ANALYSES AND MODELING OF OVARIAN CANCER MICROARRAY DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

BARIŞ SU KARAKELLE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
BIOMEDICAL ENGINEERING

DECEMBER 2019

Approval of the thesis:

THESIS TITLE

submitted by **BARIŞ SU KARAKELLE** in partial fulfillment of the requirements for the degree of **Master of Science** i**n Biomedical Engineering, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Vilda Purutçuoğlu
Head of the Department, **Biomedical Engineering** _____

Prof. Dr. Vilda Purutçuoğlu
Supervisor, **Statistics, Middle East Technical University** _____

Assoc. Prof. Dr. Çağdaş Devrim Son
Co-Supervisor, **Biology, Middle East Technical University** _____

Examining Committee Members:

Assoc. Prof. Dr. Tülin Yanık
Biology, Middle East Technical University _____

Prof. Dr. Vilda Purutçuoğlu
Statistics, Middle East Technical University _____

Assoc. Prof. Dr. Yüksel Ürün
Medical Oncology, Ankara University _____

Assoc. Prof. Dr. Özlem Türkşen
Statistics, Ankara University _____

Assoc. Prof. Dr. Ceren Vardar Acar
Statistics, Middle East Technical University _____

Date: 10.12.2019

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 : Barış Su Karakelle

Signature :

ABSTRACT

ANALYSES AND MODELING OF OVARIAN CANCER MICROARRAY DATA

Karakelle, Barış Su
Master of Science, Biomedical Engineering
Supervisor : Prof. Dr. Vilda Purutçuoğlu
Co-Supervisor: Assoc. Prof. Dr. Çağdaş Devrim Son

December 2019, 40 pages

Ovarian cancer is one of the common cancer types among other oncological diseases. The major causes of this cancer can be listed as age, obesity, hormone therapy, material inheritance and contraceptive pills. Due to its generality and importance, many researches have been conducted from distinct labs about this illness and its plausible causes have been intensively investigated either in microarray studies, where just part of the related genes are detected, or in the pairwise correlation analyses between the disease and selected symptoms via contingency tables. Hereby, in this study, we use an ovarian cancer microarray dataset and describe gene interactions in these data via two different modelling approaches, namely, Gaussian graphical model as a parametric model and artificial neural network as a nonparametric model. From these analyses, we evaluate certain findings biologically and then, compare the performance of the model accuracies in distinct accuracies measures by controlling the true network structures of selected genes. By this way, we aim to assess the performance of these two fundamental models by using this specific oncogene data.

Keywords: Ovarian Cancer, Microarray, Gaussian Graphical Model, Neural Network

ÖZ

YUMURTALIK KANSERİ MİKRODİZİN VERİSİNİN ANALİZLERİ VE MODELLEMESİ

Karakelle, Barış Su
Yüksek Lisans, Biyomedikal Mühendisliği
Tez Yöneticisi: Prof. Dr. Vilda Purutçuoğlu
Ortak Tez Yöneticisi: Doç. Dr. Çağdaş Devrim Son

Aralık 2019, 40 sayfa

Yumurtalık kanseri, diğer onkolojik hastalıklarla birlikte yaygın olan kanser tiplerinden birisidir. Bu tip kanserin nedenleri arasında yaş, obezite, hormon terapisi, maddesel kalıtım ve doğum kontrol hapları sayılabilir. Yaygınlığı ve önemi nedeniyle bu kanserin nedenleri ya ilgili genlerin tespit edildiği mikrodizin araştırmalarıyla ya da hastalık ve seçilmiş bazı semptomları arasında olasılık tablolarının kullanımıyla yapılan ilişkilendirme analizleriyle birçok bağımsız araştırmacı tarafından incelenmiştir. Bu çalışmada, bir yumurtalık kanseri mikrodizin veri kümesi kullanmaktayız ve bu verideki gen etkileşimlerini iki farklı modelleme yaklaşımı ile ifade etmekteyiz. Bunlar parametrik bir model olarak Gaussian grafiksel modeli ve parametrik olmayan bir model olarak yapay sinir ağı modelidir. Bu analizlerden, bazı bulguları biyolojik olarak değerlendirmekteyiz ve sonrasında modellerin doğruluk performanslarını, seçilen genlerin doğru ağ yapılarıyla kontrol ederek farklı doğruluk ölçüleri ile karşılaştırmaktayız. Böylece bu iki temel modelin performansını bu özel onkojen verisini kullanarak değerlendirmeyi amaçlamaktayız.

Anahtar Kelimeler: Yumurtalık Kanseri, Mikrodizin, Gaussin Grafiksel Modeli, Yapay Sinir Ağı

To my family, including my grandmother and aunts...

ACKNOWLEDGEMENTS

I genuinely acknowledge my supervisor Prof. Dr. Vilda PURUTÇUOĞLU for giving me the chance to conduct this study with her, her incredible support, affection, compassion and patience. She did not let me give up during the hardest period of my life and did her best to see me standing up again. She believed in me when I was losing faith in myself. I would like to thank Assoc.Prof.Dr. Yeşim AYDIN SON and her assistant Ayşegül Tombuloğlu as well, for their tremendous help in data acquisition stage and BRB array tool installing and usage phase.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGEMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xii

LIST OF FIGURES xiii

1 INTRODUCTION 1

2 LITERATURE REVIEW AND METHODOLOGY 3

3 RESULTS 23

4 CONCLUSION 37

REFERENCES 39

LIST OF TABLES

TABLES

Table 3.1: List of genes for each 10-dimensional network.....25

Table 3.2: Results of accuracy measures for 10-dimensional datasets via GGM. NA denotes not avalaible 26

Table 3.3: List of genes for each 23-dimensional network.....28

Table 3.4: Results of accuracy measures for 23-dimensional datasets via GGM....30

Table 3.5: List of genes for each 23-dimensional network 32

Table 3.6: Results of accuracy measures for 50-dimensional datasets via GGM.....33

Table 3.7: Estimated binary network structure for Data 6 via ANN.....34

Table 3.8: Results of accuracy measures for 10-dimensional datasets via ANN....34

Table 3.9: Results of accuracy measures for 23-dimensional datasets via ANN.....35

Table 3.10: Results of accuracy measures for 50-dimensional datasets via ANN.....35

LIST OF FIGURES

FIGURES

Figure 2.9.1. Scheme of a feedforward artifical neural network.....20

Figure 3.1: Corresponding networks of the 6th dataset composed of 10 genes and (a) estimated by GGM and (b) the structure of its true network in STRING.....24

Figure 3.2: Corresponding networks of the 7th dataset composed of 23 genes and (a) estimated by GGM and (b) the structure of its true network in STRING2.....26

Figure 3.3: Corresponding networks of the 8th dataset composed of 23 genes and (a) estimated by GGM and (b) the structure of its true network in STRING.....28

Figure 3.4: Corresponding networks of the 13th dataset composed of 50 genes and (a) estimated by GGM and (b) the structure of its true network in STRING.....30

CHAPTER 1

INTRODUCTION

The huge amount of data in the field of genomics motivated an effort to comprehend the relations between genes and how these can be related to different conditions such as genetic diseases. Many studies use gene co-expression networks to work with big gene expression datasets with the aim of understanding biological processes. Microarray technology gives a chance for the simultaneous measurement of the expression level of thousands of genes in a single hybridization assay. Modelling co-expression with network models is significantly beneficial to obtain information about the network structure and pivotal genes and how they may interact with each other. Considering genetic illnesses like cancer, the information obtained by the networks constructed can be highly important in diagnostics and drug design. That's why, recently significant number of scientists in different fields are trying to integrate and utilize statistical approaches to decipher knowledge about genetic networks. Gaussian graphical model (GGM) is a useful, commonly used parametric method to receive information about undirected associations among the genes. The inverse covariance matrix which is also known as the precision matrix is capable of giving an idea about which genes are co-expressed or interacting based on the Gaussian assumptions. Several regression methods like the one we used –glasso- are used to obtain the estimated connections in the biological networks based on the fact that these networks are not sparse and include a number of genes that are much greater than the number of observations. Based on this issue and many other aims, the modification of current or traditional statistical methods are being considered for the analysis of cancer. Lately, artificial neural networks (ANNs), construction to understand biological networks has been used after it was

found to a powerful method in econometrics and psychology. We aim to construct genetic networks using GGM and ANN based on microarray data for ovarian cancer patients. To be able to understand if they are compatible with literature, we proposed to use STRING (Szklarczyk, D., et al., 2018) based networks as a reference, taking them as true networks.

- Is GGM successful in the analyses of actual biological systems even though its application via simulated datasets and bench-mark systems indicate acceptable accuracy in terms of distinct accuracy measures?

- Can ANN be applicable as an alternative of GGM in a nonparametris setting? Because ANN became very popular in recent years. Whereas it is implemented typically in the classification analyses or very big dataset. Thereby, we try to adapt ANN as an alternative of the parametric GGM method in the construction of the protein-protein interaction networks with relatively limited number of observations.

While answering both questions, we apply the ovarian cancer dataset by comperehensively investigating their estimated structures with the quasi true networks.

CHAPTER 2

LITERATURE REVIEW AND METHODOLOGY

2.1 Description of Cancer

Cancer is a combination of diseases that develop due to genetic and epigenetic changes which affects cellular development and cell death. Certain molecular changes which trigger a healthy cell to turn into malignant, identified as being able to invade and spread from the initial lesion, was found out, yet the range of those transformations change significantly with respect to cancer type. It is currently known that metastasis relies on a balance of triggering factors and inhibitory signals, that must be higher for the stimulating signals to cause metastasis so it is convenient to state that cancer is a process where the equilibrium among cell growth and death is shifted and cell proliferation is not controlled anymore (Liu et al., 2017). Different molecules are known to direct the invasion, metastasis and rejection towards therapy. Vast accumulation of genomic and proteomic data in the last two decades has improved comprehension of the molecular background of human cancer and applied to prognostic and treatment methods. Specifically designed therapies have been improved fundamentally for malignant, refractory or repeating cancers, based on molecular profiles. Healthy cells respond properly to the signals that control their metabolism and every type of behaviour whereas, cancer cells in a way ignore any external upcoming signal and divide continously in a non-controllable way. Throughout this process, they tend to invade normal / healthy tissues and organs and final stage as spreading to whole body. This lack of controlled growth seen in cancer cells is the consequence of many different types of pecularities in several cell regulatory systems and is manifested in different features of cell behavior that define cancer cells as different from healthy cells.

When the development of cancer is observed in a cellular level, a many-stage process having mutation and choice for cells with gradually augmenting ability for division, survival, invasion, and metastasis is clear.

2.2 Reasons for Cancer

Cancer progresses due to genetic mutations which change the normal mechanisms that control cell population. When several different mutations occur in some genes, it may disturb the critical processes for the cell like proliferation, cell death and aging. Mutations may be due to external factors such as exposure to radiation or endogenously as replication errors occuring during normal process of division. Most cancers are directly proportional with age due to accumulation of damage in genome over years. Genetic mutations of DNA are composed of amplification, insertions, deletions, rearrangements together with point mutations (Vijg J., 2014).

More than 100 different types of cancer having solitary patterns of growth and invasion are identified. Causes can be grouped as external and intrinsic. External aspects are compromised of chemicals, radiation and viruses. Chemicals which are grouped as carcinogenic are smoking, excessive usage of alcohol and asbestos, some industrial chemicals and medications. Dietary habits are also known to trigger several types indeed. Intrinsic or host related parameters consist of hormones, immunity related factors and hereditary mutations. Whereas few cancers, like breast and colon cancer, are known to include signficant 'familial predisposition', there is no striking proof that cancer is actually 'programmed' in the cells (Vijg J., 2014).

2.3 How Genes Lead to Cancer

To have the option to comprehend how some genes cause malignant growth, it might be gainful to concentrate on some essential hereditary ideas. Genes come pairwise, and work together to make a protein item. One individual from the pair

originates from the mother, while the other part is acquired from the dad. Eggs and sperm are known as "germ cells." When a change or transformation in a gene happens in the germ cells, it is named as a "germline mutation." When a germline mutation is passed into next generation, it is seen in all cells of the body. On the other hand, transformations that we are not brought into the world with, however that happen by chance after some time in cells of the body are said to be "acquired". Acquired changes are not inherited and not seen in entire cells. Acquired mutations are more commonly identified with shaping malignant growth while germline transformations give a littler rate.

In our genome, there are many different types of genes that control cell growth in a very strictly regulated and programmed manner. When these genes have a fault in their DNA code, they might not work appropiately, and are described to be "altered" or mutated. Increasing number of many mutations in different genes happening in a particular community of cells as time goes by is essential to bring out malignancy. The various types of genes, which when mutated, may result in the development of cancer are described below.

Oncogenes are "turned on" types of genes which are named as proto-oncogenes. Proto-oncogenes are DNA successions found in the genomes of healthy and neoplastic tissues. Proto-oncogenes show significant evolutionary conservation, being found in a significant number of phylogenetically separated living beings extending from Drosophila flies to people (Chial, H., 2008). Protooncogenes encode development factors, growth factor receptors, and proteins that manage cell development and separation. Proto-oncogenes are by and large characterized dependent on the main role or sequence homology of their protein items. Whenever modified or mutated, they become oncogenes and after that may trigger tumor initiation or development.

Mutations in tumor suppressor are typically acquired. The two mutations in a tumor suppressor gene pair may happen as the aftereffect of aging or external factors. A mutation in a tumor supressor gene can likewise be inherited.

When a cell is in the process of division, the DNA makes a duplicate or copy of itself. During this intricate procedure, errors may happen. Mismatch-repair genes are DNA repair genes that correct these immediately happening blunders in the DNA. At the point when these genes are changed or mutated, nonetheless, mismatches in the DNA remain without amendment. In the event that these missteps happen in tumor suppressor genes or proto-oncogenes, at long last it will bring about uncontrolled cell growth and tumor formation. There are different kinds of DNA repair genes that repair mistakes in DNA that happen from mutagenic operators like untolerable dosages of radiation.

It takes mutations in few genes for cancer to start. In high proportion of cancer occurrences, every one of the mutations are acquired. In inherited cancer, one mutation is passed down from the parent, yet the other one is acquired. Because of the way that in excess of a single mutation is expected to cause cancer, it isn't consistent to express that all people who inherit a mutation in a tumor silencer gene, proto-oncogene, or DNA repair gene will develop cancer. Proto-oncogene amplification is a moderately regular occasion in gynecologic malignancies. By and large, improved oncogene expression is seen together with oncogene amplification.

2.4 Gynecologic Cancers

These are defined as any cancer that begins in woman's reproductive organs. Five major types of cancer impacting a woman's reproductive organs are cervical, ovarian, endometrial, vaginal and vulvar and their cancers are known as gynecologic cancers. Sixth type of gynecologic cancer is the very rare fallopian tube cancer.

The major causes of this cancer can be listed as age, obesity, hormone therapy, material inheritance and contraceptive pills. Due to its generality and importance, many researches have been conducted from distinct labs about this illness and its

plausible causes have been intensively investigated either in microarray studies, where just part of the related genes are detected, or in the pairwise correlation analyses between the disease and selected symptoms via contingency tables (Jönsson, J-M., 2015).

Gynecologic cancers start in various places located in female's pelvis, which is the region beneath the stomach and in between the hip bones. Each type is exclusive, with varying tokens and symptoms, risk factors and ways to lower the probability. All women have the potential for gynecologic cancers, and probability surges with age. When gynecologic cancers are diagnosed earlier, treatment is more efficient.

2.5 Ovarian Cancer

Different from cervical cancer and endometrial most cancers, epithelial ovarian most cancers does not have a typically agreed identifiable precursor lesion. In addition, most research of molecular genetic changes in ovarian cancer were performed with tissue from advanced-degree neoplasms. For these motives, it turned harder to differentiate among molecular genetic changes that are often related to disorder development and changes which can actually replicate the inherent genetic instability of advanced neoplasms. Regardless of those problems, some genetic alterations had been discovered in epithelial ovarian cancers (Zhang et al., 2016).

Most of ovarian cancers are sporadic and pop out from an accumulation of genetic damage over years. Ovarian cancers are heterogeneous with admire to histopathology and malignant capacity; consequently, the function patterns of molecular signature is likewise no longer uniform. Serous borderline tumor and low-grade serous adenocarcinoma are regularly characterized via mutations in K-RAS and BRAF.

Sporadic high-grade serous adenocarcinomas are normally advanced at diagnosis and the forecasting for patients is weak. Ovarian cancer has common mutations in

TP53 and sometimes higher expression of HER-2/neu, AKT2 and MYC. Endometrioid and clear cell adenocarcinomas have been found to relate with endometriosis and acknowledged to have common mutations of PTEN and PIK3CA. Mucinous adenocarcinoma is normally recognized at an early level of illness and characterised via mutations in K-RAS. In general, ~10% of ovarian cancers are due to inherited mutations of most cancers susceptible genes, like BRCA1 and BRCA2 and high-grade serous adenocarcinoma is the most important histological kind in patients with inherited BRCA mutations (Sonoda, 2016).

BRCA1 and BRCA2 are located in chromosome 17q and 13q, respectively. Because the BRCA1 and BRCA2 proteins complex with Rad51 recombinase different molecules which have roles in the repair of DNA double strand breaks through homologous recombination, BRCA1 and BRCA2 are labeled as tumor suppressor genes. The lifetime chance of ovarian cancer tiers among 20–40% and 10–20% in BRCA1 and BRCA2 carriers, respectively, and the median age of patients is 40–50 years old. Poly(adenosine diphosphate-ribose) polymerase (PARP) is involved within the restore of DNA singlestrand breaks through base excision.

Mutations in KRAS and BRAF that cause aberrant activation have been identified in both endometrial and ovarian cancer and appear to play a central role in carcinogenesis by conducting signals that enhance cell proliferation during tumor development. RAS, a small GTP binding protein, activates the core unit of a cascade composed of RAF, mitogen/extracellular signal-regulated kinase (MEK1/2) and MAP Kinase (MAPK or ERK) as well as the PI3K/ AKT pathway (Sonoda, 2016).

Inactivating mutations of the tumor suppressor gene, PTEN, are detected in both endometrial and ovarian cancer, PTEN is an inhibitor of PI3K/ AKT signaling and acts to control the rate of cell division and promote apoptosis. Gain of function

mutations of the CTNNB1 gene (β-catenin) are identified in endometrial and ovarian cancer especially those with squamous differentiation.

High-grade serous ovarian carcinoma is one of the most fatal gynecological cancers. All patients diagnosed with advanced disease experience fundamentally the same as traditional treatment, which is aggressive surgery pursued by multi-cycles of platinum-based combination chemotherapy. In any case, about 30% of cases show chemoresistance and gain practically no benefit. High percentage of chemosensitive patients create acquired resistance and in the long run relapse within various timeframes unfortunately. That's the reason, it is critical to grow new tools to classify fundamentally the same as or identical patients and convince them to start progressively high resolution therapies that might be conceivably better. Gene expression profiling has been connected to the study of ovarian cancer. Studies have focused on differential gene expression among tumor and normal, differentiating between histological subtypes and identifying differences among invasive and low malignant potential tumors (Kondrashova et al., 2018).

Germline mutation in either BRCA1 or BRCA2 is linked with a higher risk of ovarian cancer, especially the most common invasive histotype — serous carcinoma. Also, serous ovarian cancers have an exceptionally high prevalence of other molecular events involving BRCA pathway dysfunction. Latest findings depict increased frequency of TP53 mutation, chromosomal instability, distinct molecular subtypes and DNA copy number-driven alterations in gene expression. These confer a model in which homologous recombination repair deficiency initiates an avalanche of molecular pathways which shapes the development of high-grade serous ovarian cancer and decide about its reaction towards therapy.

HG-SOC have increased level of DNA amplifications and deletions p53 loss and BRCA loss, leading to a deficiency in homologous recombination repair (HRR) of DNA double-strand breaks, initiate the chromosomal instability and widespread copy number (CN) changes that are common properties for high-grade serous ovarian cancers. Copy number change can be a driver of molecular subtype

specification and leads to serious changes in gene expression. Subsequent mutations provide further advantages for tumour growth but may not be molecular subtype specific. Clearly, there are many recurrent gain-of-function and loss-of-function mutations in HG-SOC other than those that have been implicated so far.

2.6. Normalization and Filtering of Data

We used the dataset of Mok et al. (2009) who performed a genome wide gene expression profile of papillary serous ovarian carcinoma using an oligonucleotide array containing 47 000 transcripts and have identified 1191 sequences that are differentially regulated between cancer and normal specimens with a significance of p-value<0.001. The team validated the microarray analysis comparing our differentially regulated genes with those from other microarray studies on serous ovarian cancer. Their results were consistent with those where the data sets overlapped. In addition, they identified a number of genes that have been independently identified as differentially expressed in ovarian cancer specimens and/or cell lines by Western, Northern and immunohistochemical techniques. All normal ovarian samples were obtained from post-menopausal women. Specimens were procured under IRB-approved protocols. All tissue samples were stored at −140°C until processed.Total RNA from each sample was extracted using Trizol (Life Technologies, Inc., Gaithersburg, MD, USA) as per the manufacturer's instructions, followed by purification using RNeasy Mini columns (Qiagen, Inc., Valencia, CA, USA). To enhance the total RNA yield, OSE samples were purified on an Rneasy Micro column (Qiagen, Inc., Valencia, CA, USA) after Trizol extraction. Human Genome U133A Plus 2.0. GeneChip oligonucleotide arrays (Affymetrix, Santa Clara, CA, USA) representing 47 000 transcripts and variants, including 38 500 well-characterized human genes, were used in this study. Biotin-labeled cRNA was prepared as described in the Affymetrix Expression Analysis Technical Manual (Affymetrix, Santa Clara, CA, USA). Briefly, 5 μg of purified total RNA template was reverse transcribed to generate double-stranded cDNA using HPLC-purified T7-$(dT)_{24}$ primer (Midland Certified Reagent Company, Inc., Midland, TX, USA) and Superscript II™ RNAse H^- reverse transcriptase

(Invitrogen Life Technologies Corporation, Carlsbad, CA, USA). Following second-strand cDNA synthesis and clean-up, biotinylated antisense RNA (aRNA) was generated by in vitro transcription using the Bioarray, High Yield RNA Transcript Labeling Kit (Enzo Diagnostics, Farmingdale, NY, USA). In all, 15 μg of each RNA preparation was fragmented and combined with a hybridization cocktail containing four biotinylated hybridization controls (BioB, BioC, BioD and Cre). Hybridization to the oligonucleotide arrays and subsequent washing and detection was performed as recommended by the manufacturer. Array images were acquired using a GeneChip Scanner 3000 (Affymetrix, Santa Clara, CA, USA) and analysed with Genechip® Operating Software (GCOS).

The authors identified 53 advanced stage, high-grade primary tumor specimens from patients with papillary serous adenocarcinomas. The average age for the cohort was 61.9 years (SD = 12.7), with an average survival time of 40.5 months following surgery (SD = 41.3 months). 10 normal healthy individuals results were also available. All specimens were subjected to laser-based microdissection and analyzed as pure, microdissected epithelial cell populations on whole-genome Affymetrix U133 Plus 2.0 GeneChip microarrays. Total of 53 snap-frozen tissue specimens and a validation set of 64 paraffin-embedded tumor samples were obtained from previously untreated ovarian cancer patients, who were hospitalized at the Brigham and Women's Hospital between 1990 and 2000. All patients had advanced stage, high-grade serous ovarian

cancer according to the International Federation of Gynecology and Obstetrics standards. All specimens and their corresponding clinical information were collected under protocols approved by the institutional review boards of the corresponding institutions. To demonstrate that pure populations of epithelial cells were obtained, they checked our arrays for the expression of endothelial cell markers (TIE-2 and VEGFR2) and T cell markers (CD8 and CD45). Low-level microarray survival analysis included array normalization and estimation of expression level using an invariant set of probe sets to adjust the overall signal

level of the arrays to the same level and then a model-based PM (perfect match)-only approach established gene expression levels using dChip software.

Each Affymetrix oligonucleotide array has a small string of DNA 25 base pairs long and used to measure amount of transcription for each gene. Each gene is defined by a set of 11 to 20 probe pairs, coming from varying regions of gene's DNA sequence. The probe denotes the single strand of a gene sequence segment attached to array surface. There are 2 components to each probe pair. First one is perfect match (PM) whose value shows the amount of transcribed perfectly matched target complementary to mRNA of the gene of interest. Other is mismatch (MM) whose value indicates amount of non –specific binding of target by altering the 13^{th} base pair of probe. To measure transcribed RNA, targets are labelled and hybridized to an array. During hybridization, part of target sequences can bind to non complementary transcripts, causing bigger intensity values of each probe. This nuisance signal is named as non specific hybridization. Apart from this two types of variational signals exist. First one is background signal: Signal measured by probe in absence of a complementary target DNA, thus totally independent of any true signal. Second non specific signal happens due to binding to surface of slide, instead of probes.

There are some methods to describe gene expression level (gene expression index preferably). MAS 50, dChip, RMA, GC-RMA are some these methods. Among many alternatşves, we utilized RMA (Robust Multi- array Average) (Irizarry et al., 2003) since it performs well without losing much information in data. Furthermore, it is the first method that critically reasseses the mismatch probe values as measure of non specific hybridization and should be ignored altogether. RMA uses a model-based background correction, quantile normalization and a robust averaging expression summary method. Thus, the properties of the RMA can be listed as:

- Intensity values range between 4 and 16,
- Intensity values are in log (base 2) scale,

- Only perfect match intensities are used,
- Variance is smaller and relatively stable across the range of intensities,
- Fold-changes are underestimated.

On the other hand, the BRB-ArrayTool is an integrated package for the visualization and statistical analysis of microarray gene expression and also used in copy number alterations and methylation changes. It was developed by proffessional statisticians experienced in the analysis of microarray data and involved in the development of improved methods for the design and analysis of microarray based experiments. The analytic and visualization tools are integrated into Excel as a package. The analytic and visualization tools themselves are developed in the powerful R statistical system, in C and Fortran programs and in Java applications. Apart from using it for the normalization, we also did the logfold change =2 filtering on BRB with the help of scatter plot traits and seperating the 53 patients and 10 normals apart as different file types (experimental vs control) provided by the BRB package itself.

These two steps gave us nearly 1100 canditate genes so now we could start analysis of those genes in 53 patients. We subsampled or in a way grouped these genes by setting sample size(gene number) to 10, 23 and 50, respectively, and having 3 datasets minimum for each group. These mini datasets were directly used for GGM network construction and precision matrix formation on the platform R with related libraries as well as performing neural network construction in R.

2.7 STRING (Search Tool for Retrieval of Interacting Genes/Proteins)

We have increasingly come to recognize that cellular regulatory processes are more complex than we had once imagined and that it is generally not individual genes, but networks of interacting genes and gene products, which collectively interact to define phenotypes and the alterations that occur in the development of a disease.

Gene networks are often described verbally in combination with figures to illustrate sometimes-complicated interrelations between network elements. Nowadays,

molecular biological methods and high-throughput technologies make it possible to study a large number of genes and proteins in parallel enabling the study of larger gene networks.

In the case of transcript data, STRING re-processes and maps the large number of experiments stored in the NCBI Gene Expression Omnibus followed by normalization, redundancy reduction and Pearson correlation (described iractions, often including annotated pathway knowledge, text-mining results, inter-organism transfers or other accessory information. The STRING database ('Search Tool for Retrieval of Interacting Genes/Proteins') belongs to this latter class. STRING is one of the earliest efforts and strives to differentiate itself mainly through high coverage, ease of use and a consistent scoring system. It currently features the largest number of organisms (5090) and proteins (24.6 million), has very broad and diverse, benchmarked data sources and provides intuitive and fast viewers for online use (Szklarczyk et al., 2018). When we formed the mini datasets out of 1100 canditate genes ,we used the quasi true networks provided by STRING when co-expression, co-occurence and being in the same sort of databases were selected as parameters after entering the gene names.

2.8 Accuracy Measures

In the analyses, we compare the performance of different models via distinct accuracy measures. Among many alternates, in this study, we applied the most common ones, namely, sensitivity and specificity. Both measures can control the accuracy of the binary classification. So, they are based on the results of the following values:

- True positive (TP): This number shows the correctly specified objects that have positive label.
- True negative (TN): This number shows the correctly classified as negative when they indicates the actually negative objects.

- False positive (FP): This number shows the misclassified objects that have labeled as wrongly positive.
- False negative (FN): This number shows the misclassified objects that have negative label.

Thus, the expression of the sensitivity and specificity can be represented as below:

Sensitivity= TP / (TP+FN) and Specificity= TN / (TN+FP).

2.9 Gaussian Graphical Model

Before conducting a network analysis, the researcher must decide on the measure of association to use: What do we mean by gene-gene association? In many modern methods, definition of association is made for us. However, it's important to make this choice based on the context of the biological question. Gaussian graphical model (GGM) captures conditional relationships that are typically visualized to infer the underlying conditional (in)dependence structure, i.e., the "network". The undirected graph is G = (V, E), and includes a vertex set V = {1, ..., p} as well as an edge set $E \subset V \times V$. Let $Y = (Y_1, ..., Y_p)$ be a random vector indexed by the graphs vertices, of dimension p, that is assumed to follow a multivariate normal distribution $N_p(\mu, \Sigma)$, with the mean vector and covariance matrix Σ.

GGMs (Friedman et al., 2007; Toh and Horimoto, 2002) are closely linked to the precision matrix $\Theta = \Sigma^{-1}$, which describes the graphical structure of the corresponding Gaussian graph. In the high dimensional setting, the challenge for gene association networks arises is to obtain reliable estimate for the population covariance matrix. This problem comes from the actual nature of the genetic networks, which has a massive number of variables, yet very much less samples. Thus, the empirical covariance matrix, S, cannot be used as an unbiased predictor to obtain the population covariance matrix. Furthermore, it is clearly difficult to construct biological networks through GGM in high-dimensional setting, in which

the number of variables are found to be much larger than the sample size. The basic idea behind it, is that high-dimensional biological data are sparse in the sense that only a small number of genes will regulate one specific gene of interest. This scenario leads to the construction of an undirected graph of conditional dependencies which is sparser than a correlation network.

So as in many applications including biological data, the number of variables (genes in this case) is much larger than the number of observations, this problem is seen in the analyses of the microarray gene expression datasets with a view to investigate the relation between pairs of genes. In such a case, the maximum likelihood estimate of the inverse covariance matrix does not exist. To overcome this problem, there has been a great deal of interest in L_1 - regularization, such as LASSO (Friedman et al., 2007; 2010) which is used for estimating the sparse inverse covariance matrix. Lastly, Friedman et al. (2010) studied the graphical lasso, which involves maximizing the penalized log-likelihood function via the LASSO. Among several alternative methods, the graphical LASSO (glasso) has been the most popular because of its computational efficiency and desirable statistical properties in high-dimensional datasets. Because without loss of generality, let $\mu = 0$ and $Q = \Sigma^{-1}$, the covariance matrix indicates a correlation network and the zero elements $\Sigma i_{, j} = 0$ indicates that two variables are only conditionally independent given the observations of other nodes. Also, this means graphically there is no edge connecting the two nodes, making it much more convenient to be represented.

Accordingly, in the LASSO regression, the target node is regressed on other nodes in each time and this process is repeated for each node in the graph. Mathematically, this regression problem at each node is formulated as solving the following problem:

$$\beta_1 = \arg\min_{\beta_1} \|Y - X\beta_1\|_2 + \lambda\|\beta_1\|_1, \quad (1)$$

where Y is an N dimensional vector with each entry being a node at a specific dimension from the N sets of data observations. Furthermore, $\| \cdot \|_1$ and $\| \cdot \|_2$ represents the L_1 –norm and L_2 – norm of the given term. On the other hand, in the representation of the LASSO regression above, each set of data observation has p nodes. Thus, X is an $N \times (p-1)$-dimensional matrix with each column being the rest of the nodes in the N sets of observations. Additionally, λ denotes the penalty constant which is near 0. Thus, if λ is bigger the network becomes sparser. Finally from Equation 1, it is seen that λ controls β_1. So, β_1 is a $(p - 1)$ - dimensional coefficient weight vector one wants to estimate. Accordingly, the last L_1 - term in the equation is the sparse promoting penalty term that enforces sparsity.

Ideally, we want to use $\|\beta_1\|_0$ which is the L_0th term to minimize the number of supports. Unfortunately, this makes the problem hard to solve. Thus, we need to focus on the L_1 case. Hence, in the estimation of β_1 via Equation 1, we have the multivariate Gaussian density function which can be written as:

$$p(x|\mu, \Sigma) = |Q| \; 1/2 \; (2\pi) \; n/2 \; \exp(-(1/2) \; (x - \mu) > Q(x - \mu)). \quad (2)$$

In this expression, we want to estimate the precision matrix Q via an L_1-regularized maximum likelihood estimation method such that Q is sparse. Accordingly, under the multivariate normality of assumption of Y, while Y represents a joint multivariate vector as $Y = (Y_{-p}, Y_p)$ and $Y_{-p} = (Y_1, \cdots, Y_{p-1})$ showing that the vector has all nodes, but not the final one. Thus, the conditional distribution for Yp is formed as $Yp|Y-p = y$ and

$$y \sim N(\mu p + (y - \mu_{(-p)}) \; (\Sigma^{-1})'_{(-p,p)} \; \sigma_{(-p,p)} \; , \; \sigma(p,p) - \sigma_{(-p,p)} \; (\Sigma^{-1})`_{(-p,-p)} \; \sigma_{(-p,p)}) \quad (3)$$

in which $\Sigma_{(-p,-p)}$ refers to $((p - 1) \times (p - 1))$ - dimensional covariance matrix except the last nodes and $\sigma_{(-p,p)}$ and $\sigma_{(p,p)}$ indicate $((p - 1) \times 1)$ - dimensional covariance

vector associated with Y_{-p} and Yp, and variance for Y_p, respectively. Moreover (.)` shows the transpose of the given matrix. Hereby, the GGM regression has the regression coefficients $\beta = (\Sigma^{-1})` _{(-p,-p)} \sigma_{(-p,p)}$ that makes the conditional independence property valid. So $\beta_j = 0$ means that Y_p and Y_j are conditionally independent given all the other nodes. LASSO aims to shrink the coefficients towards zero by applying a regularization parameter as a constraint. Through this, it tends to decrease the complexity of the network by variable selection and shrinkage.

2.9.1 R Programming for GGM

Gene expression profiling provides unprecedented opportunities to study patterns of gene expression regulation, for example, in diseases or developmental processes. Bioinformatics analysis plays an important part of processing the information embedded in large-scale expression profiling studies and for laying the foundation for biological interpretation. Over the past years, numerous tools have emerged for microarray data analysis. One of the most popular platforms is R, an open source and open development software project for the analysis and comprehension of genomic data, based on the R programming language. As R has many functions available in a single environment, minimum effort is needed to write programs for data handling. One can, then, concentrate on the statistical algorithm and analysis (Zhao and Tan, 2006).

Hereby, in the analysis via GGM, we used the following R packages:

- Huge: a general framework for high-dimensional undirected graph estimation. It integrates data preprocessing, neighborhood screening, graph estimation, and model selection techniques into a pipeline.

- glasso: Estimation of a sparse inverse covariance matrix using a LASSO (L_1) penalty. Learning the structure of GGMs from data that

contain the measurements on a set of variables across samples has significantly facilitated data-driven discovery in a diverse set of scientific fields. For example, biologists can gain insights into how thousands of genes interact with each other in various disease processes by learning the GGM structure from gene expression data that measure the mRNA expression levels of genes across hundreds of patients. The graphical lasso approach attempts to learn the structure of a Gaussian graphical model (GGM) by maximizing the log likelihood of the data, subject to an L_1-penalty on the elements of the inverse covariance matrix(precision matrix) as described in the prevous part.

- Qgraph: main function of qgraph is creating an appropriate network and sending it to the plotting method.

2.9 Artifical Neural Network

Tremendous growth in current data encouraged researchers among the areas of statistics, artificial intelligence, and data mining to construct basic, flexible, strong procedures for data modeling that may be used in really big data sets.

The artifical neural network (ANN) model (Bishop, 1995; Kosko, 1992) can be named as a nonparametric statistical model which generally works better in estimating future responses when compared to a standard regression model. The single layer ANN consists of an input layer, a single hidden layer and one output layer. The nodes in the input layer correspond to genes in our study. Below, we explain each element of ANN with details. In Figure 2.1, we draw this structure in simple graph.

1. **Input Nodes –** The Input nodes provide information from the outside world to the network and are together referred to as the “Input Layer”. No

computation is performed in any of the Input nodes – they just pass on the information to the hidden nodes.

2. **Hidden Nodes –** The Hidden nodes have no direct connection with the outside world. They perform computations and transfer information from the input nodes to the output nodes. A collection of hidden nodes forms a "Hidden Layer".
3. **Output Nodes –** The Output nodes are collectively referred to as the "Output Layer" and are responsible for computations and transferring information from the network to the outside world.

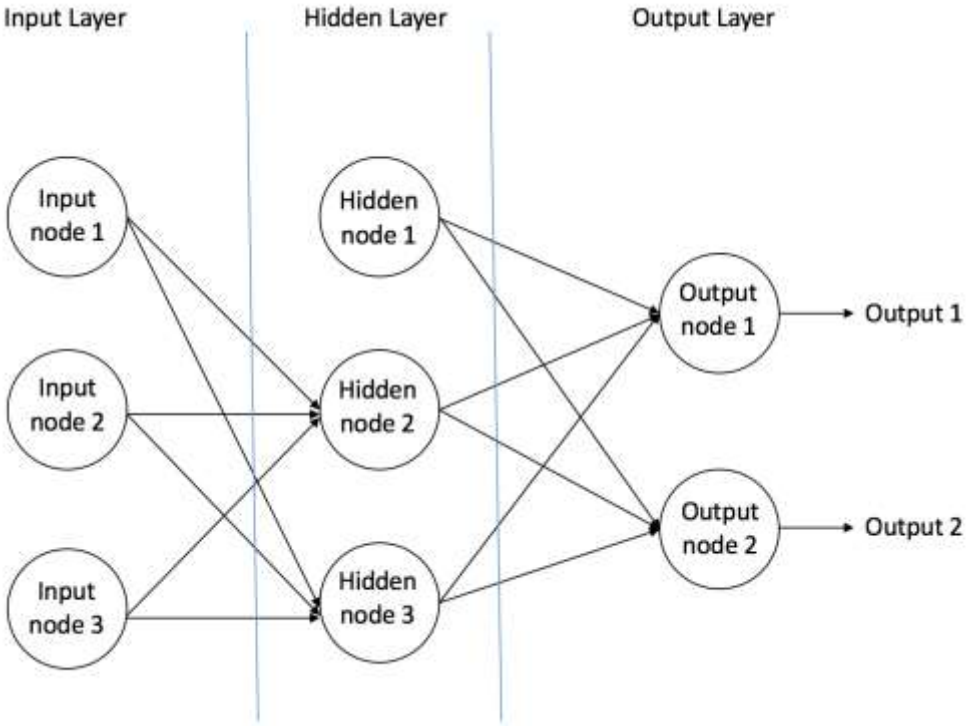


Figure 2.9.1. Scheme of a feedforward artifical neural network.

Hence, the purpose of the activation function is to introduce non-linearity into the output of a neuron. This is important because most real world data is non-linear and we want neurons to learn these non-linear representations. The feedforward neural

network is the first and simplest type of artificial neural network designed and the one which we used.

Accordingly, as an alternative method to construct gene networks and understand how they interact, we proposed a non-parametric method which has been commonly used recently in many different areas. The ANN approach does not require any assumptions like GGM, but, indeed it performs highly complex mathematical procedures during the network formation even if there are 10 or 25 genes in the dataset. Working univariately (opposed to GGM logic), which meant attempting to write every single gene in terms of others in the dataset requires many more steps than predicted. Fortunately, R programming worked efficiently with qualified libraries. The ANN models are generally preferred to study big data but it can also be implemented in small or moderate datasets as well. In this thesis, we originally adapted this complex model as the alterate of GGM in the sense that each output layer was taken as a response similar to Equation 1 and other nodes apart from reposnse node, were presented as the inputs layes. Then, by letting both linear and non-linear functional forms of weighted coefficients of the activation function which can be considered as the regression coefficient in GGM, we constructed a regression for each gene/node over other nodes, again similar to the GGM definition.

In a simplest ANN model, the single-hidden-layer, feedforward neural network is the one which we used with R programming platform with 2 necessary libraries called “catools “and “neuralnet”. The catools package can be implemented for data partitioning. Because in the computation of the activation function and associated weights, the data are partitioned as train and test sets. The former is used to estimate the weights and the latter is applied to control the performance of the estimated weights and model. This ratio can be taken differently per data like 10 %, 20% or 30% testing ratio. Then, the train data is scaled for each variable so that they can be invariate from the range of the measurements. In our analysis, we scaled each observation via the range of the associated variable by computing max-min of observation per variable and then, dividing each observation to this number.

By this way, each observation within each variable can be comparable and can lie within the range of 0 to 1.

Later, we called the NeuralNeT package in R to construct the estimated model. In this modeling, we performed the single-layer neural network which presents the simplest form of ANN where there is only one layer of input nodes that send weighted inputs to a subsequent layer of receiving nodes, or in some cases, one receiving node. The single-layer neural networks can also be thought of as part of a class of feedforward neural networks, where information only travels in one direction, through the inputs to the output.

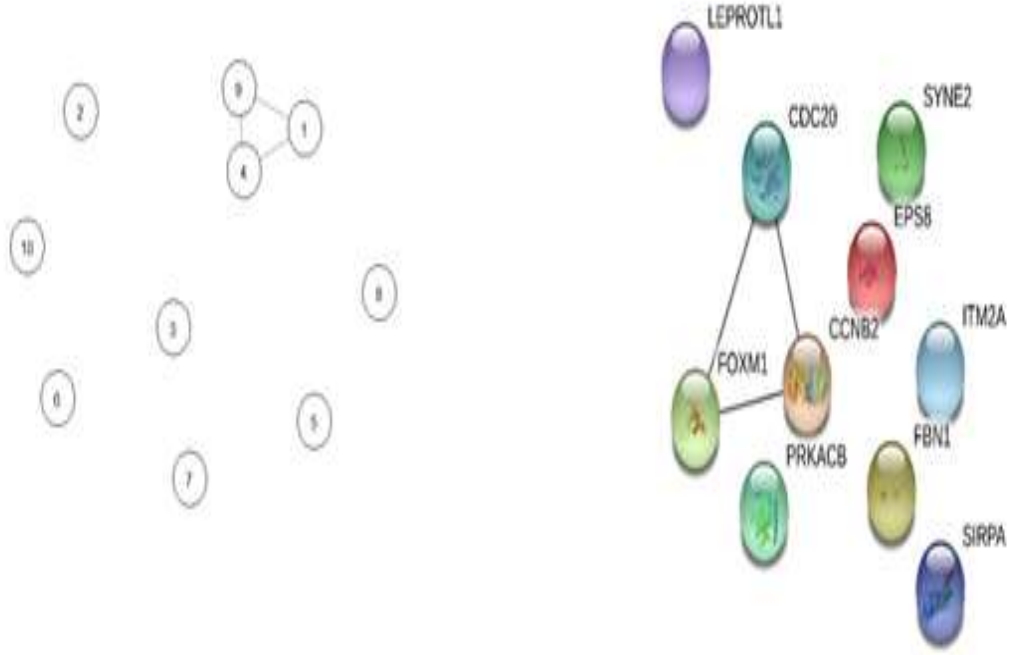
CHAPTER 3

RESULTS

We used the data from Mok et al. (2009) with accession number GSE18520 in GEO. Before we move on with the fold change filtering to come up with differentially expressed genes from 63 arrays (53 patients and 10 normal), we had to perform normalization. Here, the RMA method was done by using BRB array tool. In the analyses, the first 60 genes from 1100 canditates, were grouped into 10 genes. In Table 3.1, we presented the list of genes in each 10-dimensional network. Among these systems, in Figure 3.1, we shown the estimated and true networks for the 6th dataset for illustration. In this figure, the enumerated genes 1, 9, and 4 in the GGM estimated system denote FOXM1, CDC20 and CCNB2 genes, respectively.

Furthermore, from the biological interpretation of this system, we obtained the following outputs: Cyclin B2 (CCNB2) is a belongs to cyclin family, to be more specific- the B-type cyclins. It is found that stronger expression of cyclin B2 mRNA in tumor cells was an independent indicator of a poor prognosis in patients having adenocarcinoma of the lung . Apart from that, it is also reported that CCNB2 may serve as an oncogene and might be as a canditate biomarker of weak prognosis over short-term follow-up in breast cancer. Cell Division Cycle 20 (CDC20), serve as regulatory protein interacting with some other proteins at several points in the cell cycle. Many researches showed that CDC20 can function a promising biomarker for both therapeutic target and prognosis in many cancers . Glioblastoma is the most commo and deadly major intrinsic brain tumor. Glioblastoma shows hierarchical organization with a group of self-renewing and tumorigenic glioma tumor initiating cells (TICs), or cancer stem cells. Whereas, non-neoplastic neural stem cells are mostly quiescent, glioblastoma TICs are generally proliferative with mitotic control giving a potential problem of fragility.

Repression of CDC20 stabilizes $p21^{CIP1/WAF1}$, causing inhibition of some genes vital to tumor growth and survival, like CDC25C, c-Myc and Survivin. Transcriptional control of CDC20 is controlled by FOXM1, a primary transcription factor in TICs. This implies that CDC20 is a vital regulator of TIC proliferation and survival, bridging two key TIC nodes – FOXM1 and $p21^{CIP1/WAF1}$ — suggesting a ciritical point for therapeutic intervention.



(a) (b)

Figure 3.1: Corresponding networks of the 6th dataset composed of 10 genes and (a) estimated by GGM and (b) the structure of its true network in STRING.

Table 3.1: List of genes for each 10-dimensional network.

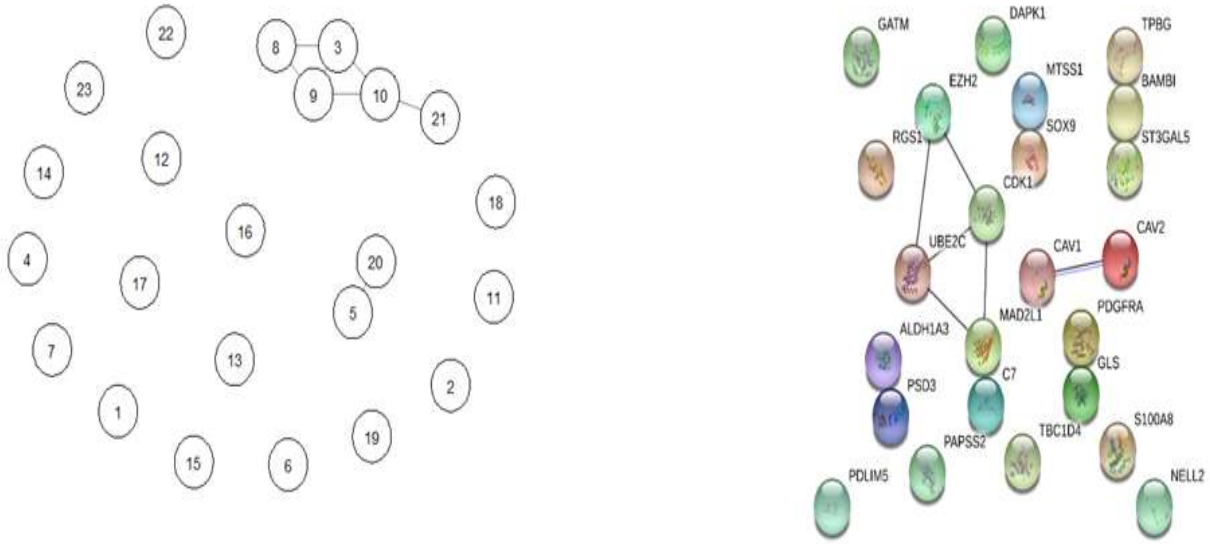
Name of Dataset	**List of Genes**
Data 1	CTCFL, BNC1, KLHDC1, PROM2, SIGLEC11, LRRN4, CBR4, CBS, NT5E, BICD1
Data 2	LINS1, TMEM37, FAM13C, TACC1, ARHGEF10, C1S, CFI, LYZ, DNAJC7, NR2F1-AS1
Data 3	SERPINB6, WDR17, KIDINS220, C8orf88, CP, ZNF493, FBXO22, LOC613266, C16orf54, LOC646762
Data 4	DLGAP1, PRKAR1A, DPYSL2, ANXA5, SPARCL1, TACC1, CCND2, SERPING1, TXNIP
Data 5	CPE, PJA2, GPNMB, TIMP3, HTRA1, PLS3, SLC2A1, DAB2, TOP2A, WSB1
Data 6	FOXM1, LEPROTL1, EPS8, CCNB2, PRKACB, ITM2A, SYNE2, FBN1, CDC20, SIRPA

From the analyses of these 6 datasets, the computed accuracy measures for GGM are presented in Table 3.2. From the results, it is seen that GGM is successful in both sensitivity and specificity where it is calculated. Because in certain networks, as there is no interaction between genes, the sensitivity cannot be applicable.

Table 3.2: Results of accuracy measures for 10-dimensional datasets via GGM. NA denotes not avalaible.

Accuracy Measures	**Data 1**	**Data 2**	**Data 3**	**Data 4**	**Data 5**	**Data 6**
Sensitivity	NA	NA	NA	NA	NA	0
Specificity	1	1	0.92	1	1	1

On the other hand, when we analyzed the 23 genes groups, we used totally 6 datasets. Among them, we illustrated the findings of Data 7 and Data 8 with their true network structures in STRING. The network from 7th data belonging to GGM showed interactions among gene numbers 8-9-3-10-21 which corresponds to the group containing CDC20 and CDK1, whereas 5-20 represents CAV1 and CAV2.

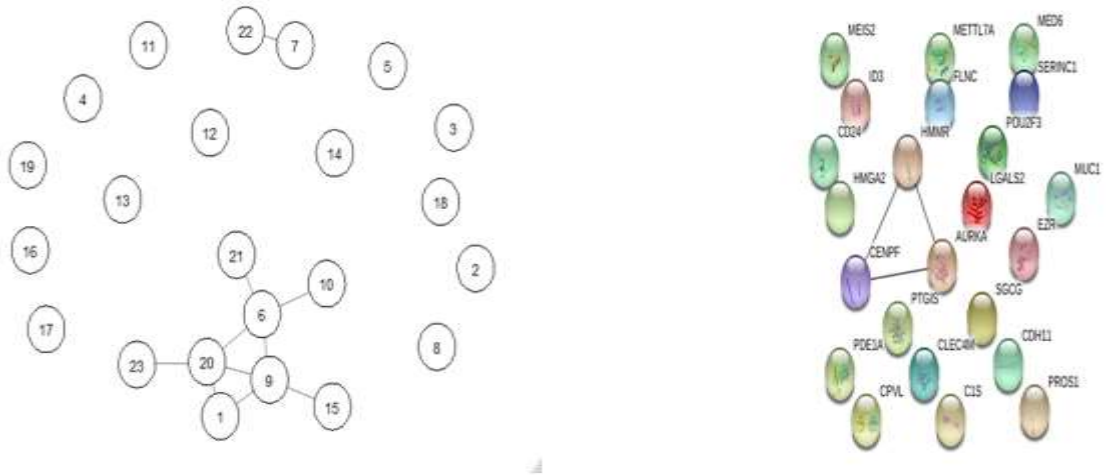


(a) (b)

Figure 3.2: Corresponding networks of the 7th dataset composed of 23 genes and (a) estimated by GGM and (b) the structure of its true network in STRING.

Regarding the outcomes from Figure 3.2, it is seen that the enhancer of zeste homologue 2 (EZH2) is the catalytic component of Polycomb repressive complex 2 (PRC2) and catalyses the trimethylation of histone H3 on Lys 27 (H3K27), that supresses gene transcription. EZH2 augments cancer-cell invasiveness and modulates stem cell differentiation. It was shown that EZH2 may be phosphorylated at Thr 487 via activation of cyclin-dependent kinase 1 (CDK1). The phosphorylation of EZH2 at Thr 487 causes an impairment regarding EZH2 binding to other PRC2 components SUZ12 and EED, thus, inhibition of EZH2 methyltransferase activity, leading to stalling of cancer-cell invasion. In human mesenchymal stem cells, the activation of CDK1 triggers the mesenchymal stem cell differentiation into osteoblasts by phosphorylation of EZH2 at Thr 487. This implies a signaling association between CDK1 and EZH2 which can mean a vital function in several different biological cascades, such as cancer-cell invasion and osteogenic differentiation of mesenchymal stem cells. CAV1 (caveolin-1) and CAV 2 (caveolin 2) are the main structural proteins of caveolae, sphingolipid and cholesterol-rich invaginations of the plasma membrane functioning in vesicular transport and signal transduction. Lately, a debate started about their function in breast cancer and their potency as markers of basal-like phenotype. CAV1 and CAV2 protein expression were examined on a tissue microarray containing 880 unselected invasive breast cancer cases, through immunohistochemistry. CAV1 and CAV2 expression were observed in 13.4 and 5.9% of all breast cancer, respectively. Their expression was highly linked with high histological grade, absence of steroid hormone receptor positivity (ER and PR), and expression of basal markers (basal cytokeratins, P63, P-cadherin).

On the other side, the results of Data 8 is represented below. In this system, number 6 represents AURKA, 20 is for HMMR and CENPF is shown by 9.



(a) (b)

Figure 3.3: Corresponding networks of the 8th dataset composed of 23 genes and (a) estimated by GGM and (b) the structure of its true network in STRING.

From the comparison of estimated and true structue of the system 8, it was found out that the HMMR mislocalization is due to an indirect impact modulated by the overexpression of AURKA. Repressing of HMMR, elevates the AURKA activity. Also, it was revealed that the AURKA activity is vital in division and self-renewal of sphere-enriched MPNST cancer stem-like cells. To continue, supressing HMMR was enough to give MPNST cells the skill to start and sustain sphere culture. Many scientists claim that AURKA could function as a step in therapy related to MPNST and tumour cell reactions to AKI, that can be differentiation, are regulated via the affluence of HMMR.

We listed the complete list of genes for all 23-dimensional systems in Table 3.3. In Table 3.4, we reported the sensitivity and specificity values obained from GGM for these 6 datasets. The results support the previoıs findings in the sense that GGM can successfully infer the systems.

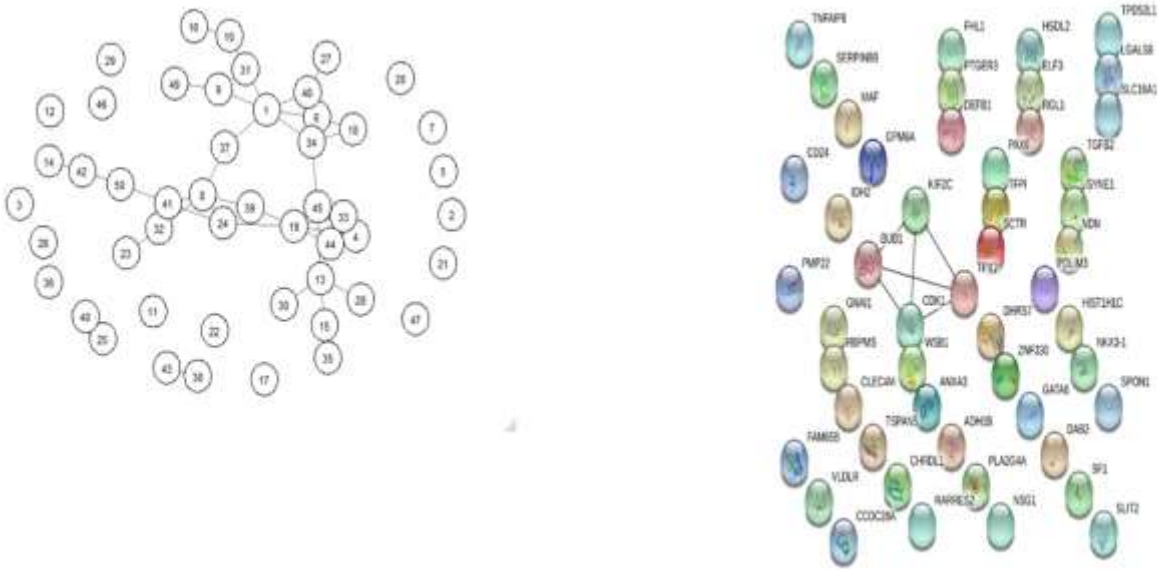
Table 3.3: List of genes for each 23-dimensional network.

Name of Dataset	**List of Genes**
Data 7	S100A8, SOX9, UBE2C, RGS1, C7, MTSS1, PAPSS2, CAV1, PDGFRA, DAPK1, GLS, GATM, ALDH1A3, CDK1, ST3GAL5, PDLIM5, BAMBI, CAV2, PSD3 EZH2, MAD2L1, TBC1D4, NELL2, TPBG
Data 8	MEIS2, METTL7A, MEDG, ID3, FLNC, SERINC1, POU2F3, HMMR, HMGA2, CENPF, AURKA, EZR, SGCG, PTGIS, PDE1A, CLEC4M, CDH11, CPVL, C1S, PROS1, CD24, LGALS2, MUC1
Data 9	S100A8, SOX9, UBE2C, RGS1, C7, MTSS1, PAPSS2, CAV1, PDGFRA, DAPK1, GLS, GATM, ALDH1A3, CDK1, ST3GAL5, PDLIM5, BAMBI, CAV2, PSD3, EZH2, MAD2L1, TBC1D4, NELL2
Data 10	DMD, THBD, SCG5, WFDC2, SLC4A4, KDR, NT5E, CLDN3, FABP4, FCGR3B, ZWINT, TRIP13, MAOB, WASF3, ME1, PRKX, FRY
Data 11	HSD17B2, TTK, NAV3, MELK, FGL2, GLDC, CP, IL6ST, TLE4, PTGER4, TCF21, ALDH3B2, MNDA, DOCK4, MAP3K8, EFNB3, MPDZ, AOX1, GFPT2, FGF13, PLCE1, FGF1, RNASE4
Data 12	WNT5A, PRG4, DSC3, ECM2, GINS1, KLK8, LHX2, PTX3, ARHGAP6, MEOX2, IL18, IFI16, CXCL6, MAF, RARRES1, FGF9, WNT2B, SMARCA2, LY96, NR0B1, HP, EPB41L3, PTH2R

Table 3.4: Results of accuracy measures for 23-dimensional datasets via GGM.

Accuracy Measures	**Data 7**	**Data 8**	**Data 9**	**Data 10**	**Data 11**	**Data 12**
Sensitivity	0	0	0	0	0	0
Specificity	1	1	1	1	1	0.913

Finally, we evaluated the results based on 50-dimensional systems. Hereby, when 50 genes including subsets were analyzed, some similarities validated by previous researches also found. We described the estimated system and its true structure of Data 13 in Figure 3.4 for illustration and discussed the biological findings for this system as below.



(a) (b)

Figure 3.4: Corresponding networks of the 13th dataset composed of 50 genes and (a) estimated by GGM and (b) the structure of its true network in STRING.

From Figure 3.4, it is found that Cytoplasmic CDK1 overexpression is linked with cancer growth and weak overall survival in 249 epithelial ovarian cancers (EOC)

before. CDK1 is known to be a canditate target of transcription factors to adjust paclitaxel resistance in EOC patients. Overexpression of BUB1 was also observed in non-small cell lung and breast cancer . TPX2 (Targeting Protein for Xklp2) is a hub gene as well. Despite the fact that, the role of TPX2 in EOC pathology is unknown, it was revealed as a biomarker of low survival in 143 EOC patients. In cervical cancer, the expression of TPX2 is associated with histological grading and lymph node metastasis. Additionally, TPX2 was identified as a target gene of microRNA-491 in esophageal cancer and serves a major role in cancer cell invasion in both esophageal and colon cancer. Mutational analysis performed so far implies that the mitotic phosphorylation of importin-α1 inhibits its binding to importin-β and allows the discharge of TPX2 which is later targeted like importin-β to the spindle. Losing importin-α1 or expression of a non-phosphorylated mutant of importin-α1 causes construction of shortened spindles having a lower microtubule density and promotes a longer metaphase, while phosphorylation-mimicking mutants are functional in mitosis. It is suggested that phosphorylation of importin-α1 is a common process for the spatial and temporal control of mitotic spindle assembly by CDK1–cyclin B1 which serves by the release of SAFs such as TPX2 and KIFC1 from inhibitory complexes that hampers spindle formation.

We listed the name of genes for three 50-dimensional system in Table 3.5 and represented the accuracy measures in Table 3.6.

Table 3.5: List of genes for each 50-dimensional network.

Name of Dataset	**List of Genes**
Data 13	N4BP2L1, CHN2, WNT5A, BICC1, STX2, CNR1, PCBP2, CCNE1, TRA2A, PCSK5, PARVA, MUC1, CFH, IRAK3, HOXA5, SST, TTBK2, PTGER3, RUFY3, LYZ, DDX17, RYR2, XIST, CYP3A5, CP, SELENBP1, HSPA12A, CLK1, CCNB1, SLC46A3, DIXDC1, SBSPON, N4BP2L2, ADAMTS3, SLITRK5, FEZ2, DST, CD163, DPY19L2P2, TRAPPC10, DIRAS3, FCF1, VGLL1, FAM69A, MST1, CD24, ARHGEF10, GATM, RGS1, APOA1
Data 14	LGALS2, EZR, CD24, SERINC1, C1S, UBE2I, ADH5, GABARAPL1, LGALS8, ID1, TGFBR2, IFI16, AQP1, HNRNPDL, RGS5, SH3GLB1, HBB, PLIN2, PLPP1, GPM6B, CENPF, MEF2C, FERMT2, PEG3, SLC39A8, TFPI2, CDC42EP3, PLAGL1, DCN, MAF, ANXA3, HIST1H1C, KIF2C, SPON1, SYNE1, GPM6A, CCDC28A, RBPMS, RARRES2, HSDL2, NDN, PAX8, RGL1, NSG1, GNAI1, ADH1B, BUB1, NKX3-1, PDLIM3, BUB1
Data 15	N4BP2L1, CHN2, WNT5A, BICC1, STX2, CNR1, PCBP2, CCNE1, TRA2A, PCSK5, PARVA, MUC1, CFH, IRAK3, HOXA5, SST, TTBK2, PTGER3, RUFY3, LYZ, DDX17, RYR2, XIST, CYP3A5, CP, SELENBP1, HSPA12A, CLK1, CCNB1, SLC46A3, DIXDC1, SBSPON, N4BP2L2, ADAMTS3, SLITRK5, FEZ2, DST, CD163, DPY19L2P2, TRAPPC10, DIRAS3, FCF1, VGLL1, FAM69A, MST1, CD24, ARHGEF10, GATM, RGS1, APOA1

Table 3.6: Results of accuracy measures for 50-dimensional datasets via GGM.

Accuracy Measures	**Data 13**	**Data 14**	**Data 15**
Sensitivity	0	0	0
Specificity	1	1	0.913

On the other hand, the artifical neural network (ANN) method as a non parametric approach alternative to GGM was also conducted for all datasets. Hereby, ANN was used under a univariate approach which means one single gene was defined in terms of other genes in the network. In the calculation, firstly, we calculated the range for all genes by taking the difference between minimum and maximum values of their observations. Then, each observation was divided by this range value to have a scaled observation. To be able to reach a reliable β (regression coefficient/weight), we divided each value to the maximum coefficient value in the network. If the fraction was less than 0.5, we claimed the correlation did not exist(so "0" value was assigned),If that fraction was above 0.5, we took it as 1 and declared the pair to be correlated. This had to be done in order to have the skill to compare with STRING results or to be able with to discuss performances of GGM and ANN. This was the way which we applied for the binary classification as a final goal. Finally, the test set was assigned 90 percent of the entire data, whereas, 10 percent was taken as the training set. While we implemeted ANN to the 6th dataset having 10 genes, we obtained the following estimated adjacency matrix in Table 3.1. From the true structure of this system shown in Figure 3.8, it is seen that STRING indicates no edges, whereas, ANN estimated certain links. Such low accuracy can be caused by different selections of the percentage for train and test sets or law number of observations. Whereas, since the result is based on the application in a single dataset, it cannot be generalized without a comprehensive analyses of ANN under distinct dimensional protein interaction systems.

Table 3.7: Estimated binary network structure for Data 6 via ANN.

	CTCFL	BNCI	KLHDC1	PROM2	SIGLEC11	LRRN4	CBR4	CBS	NT5E	BICD1
CTCFL	0	0	0	0	0	0	0	0	0	0
BNCI		0	0	0	0	0		0	1	0
KLHDC1			0	1	0	1	0	0	0	0
PROM2				0	0	0	0	1	0	0
SIGLEC11					0	0	0	0	0	0
LRRN4						0	1	0	0	0
CBR4							0	0	0	0
CBS								0	0	0
NT5E									0	0
BICD1										0

We listed the outcomes of the accuracy measure for 10-dimensional system in Table 3.8. Furthermore, the results under 23-dimensional and 50-dimensional systems weere seen in Table 3.9 and Table 3.10, respectively.

Table 3.8: Results of accuracy measures for 10-dimensional datasets via ANN.

Accuracy Measures	**Data 1**	**Data 2**	**Data 3**	**Data 4**	**Data 5**	**Data 6**
Sensitivity	NA	NA	NA	NA	NA	0
Specificity	0.91	0.91	0.91	0.91	0.91	1

Table 3.9: Results of accuracy measures for 23-dimensional datasets via ANN.

Accuracy Measures	**Data 7**	**Data 8**	**Data 9**	**Data 10**	**Data 11**	**Data 12**
Sensitivity	0	0	0	0	0	0
Specificity	0.83	0.83	0.83	0.83	0.83	0.90

Table 3.10: Results of accuracy measures for 50-dimensional datasets via ANN.

Accuracy Measures	**Data 13**	**Data 14**	**Data 15**
Sensitivity	0	0	0
Specificity	1	1	1

The results indicate that ANN is also a promising approach for the estimation of biological networks. Its performance improved when the dimension of the system increased and it is not highly affected by the number of observations per gene.

In conclusion, we observed that both GGM and ANN can be used for the construction of biological networks as the selected accuracy measures are close to each other.

CHAPTER 4

CONCLUSION

Many researches use gene co-expression networks to deal with large gene expression datasets in order to comprehend complex biological processes . Modelling co-expression with network models is useful for giving an idea of the co-expression relationships between genes and enables a set of genes to be examined with specific network tools.

In this study, our aim was to construct networks of differentially expressed genes in ovarian cancer using two different methods : Gaussain graphical model (GGM) and artificial neural network (ANN) modelling.

After giving background information about the progress of cancer, ovarian cancer and computational tools which we applied, the parametric approach /method (GGM) was explained with basic information, how and why we used this method, the reasons behind choosing precision matrices for comparison with true network and what was the way to overcome the main problems such as dimensionality. After making necessary comparisons with STRING output, it was seen that GGM to some extent managed to show similarities for gene pairs and their interactions and the accuracy measures support this outcome. On the other hand, despite the fact that, in the literature, there is no solid evidence for "explained interaction" among any gene pair impacting ovarian cancer that we suggested in GGM networks, some of them were found to be co-expressed in some cancer types, indeed even some of them were known to be differentially expressed in ovarian cancer cases though not co expressed.

The ANN modeling, on the other side, found some interactions among genes which are not confirmed by STRING. This may be due to small sample size (53

observations) as neural networks are generally utilized for big data including thousands of inputs. In the ANN algorithm we used it as the alternate of GGM and converted the original model construction to the regression model of GGM.

As an extention of this study, we can evaluate the performance of ANN in different dimensional systems and assess comparative studies based on other accuracy measures such as F-measure and Matthew's correlation coefficient which can control both the presence and non-presence of edges simultaneously.

REFERENCES

Baratloo, A., Hosseini, M., Negida, A., and El Ashal, G. 2015. Part 1: Simple definition and calculation of accuracy, sensitivity and specificity. Emergency (Tehran, Iran), 3 (2), 48–49.

Bishop, CM. 1995. Neural Networks for Pattern Recognıtıon. Oxford University Press.

Chial, H. 2008. Proto-oncogenes to oncogenes to cancer. Nature Education, 1 (1), 33.

Friedman, J., Hastie, T., and Tibshirani, R. 2007. Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9 (3), 432-441.

Friedman, J., Hastie, T., and Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statical Software, 33 (1), 1-22.

Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, and Speed, TP. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 4 (2), 249-264.

Jönsson, JM. 2015. Intrinsic subtypes and prognostic implications in epithelial ovarian cancer. Division of Oncology and Pathology, Lund University.

Kondrashova, O., Topp, M., Nesic, K., Lieschke, E., Ho, G. Y., Harrell, M. I., and Scott, C. L. 2018. Methylation of all BRCA1 copies predicts response to the PARP inhibitor rucaparib in ovarian carcinoma. Nature Communications, 9 (1), 3970.

Kosko, B. 1992. Neural Networks and Fuzzy systems. Prentice Hall, 1992.

Liu, Q., Zhang, H., Jiang, X., Qian, C., Liu, Z., and Luo, D. 2017. Factors involved in cancer metastasis: a better understanding to "seed and soil" ypothesis. Molecular Cancer, 16 (1), 176.

Mok, SC, Bonome, T, Vathipadiekal, V, Bell, A. et al. 2009. A gene signature predictive for outcome in advanced ovarian cancer identifies a survival factor: microfibril-associated glycoprotein 2. Cancer Cell, 16 (6), 521-32.

Sonoda, K. 2016. Molecular biology of gynecological cancer. Oncology Letters, 11(1), 16–22.

Szklarczyk, D., Gable, A., and Junge, A. 2018. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research, 47 (D1), 607-6013.

Toh, H. and Horimoto, K. 2002. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. Bioinformatics, 18, 287-297.

Vijg, J. 2014. Somatic mutations, genome mosaicism, cancer and aging. Current Opinion in Genetics and Development, 26, 141–149.

Zhang, Y., Cao, L., Nguyen, D., and Lu, H. 2016. TP53 mutations in epithelial ovarian cancer. Translational Cancer Research, 5 (6), 650–663.

Zhao, JH., and Tan, Q. 2006. Integrated analysis of genetic data with R. Human Genomics, 2 (4), 258–265.