function of Network-Cox model can be written as:

$$l_{\text{pen}}(\eta, h_0) = \Sigma_{i=1}^{n}(-exp(Z_i'\eta)H_0(t_i) + \delta_i[\log(h_0(t_i)) + Z_i'\eta]) - \frac{1}{2}\lambda\eta' Z\eta. \tag{4.16}$$

In equation 4.16, $Z = X\Gamma^{-1}X'$. This equation is the dual form of the equation 4.12, and it is obvious that those two equations are equal but the problem dimension is reduced from p to n. So, using the Algorithm 1 and equations 4.15 and 4.16, it is possible to optimize β and solve Net-cox model. Figure 4.2, shows the overall flow, structure and the construction of the Network-Cox model explained earlier in the subsection.

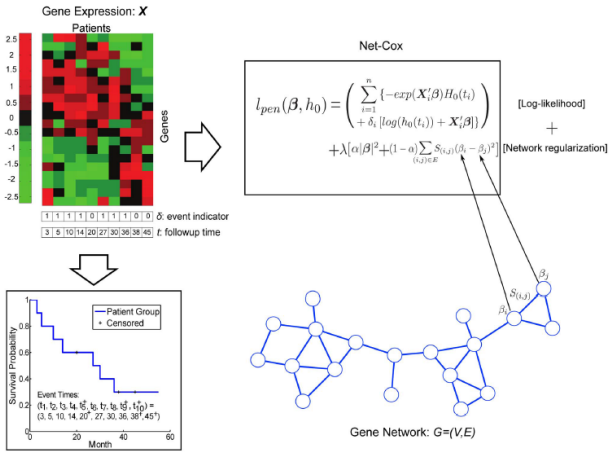

Figure 4.2: Overview and the overall flow of Network-Cox model, taken from [7]

5 Regression

Survival analysis mainly deals with the prediction of an event happening or not and if it does, time it takes for an event to happen given some covariates. In the cardiovascular research, it is equally interesting to look into other variables/symptoms that might be a cause of a disease or an event. Some examples of these symptoms are diabetes, blood pressure and dyslipidemia. As these variables are not survival data, different approach should be used for exploring the effects of different genes or gene sets. One method that is used in this paper is logistic regression as most of the clinical patient history data is binary and logistic regression is the common technique used for evaluating the effects and significance of genes as the predictors of the outcome class. Next subsection explains logistic regression in more detail.

5.1 Logistic regression

Logistic regression is a statistical model that fits the non-linear model representing the relation between predictor

and discrete variable. Discrete variable in case of this research, is the outcome class of interest(diabetes, dyslipidemia etc.) and predictors are genes and their expression profiles over all patients. In logistic regression analysis, outcome class is binary(true/false, 1/0, yes/no etc.) and main purpose of analysis is to assess the effect of multiple explanatory variables on the outcome class. As values of diabetes and dyslipidemia are logistic (yes/no), it is only fitting to exploit logistic regression.

As name already suggests, logistic regression uses logistic function for prediction step. Logistic function is defined as follows:

$$F(t) = \frac{e^t}{e^t + 1} \tag{5.1}$$

, where t can be viewed as being a linear function of an predictor x. Then, logistic function can be written as:

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{5.2}$$

Equation 5.2 determines the probability assigned by logistic regression for a probability of having a success or not. Figure 5.1, shows the plot of simple logistic function. Logistic function is extra useful because it can take any value between negative infinity and positive infinity as its input, and return the value between 0 and 1, which in its turn is interpretable as a probability. $F(x)$ illustrates that the probability of outcome class being equal to a case is equal to the value of the logistic function of the linear regression expression.

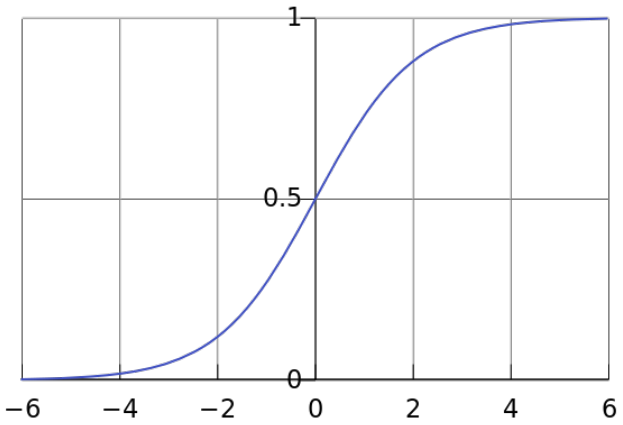

Figure 5.1: Simple plot of a logistic function

Equations shown above represent the logistic function for logistic regression where there is only one predictor. It is possible to adopt those equations for a case of multiple explanatory variables. In case of multiple explanatory variables, $\beta_0 + \beta_1 x$ is replaced by $\beta_0 + \beta_1 x_1 + ... + \beta_i x_i$. Here, β_0 is the value of the criterion when the predictor values are all zero and $\beta_1, ...\beta_i$ represents the contribution weights of each feature/predictor to the whole model. In general, for linear regression, β values are found using maximum likelihood estimation. In case of logistic regression, it is not possible to find a closed-form expression that maximized the likelihood value. Because of this, iterative methods for numerical approximations are used. Most common techniques used for numerical estimations is Newton's method which is explained elsewhere [16].

Logistic regression is combined with feature selection explained in Section 4, in order to fight the same kind of overfitting problem explained in Cox models. Predictor genes that are examined using logistic regression is the set of genes ranked highest by Feature selection methods. Section 6 will delve into different uses of logistic regression.

6 Experiments

For purposes of experimentation, R programming language combined with Bio-conductor and WGCNA [3] package were used for Module Detection and network construction.

Next, main focus of experiments was shifted from correlation among genes to correlation between genes and clinical data of patients. Java-ml [17] Ensemble Feature selection tool combined with Matlab implementation of Network-Cox regression were used as tools for finding possible target genes predicting the medical patient data. Examples of medical data include diabetes, dyslipidemia and stroke.

Results of each of these experiments is discussed in next subsections.

6.1 Module detection

For purposes of module detection, WGCNA package available for R programming language has been used. After the discussion with specialists in the field, it was agreed to set minimal allowable size of the module to be at 80. Figure 5.1, shows the dendrogram of the module detection. As it can be seen from Figure 5.1, there were 13 modules detected by Weighted Gene Co-expression Network analysis. Grey module represents the genes that were not included in any of the modules.

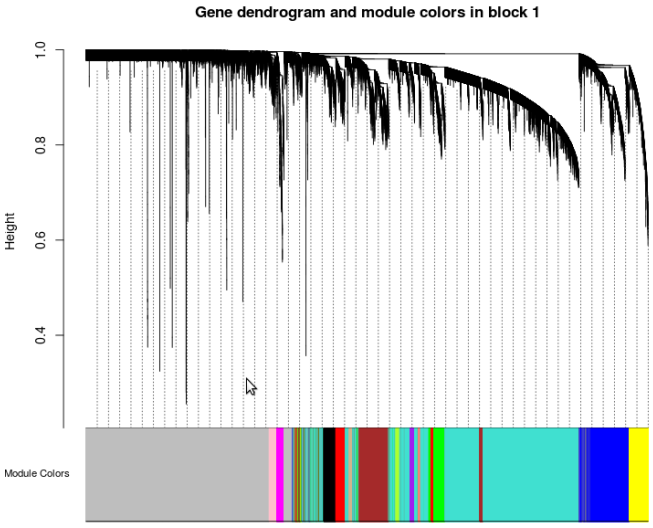

Figure 6.1: Module detection dendrogram

These modules represent possible biological connections among genes, as already mentioned in Section 3. To get better understanding of these results, further analysis can be applied on modules separately, which is out of scope of this paper.

6.2 Feature Selection

First experiment was performed in order to rank the features considering their contribution and weight at prediction of different clinical data. One variable of interest in feature selection experiments was diabetes. Diabetes feature is a binary one, where 1 represents patient having diabetes and 0 otherwise. As mentioned in Section 4, Ensemble feature selection technique was applied on the data set. For this experiment, 10 bootstrapped data sets were used to train RFE model on them and the results of these 10 feature selectors were aggregated. Following table shows top 10 genes (Associated diseases have been retrieved from gene card website):

Ranking	Gene ID	Already related diseases/disorders
1.	UTS2	Atherosclerosis and cardiovascular diseases
2.	PRKAR1A	Hypocalcemia
3.	IPO8	Alcoholism
4.	HIST1H4C	Cancer and non-hodgkin lymphoma
5.	GRHPR	Renal failure
6.	HLA-C	Psoriasis and Leukemia
7.	CLEC4F	Unknown
8.	KIAA0226L	Cervical intraepithelial neoplasia and esophagitis
9.	NQO2	Agranulocytosis
10.	TIMM22	Unknown

From results, it can be seen that, top ranked gene *UTS2* has already been associated with cardiovascular diseases and atherosclerosis. Additionally, most of the genes in top 10 have been shown to have relevance with atherosclerosis. For example, *PRKAR1A* has been associated with Hypocalcemia, which in its turn has chronic renal failure as its cause. Renal failures are known to be of importance for diagnosis of atherosclerosis and is a factor taken into account by physicians and doctors. Furthermore, *GRHPR* gene directly relates to renal failure. *KIAA0226L* is associated with esophagitis which is shown to lead to higher tendency for atherosclerosis, diabetes and heart failure [18].

Additionally to the facts mentioned above, there are also two genes in top 10, namely *CLEC4F* and *TIMM22*, are not known to be associated with any disease, making their possible future analysis more interesting.

Next to the diabetes, ensemble feature selection with the same parameters as mentioned above was applied to find the most important genes contributing to the prediction of dyslipidemia. Same as in case for diabetes, top 10 genes are shown below:

Ranking	Gene ID	Associated disease/disorder

Above mentioned results show that feature selection on its own can determine important genes in the study of cardiovascular diseases. Next to the ability of finding target genes already associated with disease, it also offers possible new target genes that should be further analysed. These conclusions suggest that it is an interesting alternative to include the feature selection techniques in future research.

6.3 Network-Cox Regression

One of the main parameters analysed in cardiovascular, as well as cancer, research is the survival data of the patients. This data represents the time it took for an event to happen, if that event happens. Details about survival data has been explained in earlier sections. Network-Cox model has been applied on the data set in order to assess the significance and importance of the genes on the prediction process. Following are the most significant genes obtained from Network-Cox regression:

Ranking	Gene ID	p-value

6.4 Logistic Regression

Next step on analysis of the genes that show high rankings during feature selection step, is the assessment of their statistical significance at the prediction of the diabetes and dyslipidemia. For this experiment, top 50 genes were assessed as being predictors of the diabetes and dyslipidemia.

Main purpose of regression analysis was to further identify more important genes or get the results already obtained by other scholars in their research.

First variable of interest as already mentioned was the diabetes among the patients. Top 50 genes obtained from ensemble feature selection were used for this experiment. As table gets very big, only more interesting genes, their p-values and their rankings obtained from ensemble feature selection are shown below:

Gene ID	p-value	Ranking
UTS2	0.016807	1
SYT10	9.28e-05	39
IPO8	0.042734	3
HIST1H4C	0.014305	4
ID1	0.000123	32
NQO2	0.005885	9
MMP9	0.000490	25
FPR1	0.000875	14
ENC1	0.005429	17
CLEC4F	0.045308	7

As it can be seen from these results, all of the above mentioned genes are significant for the prediction of the diabetes. Among them, *NQO2,UTS2,HIST1H4C* were among the top 10 genes and their relevance to overall atherosclerosis has already been mentioned in Feature selection experiments. *SYT10* is one gene that should be taken into account as it has very low p-value, meaning higher significance of this gene and also added fact of it being in top 50 genes chosen by feature selection technique and not being associated with any kind of disease or disorder, makes *SYT10* gene an interesting target for future research. Furthermore, *CLEC4F* has not been associated with any diseases so far, but it both appeared in most chosen 10 genes and it shows good enough level of contribution to logistic model, as its p-value is smaller than 0.05, which is a value adopted and used by almost every research group and scholar for assessment of significance. One more interesting gene in above table is *MMP9*, because it has been associated with atherosclerosis, inflammation and cardiovascular diseases. *ID1* has been associated with neovascularization which is related to atherosclerosis. These facts further prove that, feature selection and logistic regression are interesting alternatives and their results, especially the genes that have not been associated with atherosclerosis, should further be analysed.

Next experiment was the evaluation of significance of top 50 genes for the case where prediction class (independent variable) is dyslipidemia. As in case of diabetes, this variable is binary (0 - no, 1 - yes). Following table shows the genes with lowest p-values, their gene IDs and their rankings obtained from feature selection experiment.

Gene ID	p-value	Ranking

7 Conclusion

As a conclusion, Weighted gene co-expression networks are still an interesting approach that is used by many others. Modules detected by WGCNA package could have been further analysed for enriched pathways, ontologies or transcription factors, but this kind of analysis was out of scope of this paper.

Feature selection techniques combined with regression models are seem as an interesting alternative for future research, as their results present the genes already related to cardiovascular diseases. Next to these genes, results also suggest to further analyse some genes, namely *SYT10* and *CLEC4F*. These methods have not been used extensively by other scholars, making them more appealing to exploit. Additionally, Network-Cox models presented interesting approach for dealing with dimensionality problems of simple Cox models by including co-expression network information in their model. Similar to logistic regression and feature selection, Network-Cox model analysis offered new target genes that had not been associated with cardiovascular diseases, namely *gene1* and *gene2*. Also, Network-Cox models were able to point out the genes that have already been found by other research groups only by knowing their expression profiles, which indicate their power and makes them an extra option in future research. One thing that should be kept in mind that, these genes are possible candidate genes that can turn out to be target genes relating to the disease, but more complex analysis should be done on their functional and structural models.

8 Future Work

As already mentioned, one possible further analysis for better understanding of the biological processes happening among genes, Gene Ontology Enrichment Analysis, Transcription Factor analysis and MicroRNA analysis could have been performed on the modules detected by WGCNA modules. Furthermore, intramodular connectivity analysis can be included for finding central hub genes of each module, which has been shown to be of interest for research [10].

Additionally, Feature Selection experiments could be further extended to include different techniques in order to improve results and possibly find more target genes for future research. Also, Ensemble Feature Selection, Logistic Regression and Network-Cox models explained in this paper can be used to examine the relation between genes and different variables that are of interest for cardiovascular research that were not examined in this paper. Multi-class regression might have been added for the assessment of the link between the genes and multiple variables of interest at the same time.

Once results are obtained from above mentioned techniques, more in-detail analysis of candidates for being a target gene should be performed in order to prove their relevance to the atherosclerosis and its symptoms.

References

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

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

[5] 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"*

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

[7] 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*

[8] 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*

[9] 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)*

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

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

[12] Ravasz, E. *et al.*, "Hierarchical organization of modularity in metabolic networks", *Science, 297(Aug.), 1151-1155*

[13] Cox, D.R., "Regression models and Life Tables (with Discussion)",*Journal of the Royal Statistical Society, Series B 34:187-220*

[14] Thomas Abeel *et al.*, "Robust biomarker identification for cancer diagnosis with ensemble feature selection method", *Bioinformatics, Vol. 26, no.3, 2010, p. 392-398*

[15] Ben-Hur, A. *et al.*, "Support vector machines and kernels for computational biology", *PLoS Computational Biology, 4, e1000173, 2008*

[16] Kendall E.Atkinson, "An Introduction to Numerical Analysis", *John Wiley & Sons, Inc., 1989*

[17] Thomas Abeel, Y.V. de Peer & Y. Saeys, "Java-ML : A Machine Learning Library", *Journal of Machine Learning Research, 2009, 10, 931-934*

[18] Jau-Jiuan Sheu *et al.*, "Reflux Esophagitis and the Risk of Stroke in Young Adults: A 1-Year Population-Based Follow-Up Study", *Stroke - Journal of the American Heart Association*