A FRAMEWORK FOR GENE CO-EXPRESSION NETWORK ANALYSIS OF LUNG CANCER

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ERHAN AKDEMİR

IN PARTIAL FULFILLMENT OF REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
BIOINFORMATICS

SEPTEMBER 2013

Approval of the Graduate School of Informatics:

A FRAMEWORK FOR GENE CO-EXPRESSION NETWORK ANALYSIS OF LUNG CANCER

Submitted by **ERHAN AKDEMİR** in partial fulfillment of the requirements for the degree of **Master of Science in Bioinformatics, Middle East Technical University** by**,**

Prof. Dr. Nazife Baykal _____
Director, **Informatics Institute, METU**

Assist. Prof. Dr. Yeşim Aydın Son _____
Head of Department, **Medical Informatics, METU**

Assoc. Prof. Dr. Tolga Can _____
Supervisor, **Computer Engineering, METU**

Assist. Prof. Dr. Yeşim Aydın Son _____
Co-supervisor, **Health Informatics, METU**

Examining Committee Members:

Assist. Prof. Dr. Aybar Can Acar _____
Health Informatics, METU

Assoc. Prof. Dr. Tolga Can _____
Computer Engineering, METU

Assist. Prof. Dr. Yeşim Aydın Son _____
Health Informatics, METU

Dr. Levent Çarkacıoğlu _____
T.C. Merkez Bankası

Assoc. Prof. Dr. Vilda Purutçuoğlu Gazi _____
Statistics, METU

Date:04.09.2013

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: Erhan Akdemir

Signature:

ABSTRACT

A FRAMEWORK FOR GENE CO-EXPRESSION NETWORK ANALYSIS OF LUNG CANCER

Akdemir, Erhan
M.Sc., Bioinformatics Program
Supervisor: Assoc. Prof. Dr. Tolga Can
Co-supervisor: Assist. Prof. Dr. Yeşim Aydın Son
September 2013, 47 pages

Construction method of a gene co-expression network (GCN) is crucial in medical research aiming to reveal disease related genes. Applied similarity measure and selection of edges that represent significantly co-expressed gene pairs in the network affect directly the elements of a network and so the list of prioritized genes. Pearson correlation coefficient is a commonly used similarity measure to quantify co-expressions of genes due to its simplicity and performance compared to many complex methods. However, it is affected by outliers and may not be reliable with low sample size. On the other hand, selection of edges is generally based on an arbitrary cutoff which makes networks subjective. For a more standard and accurate analysis, reliability of a similarity measure must be ensured as well as an objective threshold determination for the selection of edges. Here, a framework is proposed for the construction of GCNs that combines a reliability measure, stability, previously applied to Pearson correlation coefficient to detect general co-expression differences between healthy and cancer state and an automatic threshold selection method, Random Matrix Theory for a standard network construction. The proposed framework was applied to lung adenocarcinoma. In the analysis part, genes were prioritized by using changes in topological and neighborhood properties of nodes in control and disease networks. Differential co-expressions of known interacting proteins and intrinsically disordered proteins were also analyzed. Results suggest that co-expression networks are topologically spoke-like and control samples are in transition phase from healthy to cancer. Thus, effects of stability on finding general co-expression differences between cancer and healthy states could not be assessed. Prioritized genes by both proposed and control methods are mostly enriched to relevant processes reflect the changes in cellular machinery as a result of a state shift to cancer and may reveal dynamical features of transition of cells to cancer state with a further detailed analysis. Furthermore, some genes were prioritized related with cilia which may have roles early phases of transition.

Keywords: gene co-expression networks, stability of correlation, Random Matrix Theory, lung adenocarcinoma

ÖZ

AKCİĞER KANSERİNİN GEN ORTAK İFADE ANALİZİ İÇİN BİR YÖNTEM

Akdemir, Erhan
Yüksek Lisans, Biyoenformatik Programı
Tez yöneticisi: Doç. Dr. Tolga Can
Yardımcı tez yöneticisi: Yar. Doç. Dr. Yeşim Aydın Son
Eylül 2013, 47 sayfa

Hastalıkla ilgili genlerin ortaya çıkarılmasını amaçlayan medikal araştırmalarda gen ortak ifade ağlarının oluşturulma biçimi önemlidir. Uygulanan benzerlik ölçüsü ve ağdaki anlamlı ortak ifadeye sahip gen çiftlerini temsil eden bağlantıların seçimi ağın öğeleri ve öne çıkarılan genleri doğrudan etkiler. Pearson korelasyon katsayısı basitliği ve diğer bir çok karmaşık yönteme karşı üstünlüğü nedeniyle genlerin ortak ifadesinin sayısallaştırılmasında sıkça kullanılan bir benzerlik ölçüsüdür. Ancak aykırı değerlerden etkilenir ve örnek sayısının azlığında güvenilir olmayabilir. Diğer yandan, genelde keyfi belirlenmiş bir eşiğin üzerindeki ortak ifadeler ağa seçilir, bu da ağı öznel yapar. Daha standart ve doğru bir analiz için benzerlik ölçüsünün güvenilirliği ve bağlantıların objektif olarak belirlenen bir eşiğe göre seçilmesi sağlanmalıdır. Burada sağlıklı ve kanser durumları arasındaki genel eş ifade farklılıklarını bulmak için kullanılabilecek, daha önce Pearson korelasyon katsayısına uygulanmış bir kararlılık ölçüsü olan *korelasyon kararlılığı* ile otomatik bir ağ oluşturma yöntemi olan *rastlantısal matris teorisi*nin birleştirildiği bir ağ oluşturma sistemi önerilmektedir. Önerilen sistem akciğer adenokanserine uygulanmıştır. Analiz aşamasında genler kontrol ve hastalık ağlarındaki topolojik ve komşuluk özelliklerinin değişimine bağlı olarak öne çıkartılmıştır. Bilinen protein-protein etkileşimleri ve düzensiz proteinlerin eş ifadeleri de analiz edilmiştir. Sonuçlar eş-ifade ağlarının tekerlek benzeri bir yapıda, kontrol örneklerinin ise sağlıklıdan kanser durumuna geçiş fazında olduğunu önermektedir. Bu yüzden kanser ve sağlıklı durumları arasındaki genel eş ifade farklılıkları değerlendirilememiştir. Hem önerilen hem de kontrol amacıyla kullanılan yöntemle öne çıkarılan genler çoğunlukla geçiş fazını yansıtan hücre etkinlikleriyle ilgilidir ve daha sonra yapılacak derin analizlerle faz geçişinin dinamik özelliklerini ortaya çıkartabilir. Dahası faz geçişinin erken dönemlerinde etkisi olabilecek siliya ile ilgili bazı genler öne çıkartılmıştır.

Anahtar sözcükler: gen ortak ifade ağları, korelasyon kararlılığı, rastlantısal matris teorisi, akciğer adenokanseri

To my family

ACKNOWLEDGEMENTS

I want to thank my supervisors Assist. Prof. Dr. Yeşim Aydın Son and Assoc. Prof. Dr. Tolga Can for their support during the course of the thesis.
I also want to thank my friends Olcay Öztürk and Burak Demiralay for their help on statistical and theoretical part of the thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

GCN: Gene Co-expression Network
PPI: Protein-Protein Interactions
S: Stability of Correlation
BC: Betweenness Centrality
CC: Clustering Coefficient
NC: Neighborhood change
NS: Neighborhood similarity

CHAPTER 1

INTRODUCTION

System level understanding of the biological phenomena is a crucial step through answering questions about life and developing new solutions to everyday life problems such as design of microbial factories for production of substances or new therapeutic approaches to complex diseases. Emergence of high throughput technologies has brought valuable opportunities in this way by providing an overall picture of biological entities in cells. Although technical problems (e.g. noise, number of samples, computational restrictions) still exist, enormous amount of data we have is a big source for the new aspects of biological information. Studies focused on biological networks and integration of different data types in recent years are examples of efforts for gaining new insights from present data [1-4].

Gene co-expression study of microarray data is a way of understanding co-regulation patterns of gene clusters under certain condition. Differential co-expression network analysis of genes between disease and non-disease states is popular in medical research and can reveal disease related functional interactions of proteins in a network [5 -7]. However, problems in network construction restrict the analysis and cause incompleteness of the knowledge about disease mechanisms.

In the case of co-expression networks, one problem is quantification of pairwise relationships between expressions of genes. Since collecting samples from an individual in a time course manner is not possible in many cases, extracting dynamical and partial relationships is not easy. To quantify the level of pairwise co-expression between genes, Pearson correlation coefficient is a common similarity measure used in disease related studies especially in one individual-one sample cases. Its reason is that Pearson correlation coefficient is computationally easy to be calculated and also outperforms many other existing methods. However, it is sensitive to outliers.

Selection of significant pairwise co-expressions, each of which corresponds to an edge in the network, is another construction problem. Traditional approaches are based on an arbitrary cutoff selected to be high enough to avoid random noise. It makes constructed networks subjective rather than objective and results in missing some true co-expressions. At this point, it is worth to note that high Pearson correlation coefficient between expression levels does not always mean a functional relationship between proteins and vice versa [8].

In this thesis, a framework for construction and differential analysis of co-expression networks between two conditions is proposed and was implemented to lung adenocarcinoma. In the framework, a stability measure is used to detect general co-expression differences between control and disease states by filtering unstable correlations across samples.

Thresholds for correlation and stability measures were calculated with an application of Random Matrix Theory, which has been previously shown to be effective in differentiating system specific correlations from random ones [9]. In the analysis part, genes were prioritized by using changes in betweenness centrality, clustering coefficient and neighborhood properties of nodes between control and disease networks. Known interacting proteins as well as intrinsically disordered ones were also integrated into analysis to benefit from their differential co-expression pattern information.

CHAPTER 2

BACKGROUND

2.1 Biological Background

A living system can be defined with its three characteristics: a body that distinguished itself from the environment, a metabolism, a process by which it can convert resources from the environment into building blocks so that it can maintain and build itself, and inheritable information passed to its offspring. In the following sections, basic concepts related with primary molecules in a cell as a living system and lung adenocarcinoma are explained.

2.1.1 DNA and Gene

DNA is the hereditary material of all known living organisms and many viruses. It stores the information necessary for the synthesis of molecules that are parts of the structure or metabolism of a cell and passed to offspring through cell division. DNA is a double helix structure consisting of two polymers of nucleotides. Each nucleotide is composed of a base (guanine (G), adenine (A), thymine (T), and cytosine (C)) as well as a backbone made of alternating sugars (deoxyribose) and phosphate groups, with the bases attached to the sugars. In the structure of DNA, a nucleotide is bound to another one in the opposite strand with hydrogen bonds between their bases forming A-T and G-C pairs.

A region on a DNA strand coding for a polypeptide is called gene. A polypeptide is an amino acid chain synthesized through gene expression and becomes a protein or a subunit of a protein complex after a number of structural modifications.

2.1.2 Proteins and Intrinsic Disorder

Proteins are the biological macromolecules that catalyze the metabolic reactions in a cell, transport other molecules and are the building blocks of cellular structures. A protein is the resulting three-dimensional (tertiary) structure of a single polypeptide chain (monomeric proteins) or more than one polypeptide chains (polymeric proteins) after post-translational modifications following gene expression.

According to the traditional protein structure paradigm, a unique, stable tertiary structure was thought to define the function of a protein and be required by it to function correctly. However, it has been found that this is not the case for intrinsically unstructured proteins. They are often referred to as naturally unfolded proteins or disordered proteins and remain functional despite the lack of a stable tertiary structure when the protein exists as an isolated polypeptide chain [10, 11].

In cells, physical interactions between proteins are crucial for many distinct processes. Many important complex processes such as DNA replication and signal transduction are carried out by a large number of interacting proteins. These interactions can be long standing to form protein complexes or molecular machines in the case of DNA replication or transient to transport or modify the interacting partner in the case of signal transduction.

Intrinsically disordered proteins are enriched in regulatory and signaling functions showing that disordered regions provide interactions with many proteins in multiple pathways [12]. Several disordered proteins were shown to be associated with diseases such as cancer, cardiovascular disease, diabetes, and others [13].

2.1.3 Cancer and Lung Adenocarcinoma

In multicellular organisms, the regulatory mechanisms that control cell division sometimes become irregular and uncontrolled growth begins resulting in loss of tissue stability. An example of this situation is the formation of a structure called tumor or neoplasm. Neoplasms are classified as benign or malignant based on their spreading to other parts of the body. In cancer, malignant tumors are formed and often invade neighboring tissues and even distant parts of the body through circulatory system.

A cancer originates from cells of epithelial tissue that have secretory properties is classified as adenocarcinoma.

Lung adenocarcinoma is the most common histological subtype of lung cancer, which is the leading cause of cancer death worldwide. It is also the most common histological subtype in women, Asians, and never-smokers and has surpassed squamous carcinoma in many countries [14, 15]. Lung adenocarcinoma is increasingly recognized as a clinically and molecularly heterogeneous disease by recent reclassifications based on pathology and patient survival, observed prognostic gene expression signature profiles, as well as the increasing number of clinical trials demonstrating targeted treatments defined by molecular subtypes such as EGFR, KRAS, BRAF, and ERBB2 mutations and EML4-ALK fusions [16].

2.2 Gene Co-expression Networks

Analysis of genome-wide gene expression has become an ordinary approach with the emergence of microarray technology to find co-expression patterns of genes in different biological conditions. Gene co-expression networks have become popular after the integration of network concepts to biology. In a GCN, nodes represent the genes, while edges between nodes mean that co-expression of genes represented by those nodes are significant. Significant edges can be defined by various approaches after quantifying the co-expression level of two genes by a similarity measure.

In the subsequent sections, concepts related with the construction of GCNs and framework is explained.

2.2.1 Pearson Correlation Coefficient and Stability

In target specific co-expression studies, small number of samples may be useful since biological meaning of co-expression is straightforward. However, inferring genome-wide correlations is hard in the presence of small data set and large number of parameters. This difficulty even increases when sample related bias is high such as smoking condition of individuals in the case of lung cancer. Thus, the reliability of similarity measure becomes more crucial in disease related network studies.

Pearson correlation coefficient is a measure of linear correlation (dependence) between two variables, giving a value between +1 and -1, where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation. It is a commonly used correlation measure in co-expression network analysis due to its simplicity. It also outperforms many other complex methods. However, it is not suitable for non-linear dependencies and is sensitive to outliers. Kinoshita proposes a reliability measure for Pearson correlation coefficient, stability, for co-expression analyses in the presence of large number of samples. The idea behind the stability is that how correlations of expressions change after removing the most contributing samples.

In the study performed with *Arabidopsis thaliana* data, it was shown that high correlations with low stability values are fragile or reflect weak functional relationships [17]. In the thesis, we propose that stability can also be used to detect general co-expression differences between two cellular states by filtering unstable correlations resulted from a few samples in the presence of low sample size. Calculation of stability is explained in detail in Section 3.2.1.

2.2.2 Random Matrix Theory

Random Matrix Theory, initially proposed by Wigner and Dyson in the 1960s for studying the spectrum of complex nuclei, is a powerful approach for identifying and modeling phase transitions associated with disorder and noise in statistical physics and materials science.

It is known that only a small portion of genes are co-expressed under a certain condition. Thus, a correlation matrix of gene expressions is a combination of true co-expressions with high correlations and weak correlations representing noise. According to Random Matrix Theory, the nearest neighbor spacing distribution (NNSD) of eigenvalues of a non-random symmetric matrix follows Poisson statistics, whereas it appears as a Gaussian orthogonal ensemble (GOE) distribution if the matrix is random [9]. Thus, a correlation matrix of gene expressions, in which correlations under a certain threshold are made zero, is non-random and starts to deviate to a random matrix with the accumulation of non-zero weak correlations by decreasing the threshold. NNSD change of eigenvalues of a correlation matrix was recently proposed to find the point of this transition which would become the threshold for significant correlations of gene expressions. Moreover, networks constructed by this method were found to be sensitive in detecting the threshold [9] and topologically robust [18].

However, since Pearson correlation coefficient is affected by outliers in the presence of low sample size, the threshold calculated by RMT would be misleading. In the framework, we

applied RMT to both Pearson correlation coefficient and stability to filter out the unreliable gene pairs having correlation above Pearson threshold with stability below stability threshold.

2.2.3 Node Comparison Measures

Significance of a gene in a network can be defined by various topological properties. Here, well known measures, betweenness centrality and clustering coefficient, were used in the prioritization of genes. Each measure reflects a different biological aspect of a gene. Betweenness centrality is a measure of a node's centrality in a network. It is equal to the number of shortest paths from all vertices to all others that pass through that node. Genes with high betweenness centrality tend to connect functional modules and pass information. Clustering coefficient is a measure of degree to which nodes in a graph tend to cluster together. Genes with high clustering coefficient are generally subunits of protein complexes or members of functional modules. Neighborhood change is also used to prioritize genes and is the normalized ratio of total number of its neighbors to number of its common neighbors between two networks to define genes most differentially co-expressed.

CHAPTER 3

MATERIALS AND METHODS

3.1 Data Sets

In this section, data sets used in this thesis and their preprocessing are described. The proposed framework was applied to gene expression data of human lung adenocarcinoma and a matching adjacent non-tumor lung tissue. Then, interacting and disordered proteins were separately mapped to the resulting GCNs.

3.1.1 Gene Expression Data

Gene expression data, previously processed by log2-transformation and Robust Spline Normalization (RSN) [16], was downloaded from Gene Expression Omnibus (GEO) in NCBI [19]. One lung adenocarcinoma (3023_T, GSM813507) and one non-tumor lung sample (3035_N, GSM813519) were discarded from 116 gene expression profiles because they actually are not from the same individual. This resulted in 57 lung adenocarcinoma and 57 adjacent non-tumor lung tissue profiles. Probes representing transcribed loci (e.g. HS.388528), predicted genes (e.g. LOC441782), and open reading frames (e.g. C17ORF77) were removed from data due to both computational and certainty concerns. Expressions of genes represented by more than one probe were averaged and samples were separated into two groups for control and disease network construction. Resulting data sets were composed of 18456 gene expressions and 57 samples each.

3.1.2 Protein-Protein Interaction Data

Latest human binary protein interactions (Release 9) were downloaded from Human Protein Reference Database (HPRD) [20]. HPRD contains annotations of human proteins based on experimental evidence from the literature. This includes PPIs as well as information about post-translational modifications, subcellular localization, protein domain architecture, tissue expression and association with human diseases [21].

Interactions of proteins were extracted from binary interactions and PPI network was constructed in R by using igraph package. Resulting network is composed of 36976interactions between 9499 proteins.

3.1.3 Intrinsically Disordered Proteins Data

Latest release of intrinsically disordered proteins (Release 6.02) was downloaded from The Database of Protein Disorder (DisProt) [22]. DisProt is a curated database that provides information about proteins that lack fixed 3D structure in their putatively native states, either in their entirety or in part.

IDs of intrinsically disordered human proteins were extracted from downloaded file including FASTA protein sequences containing disordered region identifiers of different organisms. The number of intrinsically disordered human proteins was 235.

3.2 Co-expression Network Construction

Construction of a co-expression network can be divided into two phases: quantifying the co-expressions of genes and selection of edges that represent significant co-expressions into the network. In the thesis, Pearson correlation coefficient was used to quantify co-expressions for control and disease expressions. Then, stability value of each correlation was calculated. For each co-expression data set, Random Matrix Theory was applied to find the related correlation and stability thresholds separately. Pairs of genes whose correlation and stability is greater than corresponding thresholds, r_{th} and s_{th}, were represented by the edges in the networks.

3.2.1 Quantification of Co-expressions

In the framework, co-expression level of each gene pair was quantified by Pearson correlation coefficient by

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

3.2.2 Calculation of Correlation Stabilities

To calculate stability values of correlations, principal component analysis was performed in n dimensional sample space and n principal components with orthogonal basis vector were obtained. Inner products of each n dimensional expression vector (E_p) and orthogonal basis vector gives the projected expression values e_p:

$$e_{p,j} = E_p \cdot r_j^{PC}$$

Pearson's correlations in the PC space without the first *i* PCs between probe *p* and *q* were calculated for the projected expression valuesby

$$cor_i = \frac{\sum_{j=i+1}^{n}(e_{pj} - \mu_{pi})(e_{qj} - \mu_{qi})}{(n-i-1)\sigma_{pi}\sigma_{qi}} (i = 0, 1, \dots, 10)$$

Stability of correlation (*S*) is defined by

$$S = \frac{\sum_{i=0}^{N}(\max\{cor_i, 0\})}{(N+1) \times cor_{max}}$$

where $cor_i$is the correlation without the first *i* PCs, $cor_{max}$is their maximum value (*i*=0…*N*), and *N* = 10 was used in this thesis.

3.2.3 Calculation of Thresholds

Thresholds for Pearson correlation coefficient and stability values of expression data were calculated by a Java program, RMT Co-expression, which applies Random Matrix Theory to construct gene co-expression network [23]. The software first constructs, a gene expression correlation matrix M, whose elements are pairwise Pearson correlation coefficients in the range of (-1.0, 1.0). If there are missing values in the expression files, only the experiments that both genes have values are used to calculate Pearson correlation.

Then, a series of correlation matrices are constructed using different cutoff values. If the absolute value of an element in the original correlation matrix is less than the selected cutoff, it is set to 0. Eigenvalues of each correlation matrix are obtained by direct diagonalization of the matrix. Standard spectral unfolding techniques [24] are applied to have a constant density of eigenvalues and subsequently the nearest neighbor spacing distribution P(s), which is employed to describe the fluctuation of eigenvalues of the correlation matrix. Chi square test is used to determine two critical threshold values, rl at which P(s) starts to deviate from GOE at a confidence level of p = 0.001, and rh at which P(s) follows the Poisson distribution at a confidence level of p = 0.001.

The critical point rh is chosen to be the threshold used for constructing the gene co-expression network [9].

Starting threshold was set to 0.99. Default value of subtraction in each step (0.001) was used in the calculation of both Pearson Correlation coefficient and stability thresholds.

3.3 Analysis

In this section, definition of parameters used to compare networks topologically, node comparison measures used to prioritized genes, and tools used to analyze prioritized genes.

3.3.1 Network Comparison Parameters

The following parameters were used for topological comparison of the networks:

Network *diameter*, (*d*), is the distance between the two vertices which are furthest from each other.

The *average shortest path length*, (*l*), is defined as the mean distance between each two vertices of a network, being the distance between any two vertices the number of edges along the shortest path connecting them.

The *degree distribution*, ($P(k)$), gives the probability that a randomly selected node of a network has degree *k*, i.e. that it is connected to *k* other different vertices.

The *mean clustering coefficient*, (*C*), is related to the meaning of *clustering coefficient of a vertex*, which, in turn, is defined as the ratio between the number of connections existing among its neighbors and the maximal number of edges that can exist among them.

The *Degree exponent*,γ, is the exponent of power law distribution and was calculated by "power.law.fit" function of igraph in R.

3.3.2 Node Comparison Measures

To prioritize the genes related with lung adenocarcinoma, betweenness centrality and clustering coefficient and neighborhood change were calculated for each common node between control and disease networks.

Betweenness centrality of a node is the number of shortest paths from all nodes to all others that pass through that node and calculated by

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of paths that pass through v .

Clustering coefficient is the proportion of links between the neighbors of a node divided by the number of links that could possibly exist between them and calculated by

$$C_i = \frac{2|\{e_{jk}: v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}$$

in an undirected graph, where k_i is the number of neighbors of node v_i , $|\{e_{jk}: v_j, v_k \in N_i, e_{jk} \in E\}|$ is the number of edges between nodes within its neighborhood. Note that the number of edges that can exist within the neighborhood is:

$$\frac{k_i(k_i - 1)}{2}$$

Betweenness centrality and clustering coefficient values were calculated by default functions of igraph package in R.

Neighborhood similarity (*NS*) of a node was calculated by the formula:

$$NS = \frac{\#(A \cap B)}{\#(A \cup B)} = \frac{1}{fa} + \frac{1}{fb} - 1$$

where A and B are the edges of a node in control and disease networks respectively, f_a=(#($A \cap B$))/(#A) and f_b=(#($A \cap B$))/(#B).

3.3.3. Analysis Tools

Names and functions of genes were obtained by *genecards* website [25]. Prioritized genes were enriched by Functional Annotation Clustering tool of David bioinformatics database [26]. *Biological process* GO term data and default parameters (similarity term overlap = 3; similarity thresholds = 0.50; initial group membership = 3; final group membership = 3; multiple linkage threshold = 0.50; enrichment threshold = 1) were used for clustering.
Graphical representations of the networks were obtained in Cytoscape software with its Force-directed (*unweighted)* layout option.

CHAPTER 4

RESULTS AND DISCUSSION

As previously described, the framework applies a reliability measure, stability, to Pearson correlation coefficient to quantify the stable dependencies of pairwise co-expressions under a biological state and uses Random Matrix Theory (RMT) to calculate a threshold for these quantified dependencies in the construction of gene co-expression networks (GCN). Firstly, GCNs were constructed by using only Pearson correlation coefficient with RMT to reveal the effects of using stability in combination with Pearson correlation coefficient. Secondly, GCNs by combining stability with Pearson correlation coefficient were constructed as the proposed method. Resulting networks are designated as R (control/disease) network and RS (control/disease) network respectively, for simplicity, where R stands for Pearson correlation coefficient (r) and S stands for stability.

In the analysis part, genes were prioritized with both control and proposed method with respect to betweenness centrality, clustering coefficient and neighborhood changes of common nodes (i.e. genes) between control and disease networks. In addition to gene prioritization, co-expressions of the known set of both interacting and intrinsically disordered proteins were analyzed to see their role in cancer.

4.1 Topological Properties of the Networks

Biological networks share certain topological properties with many other naturally occurring networks. They are scale-free, disassortative and have small-world property [27]. We calculated the basic network measures used to quantify these characteristics. Table 4.1.1 shows thresholds calculated with RMT GeneNet software, number of nodes and edges and values of topological measures for each network constructed by control or proposed method. Calculations of measures were performed in R using igraph package. Values in parenthesis are the averages of a thousand random networks with equal node and edge numbers to each of corresponding co-expression network by using “erdost.renyi.game” function of the same package.

Table 1: Topological properties of the networks

	R Networks		RS Networks	
Quantity	**Control**	**Disease**	**Control**	**Disease**
r threshold	0.8971	0.7511	0.8971	0.7511
s threshold	-	-	0.7081	0.7281
Nodes	1087	3598	635	2414
Edges	14401	28753	3831	14958
d	20(4)	37(5)	9(5.01)	30(5.18)
l	4.86(2.49)	11.53(3.25)	2.89(3)	8.16(3.38)
C*	0.41(0.024)	0.37(0.004)	0.33(0.015)	0.38(0.005)
max deg	172(45.5)	186(33.2)	75(23.4)	165(27.3)
γ	1.83(1.86)	2(1.74)	2(1.74)	2(1.73)

d: network diameter; *l*: average path length; *C*: mean clustering coefficient; *max deg*: maximum degree; γ: exponent of power law fit. Values parenthesis are the mean values of parameters in corresponding random networks. *: C values of isolates (nodes with degree zero or one) were treated as zero.

In scale free networks, most of the nodes have only a few links and connected together by a few nodes with very large number of links. This property is characterized by power-law degree distribution, that is, the probability that a chosen node has exactly k links follows $P(k) \sim k^{-\gamma}$ where γ is the degree exponent. As seen in Figures 4.1.2, 4.1.2, 4.1.3, and 4.1.4, all networks are scale free.

Many real networks has a degree exponent between 2 and 3. Networks except from R control network have degree exponent of their power-law fit γ=2, a property of hub-and-spoke networks with the largest hub being in contact with a large fraction of all nodes. This spoke like structure of modules in biological networks has been suggested in a recent study as opposed to hierarchical network model [28]. On the other hand R control network has a smaller degree exponent. A degree exponent less than two is explained by partial duplication model in real networks [29].

Another common feature of real networks is that any two nodes can connect with a few paths. A short average path length is an indicator of this small-world property. Biological networks are generally ultra small with average path lengths between 2 and 3 [27]. Here, networks except from RS control network all have higher average path lengths than expected as their diameters. However, they all have much more higher mean clustering coefficients than those of their random counterparts, an indication of modular structure.

Modules in biological networks generally connect each other with a few links called disassortative property. High diameter and average path lengths suggest that highly clustered modules are connected to each other with long paths in a spoke-like manner in both disease networks.

Relatively low difference between average clustering coefficients of control networks and their random counterparts suggest a distortion in modular structure by lose of disassortativity with shorter multiple paths. In RS control network, average path length and diameter are low due to presence of two similar size subnetworks instead of one very big subnetwork. From the number of nodes in RS control network, it can easily be seen that a big portion of genes were filtered out by stability.

The reason of this might be an indication of a transition in cellular machinery of adjacent cells from non-cancer to cancer which cannot be detected with histological examination. Phase transitions are characterized by extreme sensitivity to small perturbations. Thus, differential effects of tumor cell signals on adjacent cells depending on their distance to or location with respect to tumorous tissue in different samples are expected to be high and explain high number of genes filtrated out by stability as compared to disease samples. This may also explain the partial duplication model of R control network. Modules connected to each other with multiple shorter paths that may be regulatory genes active in different phases of the transition.

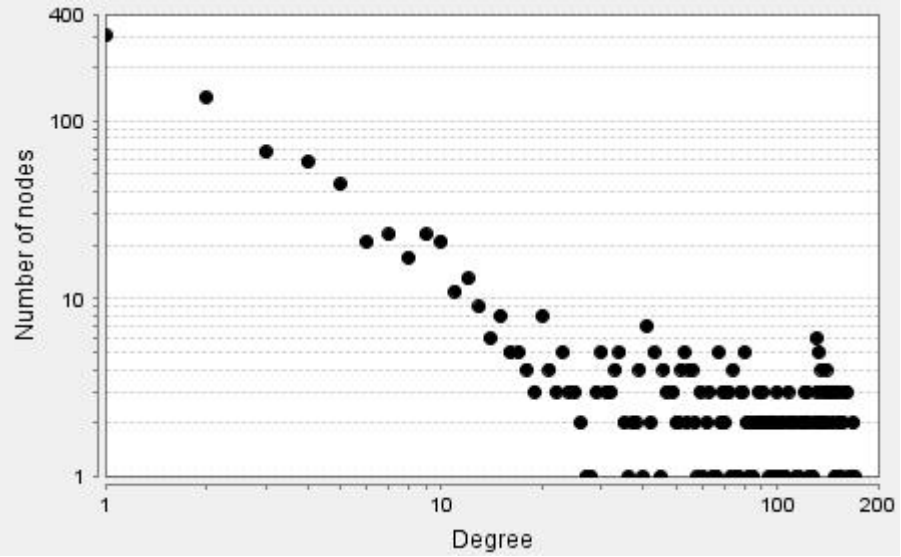

Figure 1: Degree distribution of R control network.

The x-axis shows number of neighbors of a node. The y-axis shows the number of nodes having that number of neighbors.

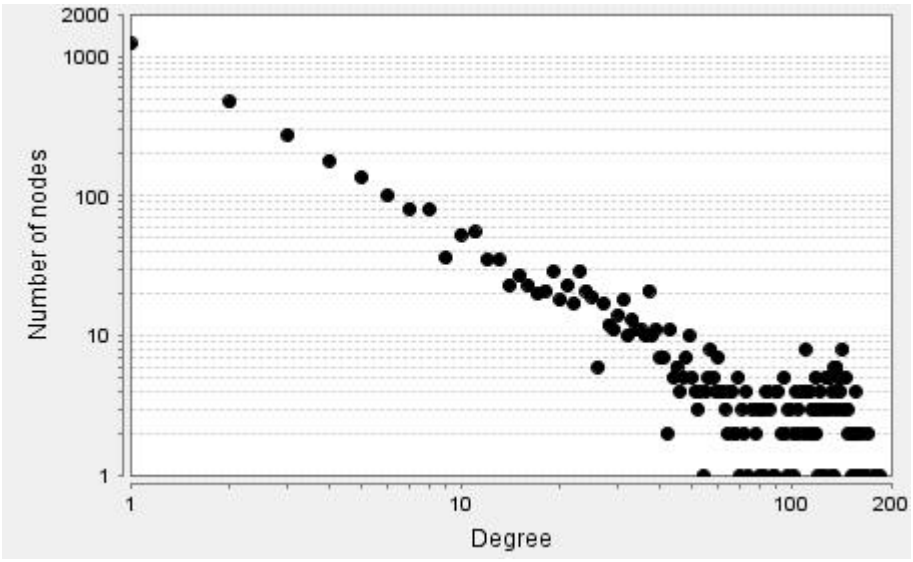

Figure 2: Degree distribution of R disease network

The x-axis shows number of neighbors of a node. The y-axis shows the number of nodes having that number of neighbors.

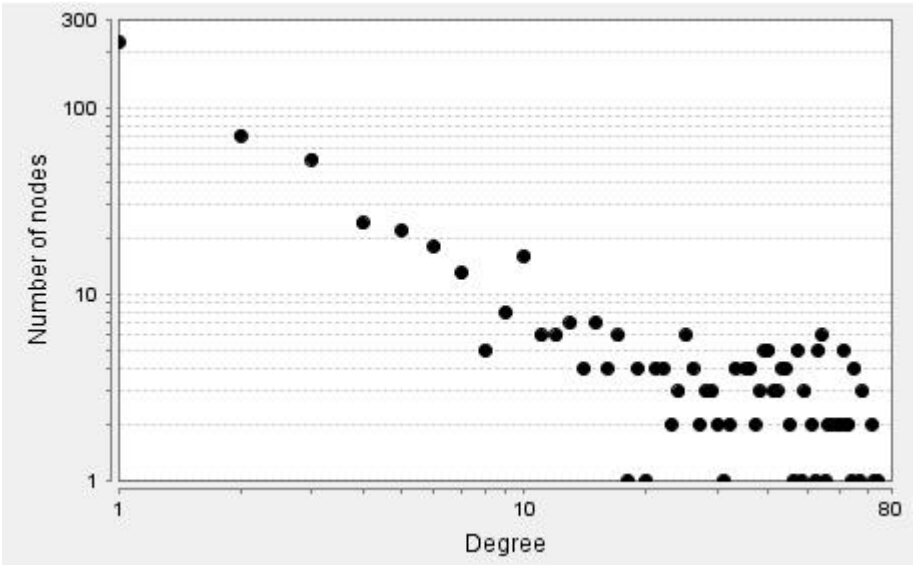

Figure 3: Degree distribution of RS control network

The x-axis shows number of neighbors of a node. The y-axis shows the number of nodes having that number of neighbors.

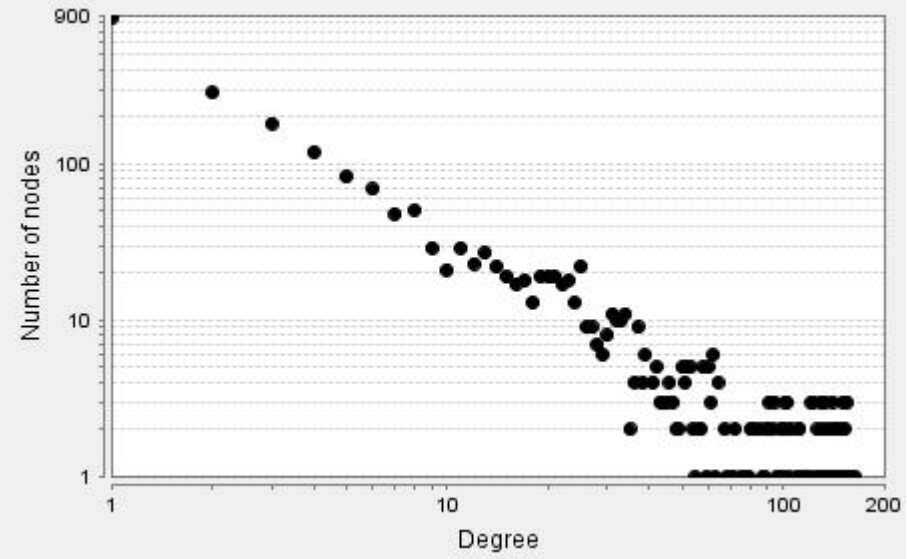

Figure 4: Degree distribution of RS disease network

The x-axis shows number of neighbors of a node. The y-axis shows the number of nodes having that number of neighbors.

Graphical representations of the networks were prepared by using Cytoscape with Force-directed (*unweighted)* layout option. Figure 4.1.5, 4.1.6, 4.1.7, and 4.1.8 are the images of the networks.

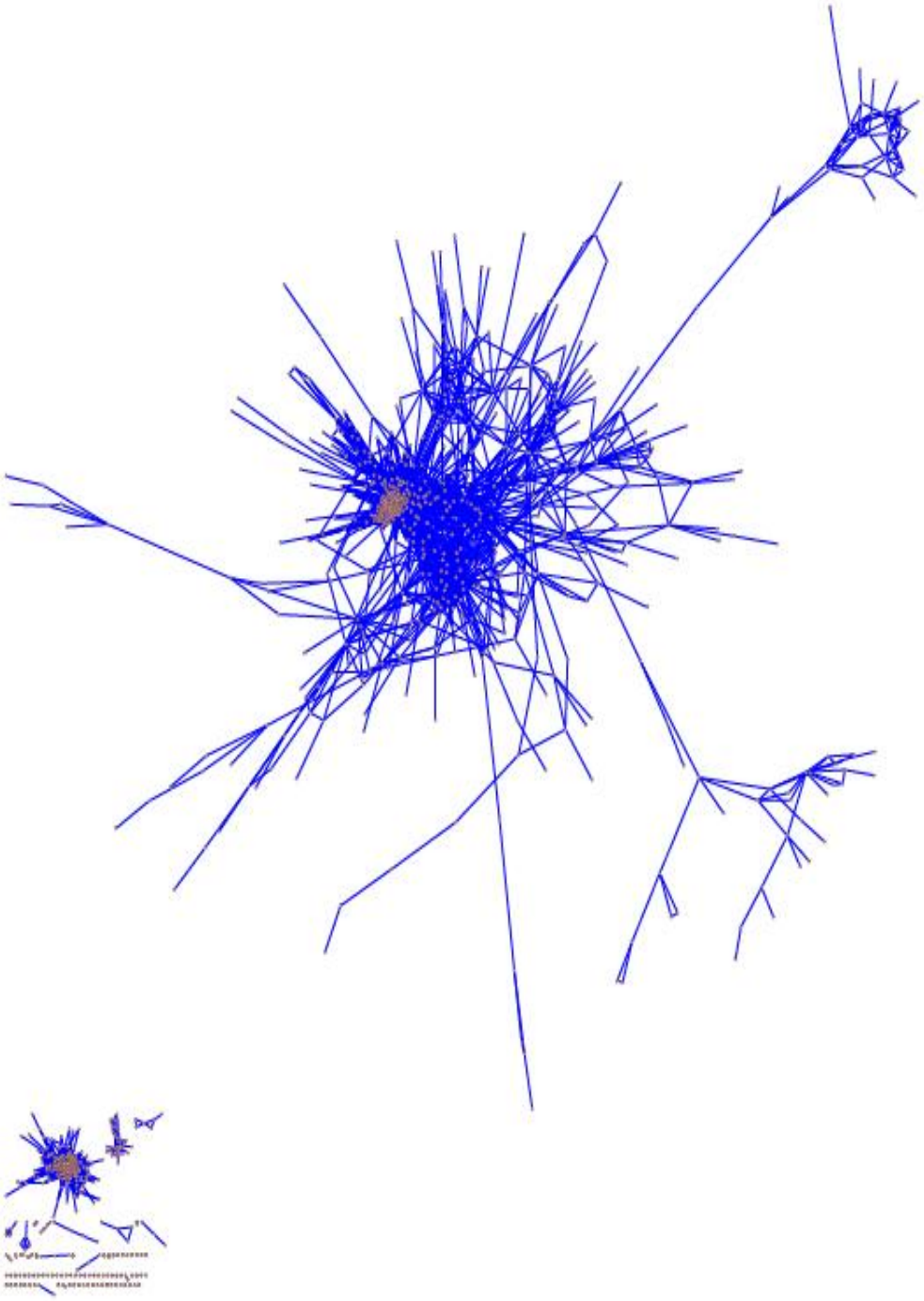

Figure 5: Graphical representation of R control network

Cytoscape force-directed (unweighted) layout.

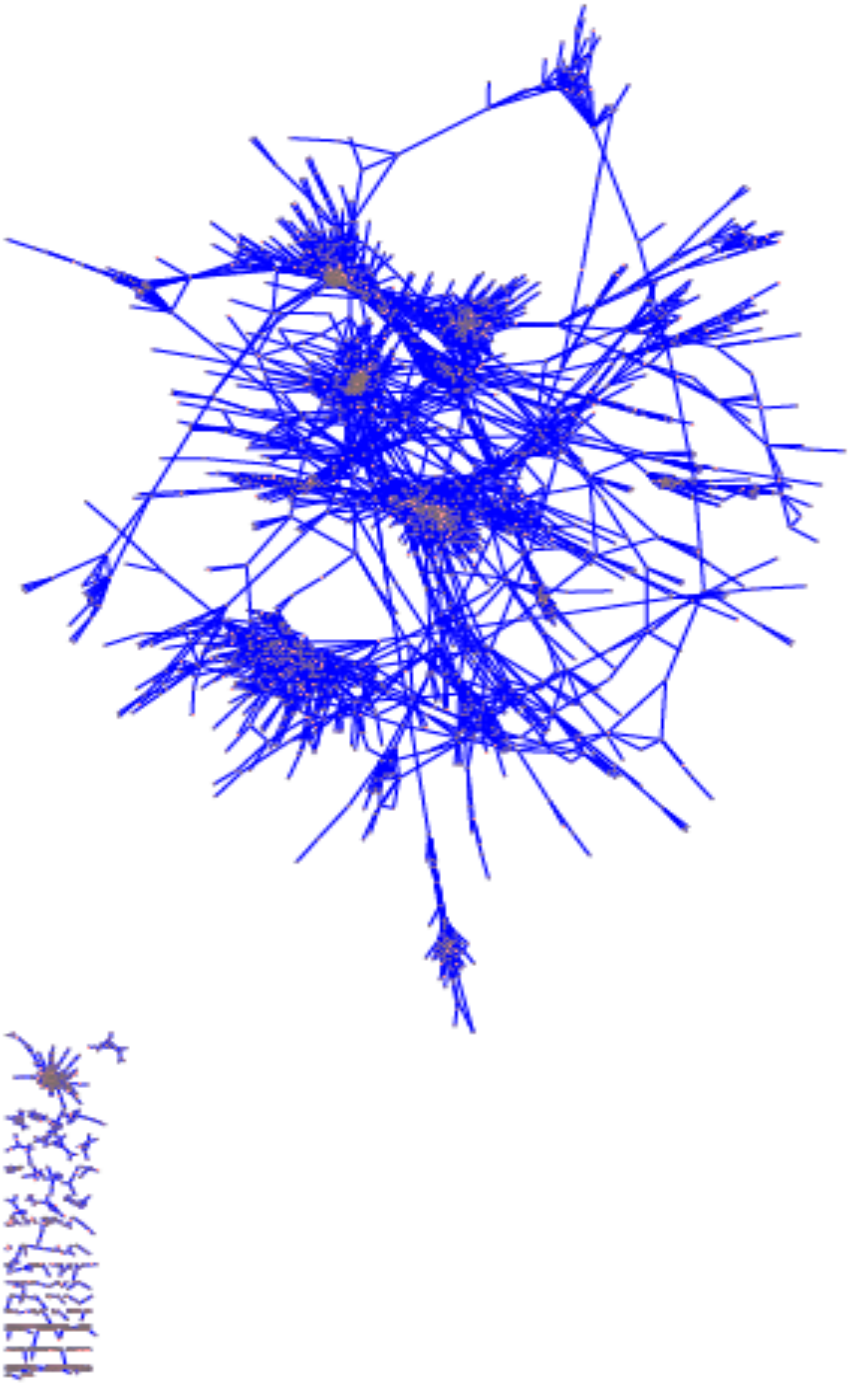

Figure 6: Graphical representation of R disease network

Cytoscape force-directed (unweighted) layout.

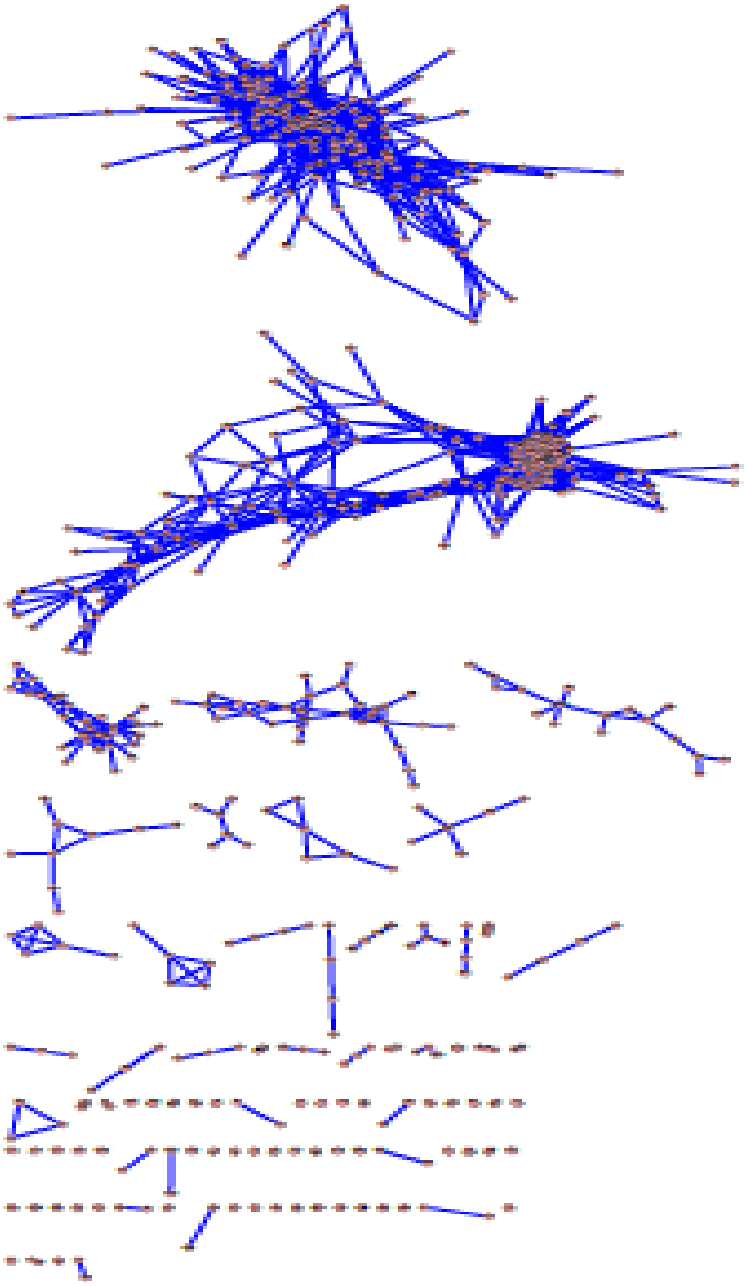

Figure 7: Graphical representation of RS control network

Cytoscape force-directed (unweighted) layout.

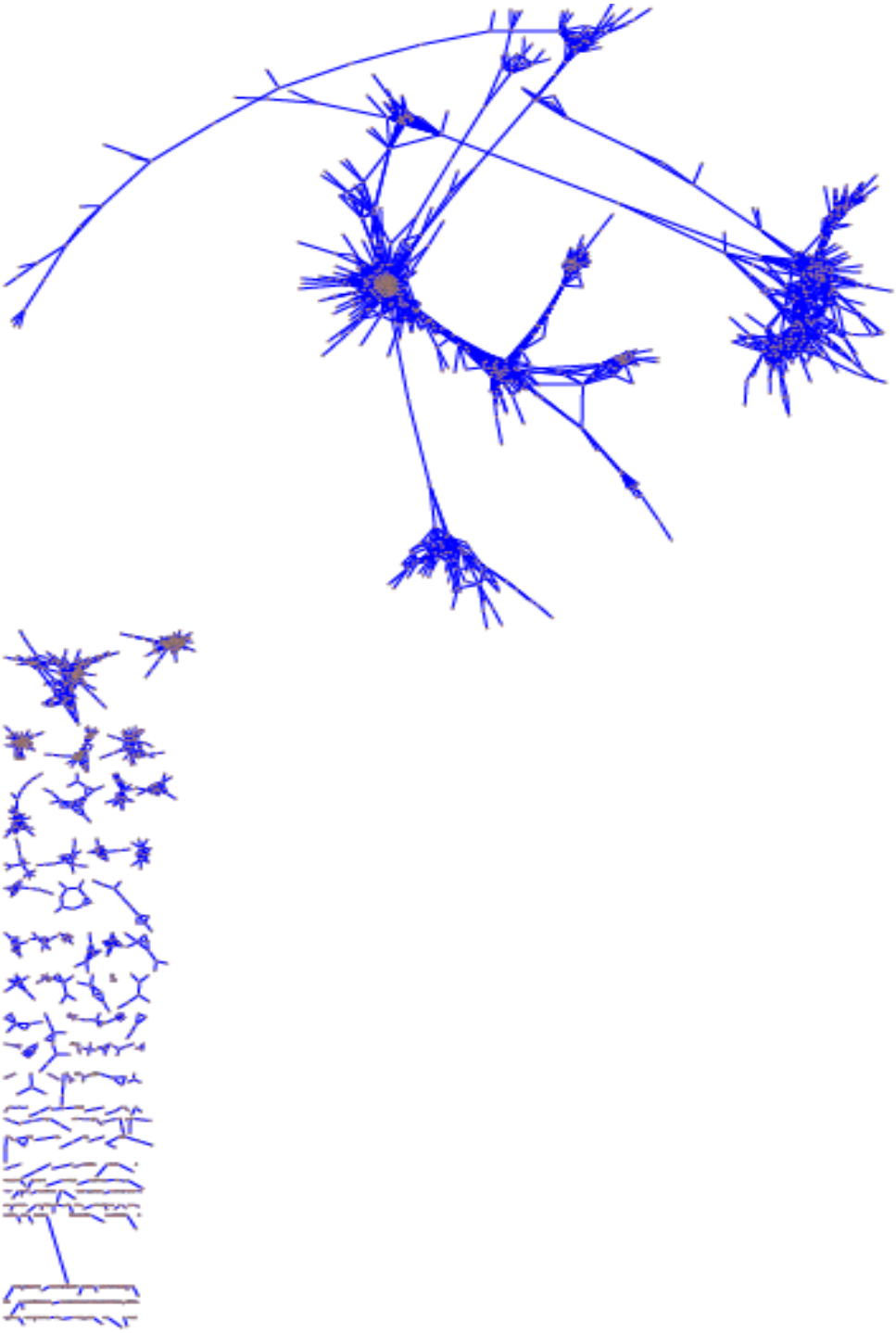

Figure 8: Graphical representation of RS disease network

Cytoscape force-directed (unweighted) layout.

4.2 Prioritized Genes

Genes were prioritized with respect to changes in three node properties between control and disease networks: betweenness centrality, clustering coefficient and neighborhood. Statistical significances of changes in the measures were analyzed by a non-parametric method, Wilcoxon signed-rank test, since values are not normally distributed. Changes in betweenness centrality and clustering coefficient were tested between control and disease networks separately as significance of increase and significance of decrease (with confidence level= 0.95) while neighborhood changes were tested against $\mu = 0$ (with confidence level= 0.95). All changes in comparison parameters were statistically significant. Prioritized genes with their official full name, associated biological processes obtained from *http://www.genecards.org* [25] and p-value of changes for each parameter are listed in Appendix A.

Functional clustering was performed in David with prioritized genes by each method for each parameter change. Lists of functional clusters can be seen in Appendix B with number of genes clustered, enrichment score and statistical significances. Note that some annotations are present in more than one cluster and can be highly general such as biological process.
Biological processes that genes were clustered with respect to betweenness centrality change in R networks are cellular process, RNA processing, and regulation of cellular process. For RS networks, they are biological regulation and multicellular organismal development. Among the prioritized genes according to betweenness centrality increase in R disease network, regulation of gene expression is the predominant cellular process with 4 genes (SSB, RPL30, METTL5, and SF3B14). As in the case of betweenness centrality increase, regulation of gene expression is the predominant cellular process for the prioritized genes according to betweenness centrality decrease in R disease network with three genes (HNRNPU, RBM6, and SUMO2). In the case of betweenness centrality increase in RS disease network, cilia related genes (DNAI1, TEKT1, and ROPN1L) are majority. There are three zinc finger genes (ZNF600, ZNF223, and ZNF 786) in the list of prioritized genes according to betweenness centrality decrease in RS disease network. High number of prioritized genes related with regulatory functions is consistent with the idea that genes with high betweenness centrality tend to connect functional modules and pass information.

Genes with high clustering coefficient change were mostly clustered to common biological processes between R and RS networks such as response to stimulus, protein complex biogenesis, signal transduction, gene expression, transport, and localization. This is not surprising since genes with high clustering coefficient are generally subunits of protein complexes, e.g. FGA and FGG, or members of functional modules which should be strictly co-regulated independent of cell type or state. On the other hand, regulation of apoptosis and multicellular organismal development are unique clusters for R and RS networks respectively. Instability in co-expression of apoptosis related genes in adjacent samples may be the reason while changes in processes related with organismal development occur in late stages of transition. The lists of prioritized genes according to clustering coefficient increase in disease networks are composed of more than 10 genes since all of their clustering coefficient increase are the same, from 0 to 1. Ten of the prioritized genes are same between two methods while this similarity is not seen in CC decrease case. This is another indicator of transition in adjacent cells.

Neighborhood change was used to detect genes that are highly differentially co-expressed. It is not possible to mention about a significant pattern for their functions. However, genes are clustered to a general process, biological regulation for both methods. Moreover, half of the genes prioritized are the same in R and RS networks. Presence of cytoskeleton related genes for both methods may explain why transforming cells were detected as non-tumorous after histological comparison with healthy samples since morphology change is a late stage in carcinogenesis and coregulation of those genes are stable across control samples of individuals.

4.3 Co-expressions of Interacting Proteins

Number of interacting proteins connected in co-expression networks is listed in Table 4.3.1. All of the PPI edges also present in R control and RS control networks are same. Number of interacting proteins also connected in disease co-expression networks is low but consistent with previous studies [30, 31]. The reason of the difference in the number of interacting proteins between control and disease networks may be that tissue samples for PPI studies are generally obtained from patients rather than healthy people. An important property of interacting proteins in disease networks is that they are mostly members of protein complexes or related with immune response and cell division. This indicates that co-regulation of interacting proteins related with these processes is more strictly controlled. Annotation clusters of interacting proteins mapped to co-expression networks are listed in APPENDIX C.

Table 2: Number of common edges between PPI and co-expression networks

	R Network		RS Network	
	Control	**Disease**	**Control**	**Disease**
PPI Network	6	79	6	43

4.4 Co-expressions of Intrinsically Disordered Proteins

Intrinsically disordered proteins were mapped to co-expression networks. None of the both members of edges in all four networks was disordered proteins. R and RS control networks had only one edge (AGO2-ZNF148) with argonaute RISC catalytic component 2 (AGO2) as the disordered protein. R disease network had eleven edges (PDGFC -YAP1, NR5A2-APOA1, APOA1-TEX11, APOA1-CYP4Z1, APOA1-GYPE, APOA1-THBD , APOA1-F11, APOA1-CCR4, AGO2-EVI5, SPIN1-RYBP, YAP1-CTSO) with disordered proteins apolipoprotein A-I (APOA1), Yes-associated protein 1 (YAP1), and RING1 and YY1 binding protein (RYBP), while RS disease network had eight edges (APOA1-CCR4, APOA1-CYP4Z1, APOA1-F11, APOA1-GYPE, APOA1-NR5A2, APOA1-TEX11, APOA1- THBD, RYBP-SPIN1) with disordered proteins apolipoprotein A-I (APOA1), and

RING1 and YY1 binding protein (RYBP). None of the disordered proteins are listed among prioritized genes. However, studies have shown that exogenous levels of apolipoprotein A-I prevent tumor development while its lowered levels are associated with ovarian cancer in mice [32]. Its differential co-expression with it interaction partners should be further analyzed. Apart from this, low number of disordered proteins is not sufficient to conclude about their co-expression patterns between two cellular states.

CHAPTER 5

CONCLUSION

Gene co-expression networks are fundamental aspects of systems biology. Identifying the true links in GCNs is a challenging task because of the nature of microarray data. This makes reliability of measure and selection of threshold crucial in network construction. In the absence of time course gene expression data, Pearson correlation coefficient seems to remain one of the mostly used measures defining the pairwise relationships. On the other hand, selection of the threshold automatically is a prerequisite for an objective and standardized network construction unless soft threshold is used.

In the thesis, a framework is proposed to deal with these two steps in network construction and applied to most common histological subtype of lung cancer, which is the leading cause of cancer death worldwide. A stability measure, *S* (stability), have been proposed to define the reliability of Pearson correlation coefficient for *A. thaliana* with hundreds of samples, was used for the same purpose with relatively low number of lung adenocarcinoma and adjacent lung tissue samples. Since stability removes the effects of samples that highly contribute to Pearson correlation, it may remove biases resulted from the sample composition independently from the number of samples. Moreover, it may reveal common co-expression dependencies in heterogeneous diseases like lung adenocarcinoma so that more effective therapies can be developed. In the detection of threshold, Random Matrix Theory was applied which has previously been used to distinguish system-specific, non-random properties of complex systems from random-noise.

Co-expression networks by using only Pearson correlation coefficient (R networks) were constructed to reveal the effects of using stability in combination with Pearson correlation coefficient (RS networks). Networks have interesting topological features such as absence of small-world property except from RS control network due to long paths between highly clustered nodes. Networks except from R control network are spoke-like and supports a recent study suggesting a heuristic spoke model rather than rigid hierarchy of deterministic hierarchical model. Co-expression networks constructed by Random Matrix Theory can be said topologically robust as shown in previous studies since filtration of a significant number of nodes by stability did not eliminate the spoke-like scale free property of the disease networks. However, R control network does not have a spoke-like structure as RS control network. On the other hand, RS control network differentiated from others such that it was composed of two big sub networks as the result of high number of genes filtrated out by stability. These differences between control networks can be explained by the nature of control samples and differential effect of stability on control networks rather than robustness issue of Random Matrix Theory.The reason of these differences might be the transition in cellular machinery of adjacent cells from non-cancer to cancer which cannot be detected with histological examination.

Phase transitions are characterized by extreme sensitivity to small perturbations. Thus, differential effects of tumor cell signals on adjacent cells depending on their distance to or location with respect to tumorous tissue in different samples are expected to be high and explain high number of genes filtrated out by stability as compared to disease samples.

A significant portion of genes prioritized with respect to betweenness centrality change are related with regulatory processes such as gene expression. This supports that genes with between high betweenness centrality serve to connect and transmit information between distinct cellular processes. The number of prioritized genes related with regulation of gene expression decreases in RS networks with respect to betweenness centrality change. This tendency is in reverse direction for genes related with cilia. Majority of prioritized genes are parts of protein complexes or members of functional modules which should be strictly co-regulated. On the other hand, more than half of the prioritized genes with respect to clustering coefficient increase in R and RS disease networks are same. This number is only one for increase in R and RS control networks. This tendency is also seen in the case of prioritized genes with respect to neighborhood change. These results indicate that stability effects betweenness centrality much more than other measures.

Presence of genes related with cilia in prioritized genes is important, especially TMEM17 (ciliogenesis; sonic hedgehog/SHH signaling), since epithelial cells that are defined as "cancer-initiating cells" generally have primary cilia, which are known to be required for activation of the sonic hedgehog/SHH signaling pathway whose abnormal activity has been reported in many cancers.

Number of interacting proteins also connected in co-expression networks was low which is consistent with previous studies. An important property of interacting proteins in disease networks is that they are mostly members of protein complexes or related with immune response and cell division. Number of disordered proteins mapped to co-expression networks was also very low to search for their co-expression patterns. Its reason is probably the low number of known disordered proteins.

The distinctions in prioritized gene similarity with respect to clustering coefficient and neighborhood change increase between control and disease networks together with the differential presence of genes related with regulation of gene expression with respect to betweenness centrality change provide valuable information about the dynamics of cancer progression and spread. However, a deep analysis that is not restricted to a couple of dozens of genes and supported with further laboratory experiments is needed for a better understanding. On the other hand, a different data set should be used with samples from healthy individuals to reveal the advantages of using stability to detect general mechanism differences between healthy and cancer state although using both methods provided different insights in this special case.

REFERENCES

[1] Zhu, X; Gerstein, M; Snyder, M. *Getting connected: analysis and principles of biological networks.*Genes & Development, 2007 May ; 21(9):1010-24.

[2] Albert, R. *Network Inference, Analysis, and Modeling in Systems Biology.* The Plant Cell.2007 November vol. 19 no. 11 3327-3338.

[3] Gopalacharyulu, PV; Lindfors, E; Bounsaythip, C; Kivioja, T;Yetukuri, L; Hollmén, J; Orešič, M. *Data integration and visualization system for enabling conceptual biology.* Bio-informatics, Volume 21, Issue suppl 1, Pp. i177-i185.

[4] Vicente, FFR; Lopes, FM; Hashimoto, RF; Cesar, RM. *Assessing the gain of biological data integration in gene networks inference.*BMC Genomics2012, 13(Suppl 6):S7.

[5] Nitsch, D; Tranchevent, LC; Thienpont, B; Thorrez, L;Van Esch, H; Devriendt, K; Moreau,Y. *Network Analysis of Differential Expression for the Identification of Disease-Causing Genes.* PLoS ONE, 2009 4(5): e5526. doi:10.1371/journal.pone.0005526.

[6] Torkamani, A; Dean, B; Schork, NJ; Thomas, EA. *Co-expression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia.*Genome Research2010 20: 403-412.

[7] Southworth, LK; Owen, AB; Kim, SK. *Aging Mice Show a Decreasing Correlation of Gene Expression within Genetic Modules.* PLoS Genetics 2009 5(12): e1000776. doi: 10. 1371/journal.pgen.1000776.

[8] Choi, YJ; Kendziorsk, C. *Statistical methods for gene set c-expression analysis.* Bio-informatics 2009 November 1; 25(21): 2780-2786.

[9] Luo, F; Yang, Y; Zhong, J; Gao, H; Khan, L; Thompson, DK; Zhou, J. *Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory.*BMC Bioinformatics 2007, 8:299 doi:10.1186/1471-2105-8-299.

[10] Dyson, HJ; Wright, PE. *Intrinsically unstructured proteins and their functions.* Nature Reviews Molecular Cell Biology 6, 197-208 (March 2005). doi:10.1038/nrm1589.

[11] Tompa, P. *Intrinsically unstructured proteins.* Biochem Science 2002 October; 27(10): 527-33.

[12] Babu, MM; van der Lee, R; de Groot, NS; Gsponer, J. *Intrinsically disordered proteins: regulation and disease.* Current Opinion in Structural Biology 2011, 21:1–9.

[13] Uversky, VN; Oldfield, CJ; Dunker, AK. *Intrinsically Disordered Proteins in Human Diseases: Introducing the D^2 Concept.* Annual Review of Biophysics June 2008, Vol. 37: 215-246.

[14] Mitchell, RS; Kumar, V; Abbas, AK; Fausto, N. *Chapter 13, box on morphology of adenocarcinoma.*Robbins Basic Pathology (8th ed.). Philadelphia: Saunders. ISBN 1- 4160-2973-7.

[15] Travis, WD; Brambilla, E; Muller-Hermelink, HK; Harris, CC. *Pathology and Genetics of Tumours of the Lung, Pleura, Thymus and Heart*. World Health Organization Classification of Tumours. Lyon: IARC Press. ISBN 92-832-2418-3. Retrieved 27 March 2010.

[16] Selamat, SA; Chung, BS; Laird-Offringa, IA. *Genome-scale analysis of DNA Methylation in lung adenocarcinoma and integration with mRNA expression.* Genome Research, 2012 July; 22(7): 1197-1211.

[17] Kinoshita, K; Obayashi, T. *Multi-dimensional correlations for gene co-expression and application to the large-scale data of Arabidopsis.* Bioinformatics. 2009 October 15;25(20): 2677-84. doi: 10.1093/bioinformatics/btp442.

[18] Gibson, SM; Ficklin, SP; Isaacson, S; Luo, F; Feltus, FA; Smith, MC. *Massive-Scale Gene Co-Expression Network Construction and Robustness Testing Using Random Matrix Theory*. PLoSOne.2013;8(2):e55871.

[19] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32867

[20] http://www.hprd.org/

[21] Mathivanan, S; Periaswamy, B; Gandhi, TKB; Kandasamy, K; Suresh, S; Mohmood, R; Ramachandra, YL; Pandey, A. *An evaluation of human protein-protein interaction data in the public domain.* BMC Bioinformatics 2006, 7(Suppl 5):S19 doi:10.1186/1471-2105-7-S5-S1.

[22] http://www.disprot.org/

[23] http://bci.clemson.edu/software/rmt

[24] Zhong, JX; Geisel, T. *Level fluctuations in quantum systems with multifractal Eigenstates.* Physical Review 1999, E 59:4071-4074.

[25] http://genecards.org/

[26] http://david.abcc.ncifcrf.gov/summary.jsp

[27] Wuchty, S; Ravasz, E; Barabasi, AL. *The Architecture of Biological Networks.* Complex Systems Science in Biomedicine. 2006, pp. 165-181.

[28] Hao, D; Ren, C; Li, C. *Revisiting the variation of clustering coefficient of biological networks suggests new modular structure.* BMC Systems Biology 2012, 6:34

[29] Chung, F; Lu, L; Dewey, TG; Galas, DJ. *Duplication Models for Biological Networks.* Journal of Computational Biology. October 2003, 10(5): 677-687

[30] Xulvi-Brunet, R; Li, H. *Co-expression networks: graph properties and topological comparisons* Bioinformatics. 2010 January 15;26(2):205-14. doi: 10.1093/bioinformatics/ btp632.

[31] Camargo, A; Azuaje, F. *Linking Gene Expression and Functional Network Data in Human Heart Failure.* PLoS ONE 2007 2(12): e1347. doi:10.1371/journal.pone. 0001347.

[32] Sua, F; Kozaka, KR; Imaizumib, S; Gaoa, F; Amneusa, MW; Grijalvab, V; Nga, C; Wagnerb, A; Houghb, G; Farias-Eisnerb, G; Anantharamaiahc, GM; Van Lentenb, BJ; Navabb, M; Fogelmanb, AM; Reddya, ST; Farias-Eisner, R.*Apolipoprotein A-I (apoA-I) and apoA-I mimetic peptides inhibit tumor development in a mouse model of ovarian cancer.*2010 Nov 16;107(46):19997-20002. doi: 10.1073/pnas.1009010107.

APPENDICES

APPENDIX A
LIST OF PRIORITIZED GENES

Table 3: Top 10 BC increase genes in R disease network

symbol	name	cellular processes*
SSB	Sjogren Syndrome Antigen B (Autoantigen La)	involves in diverse aspects of RNA metabolism
BCCIP	BRCA2 And CDKN1A Interacting Protein	may promote cell cycle arrest by enhancing the inhibition of CDK2 activity by CDKN1A
RPL30	Ribosomal Protein L30	RNA binding
METTL5	Methyltransferase Like 5	Probable methyltransferase activity
AKR1D1	Aldo-Keto Reductase Family 1, Member D1	delta4-3-oxosteroid 5beta-reductase activity
TMEM17	transmembrane protein 17	Required for ciliogenesis and sonic hedgehog/SHH signaling
SF3B14	splicing factor 3B, 14 kDa subunit	RNA binding
SNX29	Sorting Nexin 29	phosphatidylinositol binding
HSPE1	heat shock 10kDa protein 1 (chaperonin 10)	functions as a chaperonin
IER3IP1	immediate early response 3 interacting protein 1	may play a role in the ER stress response by mediating cell differentiation and apoptosis

p-value= 5.96×10^{-66}

*Cellular process information is obtained from *http://www.genecards.org*

Table 4: Top 10 BC decrease genes in R disease network

symbol	name	cellular processes*
HNRNPU	heterogeneous nuclear ribonucleoprotein U (scaffold attachment factor A)	RNA binding
CHMP5	charged multivesicular body protein 5	degradation of surface receptors formation of endocytic multivesicular bodies
PPIAP29	peptidylprolyl isomerase A (cyclophilin A) pseudogene 29	-
CCDC14	coiled-coil domain containing 14	-
RBM6	RNA binding motif protein 6	RNA binding
DPP9	dipeptidyl-peptidase 9	serine-type peptidase activity
FTH1P12	ferritin, heavy polypeptide 1 pseudogene 12	-
SUMO2	SMT3 Suppressor Of Mif Two 3 Homolog 2 (S. Cerevisiae)	nuclear transport transcriptional regulation apoptosis protein stability
BABAM1	BRISC and BRCA1 A complex member 1	DNA damage
ATP5C1	ATP synthase, H+ transporting, mitochondrial F1 complex, gamma polypeptide 1	ATP synthesis

p-value = 2.9×10^{-18}

*Cellular process information is obtained from *http://www.genecards.org.*

Table 5: Top 10 BC increase genes in RS disease network

symbol	name	cellular processes*
COL5A2	Collagen, Type V, Alpha 2	alpha chain for one of the low abundance fibrillar collagens
CALML3	Calmodulin-Like 3	calcium ion binding
LRRC48	leucine rich repeat containing 48	-
EFCAB1	EF-hand calcium binding domain 1	calcium ion binding
DNAI1	dynein, axonemal, intermediate chain 1	Part of the dynein complex of respiratory cilia
CD27	CD27 molecule	transmembrane signaling receptor activity; T cell immunity
TEKT1	tektin 1	Structural component of ciliary and flagellar microtubules microtubule cytoskeleton organization
CD19	CD19 molecule	B lymphocyte activation
CHST9	carbohydrate (N-acetylgalactosamine 4-0) sulfotransferase 9	cell-cell interaction signal transduction
ROPN1L	rhophilin associated tail protein 1-like	ciliary movement

p-value = 3.53×10^{-21}

*Cellular process information is obtained from *http://www.genecards.org.*

Table 6: Top 10 BC decrease genes in RS disease network

symbol	name	cellular processes*
ZNF600	zinc finger protein 600	May be involved in transcriptional regulation
ZNF223	zinc finger protein 223	May be involved in transcriptional regulation
DCLRE1C	DNA cross-link repair 1C	single-stranded DNA specific endodeoxyribonuclease activity 5'-3' exonuclease activity
ZNF786	zinc finger protein 786	May be involved in transcriptional regulation
STRIP2	striatin interacting protein 2	Plays a role in the regulation of cell morphology and cytoskeletal organization
SSTR2	somatostatin receptor 2	somatostatin receptor activity
CYLC2	Cylicin, Basic Protein Of Sperm Head Cytoskeleton 2	structural constituent of cytoskeleton
WDR38	WD repeat domain 38	-
GALNT3	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 3 (GalNAc-T3)	Plays a central role in phosphate homeostasis
LRRC46	leucine rich repeat containing 46	-

p-value = 1.04×10^{-18}

*Cellular process information is obtained from *http://www.genecards.org*.

Table 7: Top 16 CC increase genes in R disease network

symbol	name	cellular processes*
SNHG10	small nucleolar RNA host gene 10 (non-protein coding)	non-protein coding
RPS26P11	ribosomal protein S26 pseudogene 11	-
CD1A	CD1a molecule	natural killer T-cell activation
HINT1	histidine triad nucleotide binding protein 1	protein kinase C binding tumor suppression
EXOG	endo/exonuclease (5'-3'), endonuclease G-like	endonuclease activity may play a role in apoptosis
FGG	fibrinogen gamma chain	yields monomers that polymerize into fibrin platelet aggregation
GSDMA	gasdermin A	induction of apoptosis
RPS26	ribosomal protein S26	structural constituent of ribosome
WSB2	WD repeat and SOCS box containing 2	protein ubiquitination intracellular signal transduction
CMPK1	cytidine monophosphate (UMP-CMP) kinase 1, cytosolic	cellular nucleic acid biosynthesis
PDCD6	programmed cell death 6	calcium ion binding may mediate apoptosis
CD207	CD207 molecule, langerin	carbohydrate binding T-cell activation
SPG7	spastic paraplegia 7 (pure and complicated autosomal recessive)	metalloendopeptidase activity peptidase activity
MT1P3	metallothionein 1 pseudogene 3	-
MAFA	v-maf musculoaponeurotic fibrosarcoma oncogene homolog A (avian)	activates insulin gene expression
FGA	fibrinogen alpha chain	yielding monomers that polymerize into fibrin platelet aggregation

p-value = 1.29×10^{-24}

*Cellular process information is obtained from *http://www.genecards.org*.

Table 8: Top 10 CC decrease genes in R disease network

symbol	name	cellular processes*
STARD13	StAR-related lipid transfer (START) domain containing 13	GTPase activator activity may be involved in regulation of cytoskeletal reorganization, cell proliferation and cell motility
MRPL22	mitochondrial ribosomal protein L22	structural constituent of ribosome
ZNF528	zinc finger protein 786	may be involved in transcriptional regulation
TRMT5	tRNA methyltransferase 5 homolog (S. cerevisiae)	tRNA modification
MRPL20	mitochondrial ribosomal protein L20	structural constituent of ribosome
PSMG2	proteasome (prosome, macropain) assembly chaperone 2	promotes assembly of the 20S proteasome
UQCRH	ubiquinol-cytochrome c reductase hinge protein	ubiquinol-cytochrome-c reductase activity
SRP14	signal recognition particle 14kDa (homologous Alu RNA binding protein)	targeting secretory proteins to the rough endoplasmic reticulum membrane
ANGEL2	angel homolog 2 (Drosophila)	-
ACTL6A	actin-like 6A	involved in transcriptional activation and repression of select genes by chromatin remodeling

p-value = 7.08×10^{-23}

*Cellular process information is obtained from *http://www.genecards.org.*

Table 9: Top 15 CC increased genes in RS disease network

symbol	name	cellular processes*
RPS26P11	ribosomal protein S26 pseudogene 11	pseudogene
MT1P3	metallothionein 1 pseudogene 3	-
FGG	fibrinogen gamma chain	yields monomers that polymerize into fibrin platelet aggregation
WSB2	WD repeat and SOCS box containing 2	protein ubiquitination intracellular signal transduction
MAFA	v-maf musculoaponeurotic fibrosarcoma oncogene homolog A (avian)	activates insulin gene expression
CD207	v-maf musculoaponeurotic fibrosarcoma oncogene homolog A (avian)	carbohydrate binding T-cell activation
CCL3L3	chemokine (C-C motif) ligand 3-like 3	chemokine activity
PPM1K	protein phosphatase, Mg2+/Mn2+ dependent, 1K	Regulates the mitochondrial permeability transition pore essential for cellular survival and development
RPS26	ribosomal protein S26	structural constituent of ribosome
PIPSL	PIP5K1A and PSMD4-like, pseudogene	pseudogene
FGA	fibrinogen alpha chain	yields monomers that polymerize into fibrin platelet aggregation
EXOG	fibrinogen gamma chain	endonuclease activity may play a role in apoptosis
CDA	cytidine deaminase	cytidine deaminase activity
CD1A	CD1a molecule	natural killer T-cell activation
CCL3L1	chemokine (C-C motif) ligand 3-like 1	chemokine activity

p-value = 1.89×10^{-24}

*Cellular process information is obtained from *http://www.genecards.org.*

Table 10: Top 10 CC decrease genes in RS disease network

symbol	name	cellular processes*
ANXA2P1	annexin A2 pseudogene 1	-
BCYRN1	brain cytoplasmic RNA 1	may regulate dendritic protein biosynthesis
GSTA1	glutathione S-transferase alpha 1	glutathione transferase activity
PDE4C	phosphodiesterase 4C, cAMP-specific	glutathione transferase activity
STAR	steroidogenic acute regulatory protein	steroid hormone synthesis
CHGB	chromogranin B (secretogranin 1)	may serve as a precursor for regulatory peptides
ZNF528	zinc finger protein 528	Zinc finger
RABL2B	RAB, member of RAS oncogene family-like 2B	GTPase activity
ZNF563	zinc finger protein 563	may be involved in transcriptional regulation
PTGES3	prostaglandin E synthase 3 (cytosolic)	disrupts receptor-mediated transcriptional activation

p-value = 4.95×10^{-15}

*Cellular process information is obtained from *http://www.genecards.org.*

Table 11: Top 10 neighborhood change genes in R networks

symbol	name	cellular processes*
ARMC4	armadillo repeat containing 4	-
CCNA1	cyclin A1	control of the cell cycle
CCDC74A	coiled-coil domain containing 74A	-
KCNH3	potassium voltage-gated channel, subfamily H (eag-related), member 3	Pore-forming (alpha) subunit of voltage-gated potassium channel
MS4A8B	membrane-spanning 4-domains, subfamily A, member 8B	receptor activity
SPN	sialophorin	transmembrane signaling receptor activity; a negative regulatory role in adaptive immune response
TRIP13	thyroid hormone receptor interactor 13	This gene is one of several that may play a role in early-stage non-small cell lung cancer.
TMSB15A	thymosin beta 15a	organization of the cytoskeleton
STAG3L2	stromal antigen 3-like 2	-
CIDEC	cell death-inducing DFFA-like effector c	apoptosis

p-value = 1.33×10^{-93}

*Cellular process information is obtained from *http://www.genecards.org*.

Table 12: Top 10 neighborhood change genes in RS networks

symbol	name	cellular processes*
CCNA1	cyclin A1	control of the cell cycle
ARMC4	armadillo repeat containing 4	-
TMSB15A	thymosin beta 15a	organization of the cytoskeleton
CCDC74A	coiled-coil domain containing 74A	-
LRRC48	leucine rich repeat containing 48	-
CIDEC	cell death-inducing DFFA-like effector c	apoptosis
TMSB15B	thymosin beta 15b	organization of the cytoskeleton
CHST9	carbohydrate (N-acetylgalactosamine 4-0) sulfotransferase 9	cell-cell interaction; signal transduction
PCSK1	proprotein convertase subtilisin/kexin type 1	endopeptidase activity
CDHR2	cadherin-related family member 2	contact inhibition at the lateral surface of epithelial cells

p-value = 1.42×10^{-53}

*Cellular process information is obtained from *http://www.genecards.org*.

APPENDIX B
FUNCTIONAL ANNOTATION CLUSTERS OF PRIORITIZED GENES

Table 13: Functional annotation clusters of BC change in R networks

Annotation Cluster 1	**enrichment score : 1.46**	**count**	**p-value**	**Benjamini**
primary metabolic process		12	9.3E-3	8.5E-1
metabolic process		12	2.3E-2	8.0E-1
cellular metabolic process		11	2.9E-2	7.7E-1
macromolecule metabolic process		10	3.5E-2	7.6E-1
cellular macromolecule metabolic process		9	6.3E-2	8.1E-1
cellular process		13	1.2E-1	9.5E-1
Annotation Cluster 2	**enrichment score : 1.01**	**count**	**p-value**	**Benjamini**
RNA processing		4	1.2E-2	7.2E-1
mRNA metabolic process		3	4.4E-2	7.8E-1
RNA metabolic process		4	5.1E-2	7.8E-1
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process		6	1.9E-1	9.8E-1
cellular nitrogen compound metabolic process		6	2.3E-1	9.9E-1
nitrogen compound metabolic process		6	2.5E-1	9.9E-1
gene expression		5	2.9E-1	9.9E-1
Annotation Cluster 3	**enrichment score : 0.03**	**count**	**p-value**	**Benjamini**
regulation of cellular metabolic process		3	8.7E-1	1.0E0
regulation of metabolic process		3	8.8E-1	1.0E0
biological regulation		5	9.7E-1	1.0E0
regulation of cellular process		4	9.8E-1	1.0E0
regulation of biological process		4	9.9E-1	1.0E0

Table 14: Functional annotation clusters of BC change in RS networks

Annotation Cluster 1 enrichment score : 0.23	**count**	**p-value**	**Benjamini**
cellular macromolecule biosynthetic process	5	2.5E-1	1.0E0
macromolecule biosynthetic process	5	2.5E-1	1.0E0
biological regulation	9	3.7E-1	1.0E0
cellular biosynthetic process	5	4.0E-1	1.0E0
biosynthetic process	5	4.2E-1	1.0E0
nitrogen compound metabolic process	5	4.7E-1	1.0E0
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	4	5.0E-1	1.0E0
regulation of nitrogen compound metabolic process	4	5.0E-1	1.0E0
regulation of cellular biosynthetic process	4	5.3E-1	1.0E0
regulation of biosynthetic process	4	5.4E-1	1.0E0
cellular macromolecule metabolic process	6	5.6E-1	1.0E0
transcription	3	6.0E-1	1.0E0
regulation of primary metabolic process	4	6.1E-1	1.0E0
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	4	6.4E-1	1.0E0
regulation of cellular metabolic process	4	6.5E-1	1.0E0
macromolecule metabolic process	6	6.6E-1	1.0E0
regulation of metabolic process	4	6.9E-1	1.0E0
cellular nitrogen compound metabolic process	4	7.0E-1	1.0E0
regulation of biological process	7	7.2E-1	1.0E0
regulation of transcription	3	7.2E-1	1.0E0
regulation of macromolecule biosynthetic process	3	7.7E-1	1.0E0
regulation of gene expression	3	7.7E-1	1.0E0
gene expression	3	8.0E-1	1.0E0
cellular metabolic process	6	8.1E-1	1.0E0
regulation of cellular process	6	8.4E-1	1.0E0
regulation of macromolecule metabolic process	3	8.4E-1	1.0E0
primary metabolic process	6	8.5E-1	1.0E0
metabolic process	6	9.2E-1	1.0E0
Annotation Cluster 2 enrichment score : 0.1	**count**	**p-value**	**Benjamini**
multicellular organismal development	3	7.7E-1	1.0E0
multicellular organismal process	4	8.0E-1	1.0E0
developmental process	3	8.2E-1	1.0E0

Table 15: Functional annotation clusters of CC change in R networks

Annotation cluster 1	**enrichment score : 1.2**	**count**	**p-value**	**Benjamini**
response to calcium ion		3	2.7E-3	5.0E-1
response to metal ion		3	1.4E-2	7.1E-1
response to inorganic substance		3	3.4E-2	7.7E-1
response to chemical stimulus		5	1.0E-1	8.2E-1
response to stimulus		6	5.8E-1	9.9E-1
regulation of biological quality		3	6.3E-1	9.9E-1
Annotation Cluster 2	**enrichment score : 1.13**	**count**	**p-value**	**Benjamini**
cellular protein complex assembly		3	2.2E-2	7.6E-1
protein complex biogenesis		4	3.3E-2	8.2E-1
protein complex assembly		4	3.3E-2	8.2E-1
macromolecular complex assembly		4	6.5E-2	8.9E-1
cellular macromolecular complex assembly		3	7.4E-2	8.9E-1
macromolecular complex subunit organization		4	7.6E-2	8.7E-1
cellular macromolecular complex subunit organization		3	9.0E-2	8.2E-1
cellular component assembly		4	1.3E-1	8.6E-1
cellular component biogenesis		4	1.7E-1	9.1E-1
cellular component organization		6	2.7E-1	9.5E-1
Annotation Cluster 3	**enrichment score : 0.7**	**count**	**p-value**	**Benjamini**
signal transduction		8	8.9E-2	8.4E-1
regulation of cellular process		13	2.1E-1	9.4E-1
regulation of biological process		13	2.6E-1	9.6E-1
biological regulation		13	3.5E-1	9.6E-1
Annotation Cluster 4	**enrichment score : 0.51**	**count**	**p-value**	**Benjamini**
translation		4	1.1E-2	7.5E-1
gene expression		9	4.4E-2	8.1E-1
cellular macromolecule biosynthetic process		8	8.5E-2	8.8E-1
macromolecule biosynthetic process		8	8.8E-2	8.6E-1
cellular biosynthetic process		9	9.0E-2	8.0E-1
biosynthetic process		9	1.0E-1	8.1E-1
protein metabolic process		7	1.9E-1	9.4E-1
cellular process		18	2.2E-1	9.4E-1
cellular protein metabolic process		6	2.3E-1	9.5E-1
macromolecule metabolic process		11	2.6E-1	9.6E-1
cellular macromolecule metabolic process		10	3.0E-1	9.6E-1
cellular metabolic process		12	3.1E-1	9.6E-1
primary metabolic process		12	3.8E-1	9.7E-1
regulation of gene expression		6	3.8E-1	9.6E-1
metabolic process		13	3.9E-1	9.6E-1
regulation of RNA metabolic process		4	4.8E-1	9.9E-1
regulation of macromolecule metabolic process		6	5.1E-1	9.9E-1

Table B.3 (continued)

regulation of primary metabolic process	6	5.1E-1	9.9E-1
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	6	5.5E-1	9.9E-1
regulation of cellular metabolic process	6	5.7E-1	9.9E-1
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	5	5.9E-1	9.9E-1
transcription	4	5.9E-1	9.9E-1
regulation of macromolecule biosynthetic process	5	5.9E-1	9.9E-1
regulation of nitrogen compound metabolic process	5	5.9E-1	9.9E-1
regulation of metabolic process	6	6.1E-1	9.9E-1
cellular nitrogen compound metabolic process	6	6.2E-1	9.9E-1
regulation of cellular biosynthetic process	5	6.3E-1	9.9E-1
regulation of biosynthetic process	5	6.3E-1	9.9E-1
nitrogen compound metabolic process	6	6.5E-1	9.9E-1
regulation of transcription, DNA-dependent	3	7.4E-1	1.0E0
regulation of transcription	4	7.4E-1	1.0E0
Annotation Cluster 5 enrichment score : 0.49	**count**	**p-value**	**Benjamini**
regulation of apoptosis	3	3.2E-1	9.6E-1
regulation of programmed cell death	3	3.2E-1	9.6E-1
regulation of cell death	3	3.2E-1	9.5E-1
Annotation Cluster 6 enrichment score : 0.24	**count**	**p-value**	**Benjamini**
establishment of localization in cell	3	3.4E-1	9.6E-1
cellular localization	3	3.8E-1	9.6E-1
transport	4	7.5E-1	1.0E0
establishment of localization	4	7.6E-1	1.0E0
localization	4	8.3E-1	1.0E0

Table 16: Functional annotation clusters of CC change in RS networks

Annotation Cluster 1	**enrichment score : 1.41**	**count**	**p-value**	**Benjamini**
regulation of biological process		14	2.6E-2	8.0E-1
biological regulation		14	4.3E-2	8.9E-1
regulation of cellular process		13	5.4E-2	9.1E-1
Annotation Cluster 2	**enrichment score : 1.09**	**count**	**p-value**	**Benjamini**
regulation of biological quality		7	5.7E-3	8.2E-1
response to metal ion		3	1.0E-2	7.9E-1
response to inorganic substance		3	2.5E-2	8.5E-1
response to chemical stimulus		5	6.2E-2	9.1E-1
response to external stimulus		4	9.3E-2	9.6E-1
response to wounding		3	1.3E-1	9.5E-1
response to stress		4	3.3E-1	1.0E0
response to stimulus		6	4.2E-1	1.0E0
multicellular organismal process		4	9.3E-1	1.0E0
Annotation Cluster 3	**enrichment score : 0.89**	**count**	**p-value**	**Benjamini**
regulation of biological quality		7	5.7E-3	8.2E-1
regulation of body fluid levels		3	1.2E-2	7.1E-1
protein complex assembly		3	1.2E-1	9.5E-1
protein complex biogenesis		3	1.2E-1	9.5E-1
macromolecular complex assembly		3	1.9E-1	9.9E-1
macromolecular complex subunit organization		3	2.1E-1	9.9E-1
cellular component assembly		3	2.9E-1	1.0E0
cellular component biogenesis		3	3.4E-1	1.0E0
cellular component organization		5	3.5E-1	1.0E0
multicellular organismal process		4	9.3E-1	1.0E0
Annotation Cluster 4	**enrichment score : 0.61**	**count**	**p-value**	**Benjamini**
regulation of biological quality		7	5.7E-3	8.2E-1
transport		3	8.5E-1	1.0E0
establishment of localization		3	8.6E-1	1.0E0
localization		3	9.0E-1	1.0E0
Annotation Cluster 5	**enrichment score : 0.39**	**count**	**p-value**	**Benjamini**
regulation of biological quality		7	5.7E-3	8.2E-1
organ development		3	6.4E-1	1.0E0
system development		3	8.0E-1	1.0E0
anatomical structure development		3	8.4E-1	1.0E0
multicellular organismal development		3	8.9E-1	1.0E0
developmental process		3	9.2E-1	1.0E0
multicellular organismal process		4	9.3E-1	1.0E0

Table B.4 (cont.)

Annotation Cluster 6 enrichment score : 0.31	count	p-value	Benjamini
cellular biosynthetic process	8	9.6E-2	9.5E-1
biosynthetic process	8	1.1E-1	9.5E-1
cellular metabolic process	11	2.3E-1	9.9E-1
primary metabolic process	11	2.9E-1	1.0E0
regulation of primary metabolic process	6	3.6E-1	1.0E0
regulation of RNA metabolic process	4	3.8E-1	1.0E0
metabolic process	11	4.5E-1	1.0E0
cellular macromolecule metabolic process	8	4.5E-1	1.0E0
cellular macromolecule biosynthetic process	5	4.5E-1	1.0E0
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	5	4.5E-1	1.0E0
regulation of metabolic process	6	4.5E-1	1.0E0
macromolecule biosynthetic process	5	4.5E-1	1.0E0
regulation of nitrogen compound metabolic process	5	4.6E-1	1.0E0
gene expression	5	5.0E-1	1.0E0
cellular protein metabolic process	4	5.6E-1	1.0E0
cellular process	14	5.6E-1	1.0E0
macromolecule metabolic process	8	5.7E-1	1.0E0
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	5	6.2E-1	1.0E0
regulation of cellular metabolic process	5	6.3E-1	1.0E0
regulation of transcription, DNA-dependent	3	6.5E-1	1.0E0
cellular nitrogen compound metabolic process	5	6.8E-1	1.0E0
protein metabolic process	4	6.9E-1	1.0E0
regulation of gene expression	4	7.0E-1	1.0E0
nitrogen compound metabolic process	5	7.1E-1	1.0E0
regulation of biosynthetic process	4	7.3E-1	1.0E0
transcription	3	7.4E-1	1.0E0
regulation of macromolecule metabolic process	4	7.9E-1	1.0E0
regulation of transcription	3	8.5E-1	1.0E0
regulation of macromolecule biosynthetic process	3	8.8E-1	1.0E0
regulation of cellular biosynthetic process	3	9.0E-1	1.0E0

Table 17: Functional annotation clusters of NC in R networks

Annotation Cluster 1 enrichment score : 0.45	count	p-value	Benjamini
regulation of cellular process	5	3.1E-1	1.0E0
regulation of biological process	5	3.5E-1	1.0E0
biological regulation	5	4.0E-1	1.0E0

Table 18: Functional annotation clusters of NC in RS networks

Annotation Cluster 1 enrichment score : 0.78	count	p-value	Benjamini
regulation of biological quality	4	9.6E-3	8.4E-1
biological regulation	5	2.3E-1	1.0E0
cellular process	6	2.3E-1	1.0E0
regulation of cellular process	4	4.7E-1	1.0E0
regulation of biological process	4	5.1E-1	1.0E0

APPENDIX C
FUNCTIONAL ANNOTATION CLUSTERS OF CO-EXPRESSED INTERACTING PROTEINS

Table 19: Functional annotation clusters of co-expressed interacting proteins

Annotation Cluster 1	**enrichment score : 12.9**	**count**	**p-value**	**Benjamini**
cell cycle process		29	8.6E-17	1.2E-13
mitotic cell cycle		24	6.1E-16	2.4E-13
cell cycle		32	6.7E-16	1.8E-13
cell cycle phase		23	7.9E-14	1.7E-11
M phase		21	1.0E-13	1.9E-11
cell division		19	1.9E-12	2.6E-10
mitosis		16	3.3E-11	4.0E-9
nuclear division		16	3.3E-11	4.0E-9
M phase of mitotic cell cycle		16	4.2E-11	4.6E-9
organelle fission		16	5.8E-11	5.7E-9
organelle organization		29	7.2E-8	4.4E-6
Annotation Cluster 2	**enrichment score : 11.44**	**count**	**p-value**	**Benjamini**
immune system process		36	2.3E-16	1.2E-13
immune response		27	9.2E-13	1.4E-10
defense response		19	2.3E-7	1.2E-5

TEZ FOTOKOPİ İZİN FORMU

ENSTİTÜ

Fen Bilimleri Enstitüsü ☐

Sosyal Bilimler Enstitüsü ☐

Uygulamalı Matematik Enstitüsü ☐

Enformatik Enstitüsü ☐

Deniz Bilimleri Enstitüsü ☐

YAZARIN

Soyadı : AKDEMİR
Adı : ERHAN
Bölümü : BİYOENFORMATİK

TEZİNADI(İngilizce) : A FRAMEWORK FOR GENE CO-EXPRESSION NETWORK ANALYSIS OF LUNG CANCER

TEZİNTÜRÜ: Yüksek Lisans ☐ Doktora ☐

1.Tezimin tamamı dünya çapında erişime açılsın ve kaynak gösterilmek şartıyla tezimin bir kısmı veya tamamının fotokopisi alınsın. ☐

2.Tezimin tamamı yalnızca Orta Doğu Teknik Üniversitesi kullancılarının erişimine açılsın. (Bu seçenekle tezinizin fotokopisi ya da elektronik kopyası Kütüphane aracılığı ile ODTÜ dışına dağıtılmayacaktır.) ☐

3.Tezim bir (1) yıl süreyle erişime kapalı olsun. (Bu seçenekle tezinizin fotokopisi ya da elektronik kopyası Kütüphane aracılığı ile ODTÜ dışına dağıtılmayacaktır.) ☐

Yazarın imzası ..... Tarih .....