DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

EVALUATION OF EFFECTS OF THE DRUG TREATMENTS ON BIOLOGICAL NETWORKS

by

Seçkin ERCE

December, 2017

İZMİR

EVALUATION OF EFFECTS OF THE DRUG TREATMENTS ON BIOLOGICAL NETWORKS

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master in Computer Engineering

by
Seçkin ERCE

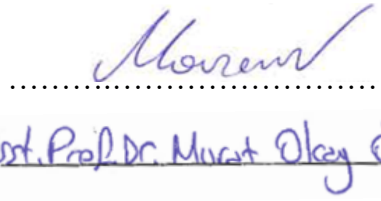
December, 2017
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

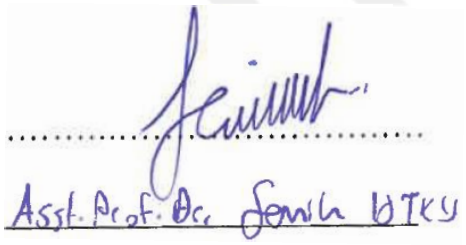
We have read the thesis entitled **"EVALUATION OF EFFECTS OF THE DRUG TREATMENTS ON BIOLOGICAL NETWORKS"** completed by **SEÇKİN ERCE** under supervision of **ASST. PROF. DR. ZERRİN IŞIK** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Zerrin IŞIK

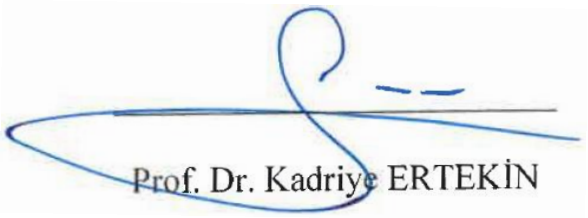
Supervisor



Jury Member



Jury Member



Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENT

I would like to thank my thesis advisor Asst. Prof. Dr. Zerrin Işık from Computer Engineering Department at Dokuz Eylül University. She consistently allowed this paper to be our work, but steered me in the right the direction whenever she thought I needed it.

Seçkin ERCE

EVALUATION OF EFFECTS OF THE DRUG TREATMENTS ON BIOLOGICAL NETWORKS

ABSTRACT

Developing a computational method can assist to figure out the expected results of a drug treatment and observe cell behaviors before doing the wet-lab experiments. This thesis aims to calculate the effects of a drug treatment with bioinformatics methods to better understand the cellular response of a patient even before s/he undertakes a drug treatment for lymphoma. This computational method consists of fundamental scientific domains such as biology, computer science, mathematics, and engineering. This calculation can be performed over a variety of biological networks to numerically interpret the changes of a drug treatment which are represented by gene expression data. The affected biological networks are constructed by integrating the KEGG signaling networks and gene expression data of 14 separate drugs treated on lymphoma cancer cells. Our studies on signaling networks showed that the highest influenced proteins are not drug targets; some other proteins, which are in the periphery parts of the networks, have changed importantly because of the drug treatments. Additionally, statistically significant proteins are generally located at the center of the signaling networks and highly connected to other proteins in the subsequent levels. We confirmed some proteins, which were mostly influenced by the drug treatment and having cancer related cellular activities based on a literature validation. In addition, similarities between drugs were observed according to statistically significant intersection of proteins, and these results were confirmed by similar studies.

Keywords: KEGG signaling pathways, cellular response, drug treatment, gene expression data, lymphoma cells, bioinformatics

İLAÇ TEDAVİLERİNİN BİYOLOJİK AĞLARA ETKİSİNİN DEĞERLENDİRİLMESİ

ÖZ

Hesaplamalı bir yöntemin geliştirilmesi ilaç tedavisinden beklenecek sonuçları bulmaya ve laboratuvar deneyleri öncesinde hücrenin davranışlarını gözlemlemeye yardımcı olabilir. Bu tez bir hastanın lenfoma tedavisi için ilaç tedavisi almadan önce hastanın hücresel tepkisini daha iyi anlamak için biyoenformatik yöntemlerle ilaç tedavisinin etkilerini hesaplamayı amaçlar. Bu hesaplamalı biyoenformatik yöntemler, biyoloji, bilgisayar bilimleri, matematik ve mühendislik gibi temel bilimsel alanlardan oluşur. Bu hesaplama, gen ekspresyon verileri ile temsil edilen bir ilaç tedavisinin etkilerini nicel olarak değerlendirmek için çeşitli biyolojik ağlar üzerinde gerçekleştirilmiştir. Bu etkilenen biyolojik ağlar, KEGG sinyal ağları ve lenfoma hücrelerinde uygulanan 14 farklı ilacın gen ekspresyon verileri birleştirilerek oluşturulmuştur. Bu sinyalizasyon ağlarının analizi sonuçlarımız, en çok etkilenen proteinlerin verilen ilaçların doğrudan protein hedefleri değil, ağın dış kısımlarındaki diğer proteinler olduğunu göstermiştir. Ek olarak, istatistiksel açıdan önemli proteinler genel olarak sinyal ağlarının merkezinde bulunmuş ve bu proteinlerin birbirlerine fazla sayıda bağ ile bağlı olduğu görülmüştür. Yaygın olarak birkaç ilacın tedavisi sonrası etkilenen ve istatistiksel olarak anlamlı olan bazı proteinlerin, kanserle ilişkili hücresel aktiviteleri olduğu literatürde de doğrulanmıştır. Aynı zamanda ilaçlar arası benzerlikler, istatistiksel açıdan önemli proteinlerin ortaklığına göre gözlemlenmiştir ve çıkan sonuçlar yapılan benzer çalışmalar ile doğrulanmıştır.

Anahtar Kelimeler: KEGG sinyal yolakları, hücre davranışları, ilaç tedavisi, gen ekspresyon verileri, lenfoma hücreleri, biyoenformatik

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

1.1 Motivation

Microarray experiments freely can be accessible for whole genomes of several organisms during the last years. Such datasets includes in the cell about the molecular level attitude of genes under different situations like drug treatment. The analysis of microarray experiments creates gene expression profiles. Classic microarray experiments detect the drug targets that are supposed to be a response from the cell in the gene level (Kerr & Churchill, 2007).But the drug targets cannot give an exactly view of the cellular response. We need to find new algorithms to solve that problem (Isik, Ersahin, Atalay, Aykanat & Cetin, 2012).

Biological pathways are collections units of proteins that cooperatively do a defined metabolic duty. Pathway-based analysis is very new view about understanding huge amount of gene expressions in the signaling and metabolic levels data. Biological Pathway represent the attitude of group of genes as a reply to an external signal (Pas, Hemert, Hulsegge, Rebel & Smits,2008).Graphical topology like tree networks of cell signaling networks and gene expression profiles may help to understand drug treatment effects in the cell attitude in the metabolic level (Zhang, Gao, Liu, Zhao & Che, 2009).

Microarrays are the summary for searching the genetic structure of each patient. Microarray experiments are very complex and must be accompanied by data analysis components. This technology offers the opportunity to obtain the exact state of gene expression and to detect genes and pathways that, are affected by the drug treatment (Isik, Ersahin, Atalay, Aykanat & Cetin, 2012). Figure 1.1 shows the drug-protein interaction in a signaling pathway. Clustering is one of the most important step of the microarray analysis (Shapiro & Tamayo, 2003). Clustering is done for organizing the samples into many clusters such that cluster samples with huge similarity belong to same cluster (Bellaachia, Portnoy, Chen & Elkahloun, 2002). Clustering of gene

profiles is done for disclosing the effect of drug treatment on genes. These proteins can act as target for researchers to discover drugs that can be very useful in drug treatment of the disease (Li & Ong, 2004). The highest influenced proteins are not drug targets; on the other proteins which is in the faraway parts of the networks have changed importantly because of the drug treatments.

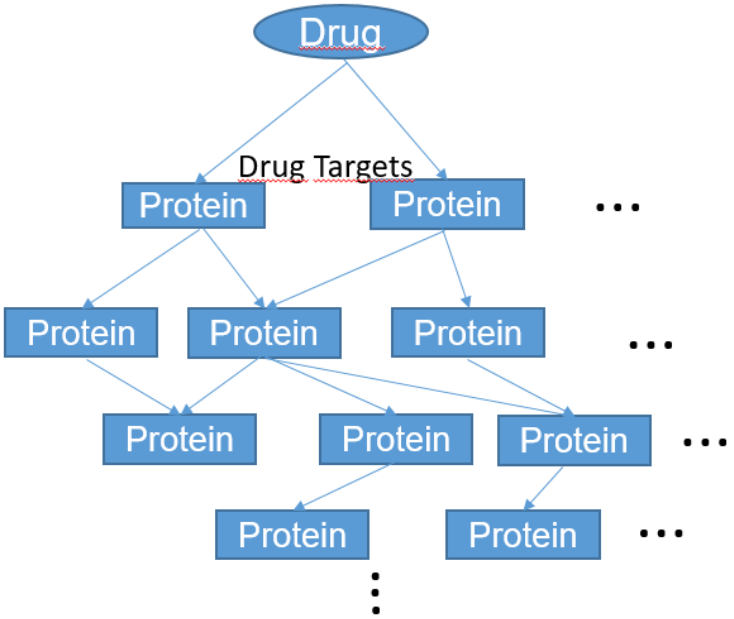


Figure 1.1 Drug-Protein interaction in a signaling pathway

1.2 Problem Definition

The objective of this thesis is to prediction of drug treatment impacts on the signaling pathways. The specific questions addressed in this study are summarized in the following.

- How to use the free databases?
- Which method will be used for getting initial scores for proteins?
- How drug targets are identified?
- How the cycles were eliminated in pathways?
- Is the analysis of only the drug targets enough for providing a completely comprehension of the molecular response of the cell?

- How does a score flow algorithm visualize response of the cell the effects of drugs on the biological network?
- How are the correctness of the results validated?
- What is the relation between the level and the effected proteins?
- What is the similarity rate used for detecting the similarity percentage of the drugs?

1.3 Organization of the Thesis

Chapter 2 covers a literature review. Chapter 3 explains the materials and methods which were used in the thesis. Chapter 4 describes results of the thesis. Chapter 5 is the conclusion part and also offers recommendations for future improvements about this work.

CHAPTER TWO
LITERATURE REVIEW

2.1 From DNA to Protein

DNA carries heritable information in the cell. Genes code information of one or more proteins. An organism as a whole, we need to understand the role and function of DNA in every cell of an organism (Alberts, Johnson, Lewis, Raff, Roberts, & Walter, 2002).

Proteins are the most important molecules of life, joining in fundamentally every chemical and biological molecule and all activity in the life. They are creating materials for the cells, seeming in the structures inside the cell and within the cell membrane. They transport oxygen, they create tissue, they transfer DNA for the children and they make all the job in any living being (Cooper, 2000).

2.1.1 Discovery and Structure of DNA

DNA (Deoxyribonucleic Acid) discovered by Gregor Mendel in 1865. Mendel noticed DNA with breeding experiments in peas, that the several phenotypes of the peas were inherited according to certain laws. Then we call that "Mendelian Laws" (Mendel, 1865). Friedrich Miescher was the first person who isolated DNA in 1869, but he did not understand the importance of his discovery. Wilhelm Johannsen expressed gene to define as a unit of heredity in 1909. From 1949 to 1953 the structure of DNA was understood.

DNA has four bases, associated by a sugar-phosphate backbone. Nucleotides are these bases. Nucleotides establish two component linearly composed strands, compose a double spiral. Two nucleotides respectively compose hydrogen bonds, which balance the double helix. The four nucleotides are Adenine, Thymine, Guanine and Cytosine. A and T can connect to each other by combining two hydrogen bonds. That's why, A and T are called to be supplementary. G and C are too supplementary: they combine

with three hydrogen. This supplementary strand is said complementary DNA (cDNA). When two proper cDNA strands bind to each other they install the double helix, referred above. The connecting of two complementary cDNA strands is named hybridization (Pray, 2008). DNA is shown with double helix form. Figure 2.1 shows the double helix form with its backbone and nucleotides.

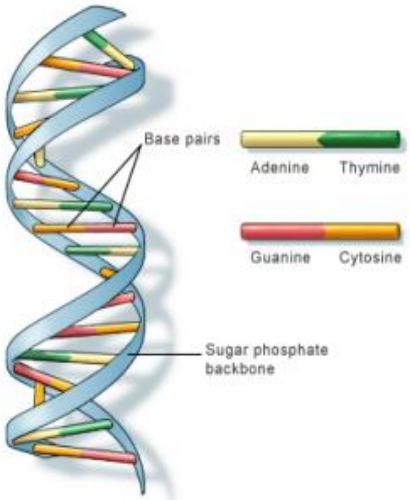


Figure 2.1 Double helix structure of the DNA

In 1957, Francis Crick discovered the flow of information from DNA to RNA to protein while Frederick Sanger, Allan Maxam and Walter Gilbert present the first methods on how to sequence DNA in 1977. This was very important for understanding the genetics of a living being. The next years many different function, structure and feature of DNA were understood. The first microarray technics were advanced in the 1980's. Scientists reported that "Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome." in 2001 (Lander et al. , 2001).

Since 1865 many more discoveries about DNA found, but to explain all of them would go outside the scope of this thesis.

2.1.2 Functional View on DNA : Genes and Genes Expression

Genes are a part of the DNA. A gene of a living being is a part of one of the living being's chromosomes. It has the hereditary information for one or more specific proteins of the organism (Cooper, 2000).

The flow of information from DNA to RNA to proteins is one of the fundamental principles of molecular biology. It is called as the "central dogma". The central dogma of molecular biology explains the two-step steps, transcription and translation. Every protein of the organism is coded in DNA. Transcription is the synthesis of an mRNA copy of a segment of DNA. RNA is synthesized by the enzyme RNA polymerase. In the translation stage, the mRNA is "decoded" to create a protein that has a specific series of amino acids by ribosomes. The flow of information from DNA to the proteins is shown in Figure 2.2. The abundance of translated mRNA is called the expression level of the gene (Zien, Schoelkopf, Tsuda, & Vert, 2004).

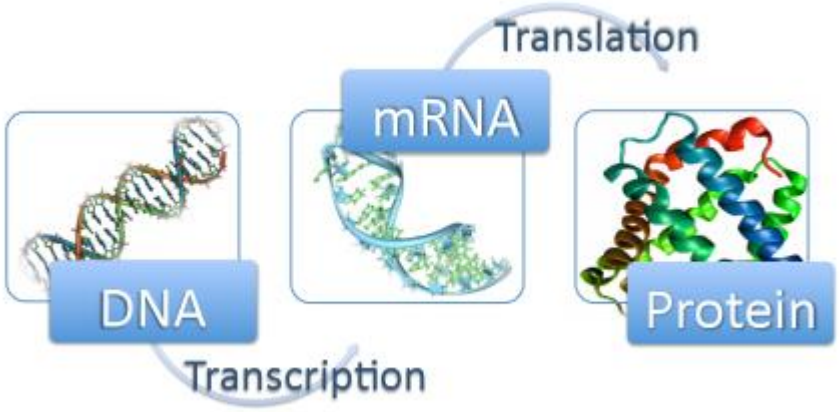


Figure 2.2 The translation and transcription of a gene into a protein. First, the DNA is transcribed into the mRNA .Second, the mRNA is translated into the corresponding protein

Humans contain about 30000 genes. At any given time, for different cell groups, some of these genes are active, others are inactive. Researchers can find an answer to this question for a cell sample or tissue by examining gene expression profiling, a microarray analysis technique (Ahnert, Fink & Zinovyev, 2008).

The genome of a living being consists of the set of all genes. These genes have a lot of biological functions. Understanding the function of a gene means that influences the gene reacts on and which might have effects on other genes. Such discoveries can be made by detecting differential expression of genes between certain conditions (Baldi & Hatfield, 2002).

Genes encode proteins and proteins dictate cell function. Therefore, the thousands of genes expressed in a particular cell determine what that cell can do. Moreover, each step in the flow of information from DNA to RNA to protein provides the cell with a potential control point for self-regulating its functions by adjusting the amount and type of proteins it manufactures. At any given time, the amount of a particular protein in a cell reflects the balance between that protein's synthetic and degradative biochemical pathways (Berg, Tymoczko & Stryer, 2002). On the synthetic side of this balance, recall that protein production starts at transcription (DNA to RNA) and continues with translation (RNA to protein). Thus, control of these processes plays a critical role in determining what proteins are present in a cell and in what amounts. In addition, the way in which a cell processes its RNA transcripts and newly made proteins also greatly influences protein levels (Alberts, Johnson, Lewis, Raff, Roberts, & Walter, 2002; Cooper, 2000 ; Wong, 2016)

2.2 Microarray Data Analysis

General steps of microarray analysis is shown in the Figure 2.3. First step is isolation that means extraction of mRNA from cells. Then labelling with color (generally fluorescent) and hybridization are applied. Hybridization and washes are done under high stringency conditions for abstaining from possibility of cross-hybridization between similar genes. The next stage is detection to generate array-images using fluorescent microarray scanner. The basis of gene expression levels is that the quantification of fluorescence measured at each sequence-specific location is mean to the quantity of mRNA. These measurements are helpful to compare the gene expression with other genes in different conditions (healthy vs. and disease) (Pulverer, Noehammer, Vierlinger & Weinhaeusel , 2012) .

Data extraction stage means the representation and extraction of data from images acquired from microarray experiments. Bioinformatics analysis is the last stage of the microarray analysis. That stage works on data normalization and statistical data analysis (Jain et al.,2002).

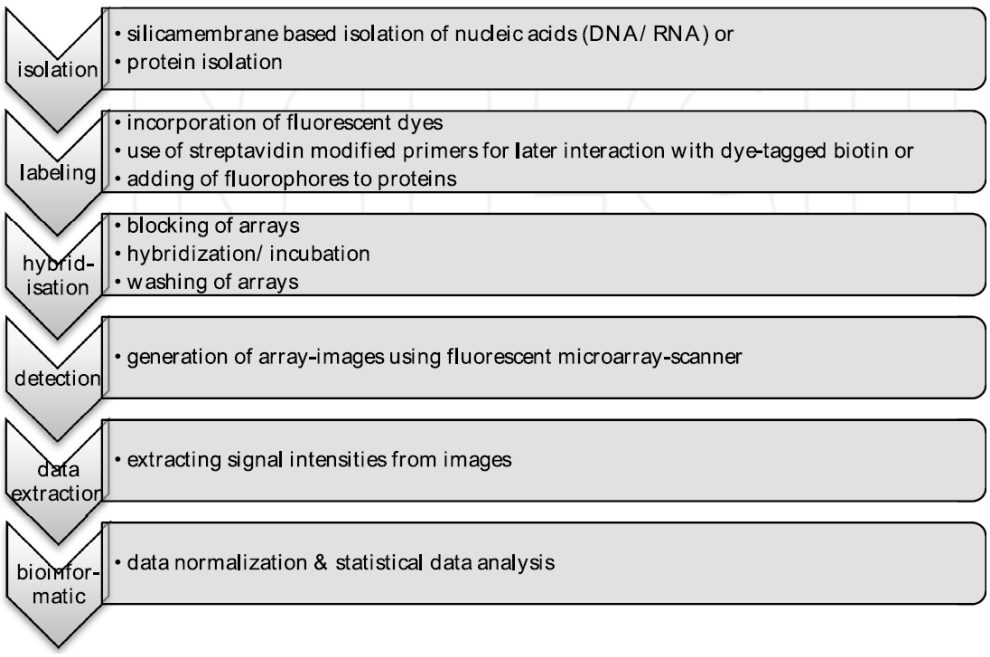


Figure 2.3 General Principles of Microarray Data Analysis

2.2.1 Data Extraction

Data extraction stage is shown in the Figure 2.4. Scanners are imaging microarrays. This is called quantitation, and is necessary for all microarray analyses. These image data must be confirmed with background correction. Background corrections are used for removing non-specific data that arises from non-specific hybridization, or coatings or other materials on the microarray analysis (Rickman, Herbert & Aggerbeck, 2003; Brown, Goodwin & Sorger, 2001; Wang, Ghosh & Guo, 2001).The mis-match adjustment can be used for removing non-specific hybridization like a secondary normalization process (Liu et al., 2002 ; Irizarry et al., 2003).

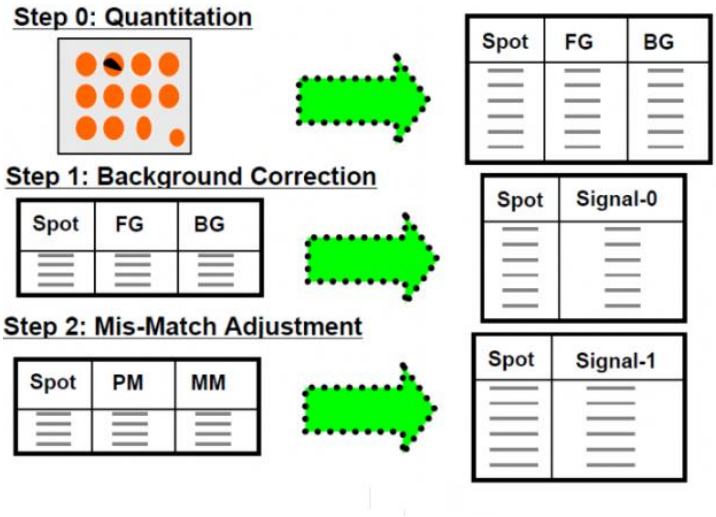


Figure 2.4 Steps of Data Extraction

2.2.2 Bioinformatics Analysis

General approach for the bioinformatics process of microarray data is shown in the Figure 2.5. Normalization is the step of removing "non-biological variability" from a dataset. Summarization converts the signal from all sequences into a single number (Yang et al., 2002).

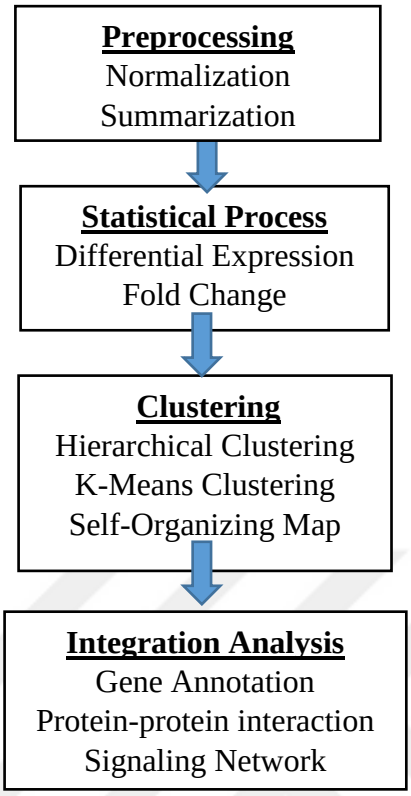


Figure 2.5 General approach for the bioinformatics process of microarray data

Standard statistical processes are usually used for every gene in the microarray. Detecting genes which change in response to a drug, one would use a t-test to search for differential expression between drug-treated and non-drug treated samples. The t-test would be done one by one for every gene on the microarray. Fold change is the most used method for differential expression (Chen-An, Yi-Ju & James, 2003).

Fold change is a measure describing how much expression level changes from an initial to a final value / condition. In the field of bioinformatics, fold changes are defined directly in terms of ratios. If the initial value is A and the final value is B, the fold change is defined as B/A. Note that this is different to the definition described above. In other words, a change from 30 to 60 is defined as a fold-change of 2. This is also referred to as a "2-fold increase". Similarly, a change from 30 to 15 is referred to as a "2-fold decrease". Log ratios are often used for analysis and visualization of fold changes. The $\log_2$ (log with base 2) is most commonly used conversion (Pingzhao, Celia & Joseph, 2009; Tusher, Tibshirani, Chu, 2001). For example, on a plot axis

showing log_2-fold-changes, an 8-fold increase will be displayed on the axis as 3 (since 2^3 = 8) (Mariani et al., 2003).

The p-value is the probability value that a gene's expression value is different between the two groups due to randomized data. A p-value of 0.05 signifies a 5% probability that the gene's mean expression value in one condition is different than the mean in the other condition by chance alone (Tusher, Tibshirani, Chu, 2001).

If error of gene x is less than %5 up to random-input running then this gene is statistically significant. We can test this case for every genes in the sample so we can detect the genes which are statistically significant (Storey & Tibshirani, 2003).

2.2.2.1 Clustering

Clustering is the step of detecting the patterns in a dataset. That means "natural trends" of the data. It does not include any knowledge about biological hypothesis, samples or gene annotation (Smolkin & Debashis, 2003). Clustering methods are used for identifying patterns of gene expressions. It helps for understanding of the relations among gene expressions. These patterns can be found with the similarity or distance metrics among the gene expression profiles (Jelili et al.,2016).

Euclidean distance and Pearson's correlation are the most used distance functions. Clustering method will be chosen up to the distance function. K-means clustering, Self-Organized Maps (SOM) and hierarchical clustering are the most used clustering techniques.

2.2.2.1.1 Hierarchical Clustering. Hierarchical clustering methods perform hierarchical decomposition of units in the data set using the distance values of the units of the data set to each other. During hierarchical decomposition, a tree diagram known as a dendrogram is used. The tree diagram provides visualization of clusters obtained by hierarchical clustering. The number of clusters is visually decided (Sibson, 1973; Defays, 1977).

Agglomerative and divisive approaches are fairly the most used techniques. Agglomerative is a "bottom up" approach. Each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy. Divisive is a "top down" approach. All observations start in one cluster and splits are performed recursively as one moves down the hierarchy (Rokach & Oded, 2005). The two types of hierarchical clustering with dendrogram is shown in the Figure 2.6.

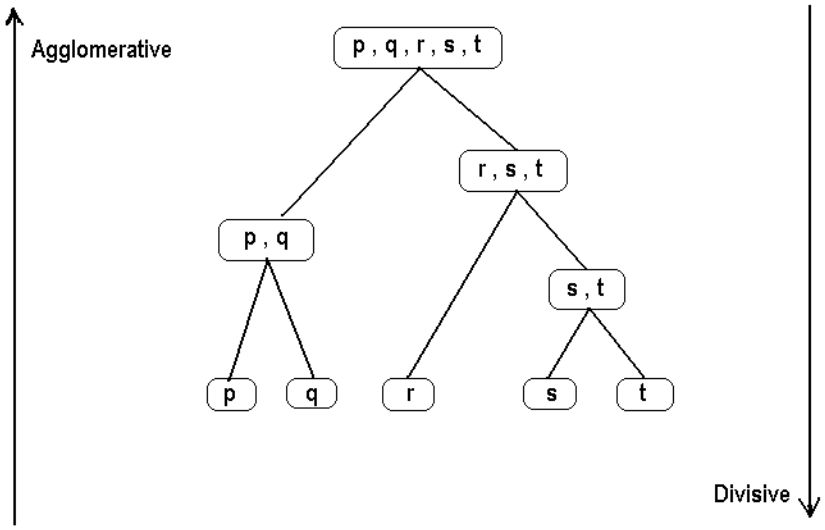


Figure 2.6 Hierarchical Clustering Dendrogram with two types

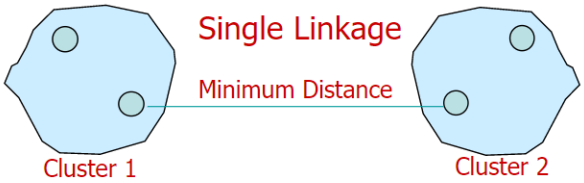
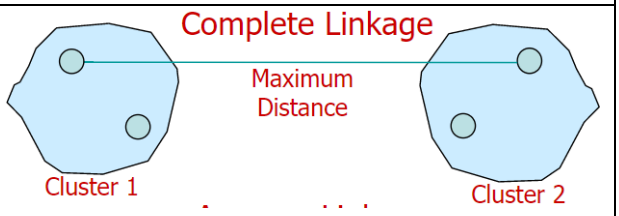
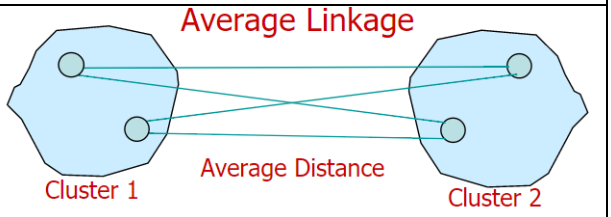
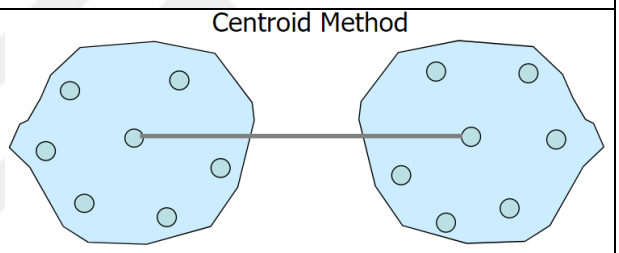
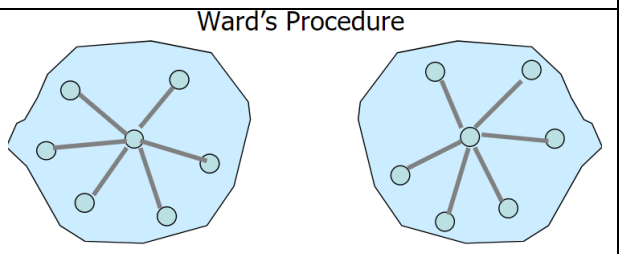
Hierarchical clustering mainly includes two phases. At first a distance matrix, which has all the pairwise distances between the genes, is calculated for detecting dissimilarity estimates via distance metrics (Meelis & Jaak, 2008). Most used metrics for measuring distances are shown in the Table 2.1. So, we have results about the number of distance measures available and their influence in the clustering algorithm results. After in every step, a new distance matrix between the newly formed clusters and the other clusters are calculated again.

Table 2.1 Measuring Distances Formulas

Methods for measuring distances	**Formula**
Euclidean distance	$d_{euc}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
Manhattan distance	$d_{man}(x,y) = \sum_{i=1}^{n} \lvert(x_i - y_i)\rvert$
Pearson correlation distance	$d_{cor}(x,y) = 1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$
Eisen cosine correlation distance	$d_{eisen}(x,y) = 1 - \frac{\left\lvert\sum_{i=1}^{n} x_i y_i\right\rvert}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$
Spearman correlation distance	$d_{spear}(x,y) = 1 - \frac{\sum_{i=1}^{n}(x'_i - \bar{x'})(y'_i - \bar{y'})}{\sqrt{\sum_{i=1}^{n}(x'_i - \bar{x'})^2 \sum_{i=1}^{n}(y'_i - \bar{y'})^2}}$
Kendall correlation distance	$d_{kend}(x,y) = 1 - \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$

The linkage criterion determines the distance between sets of observations as a function of the pairwise distances between observations (Székely & Rizzo, 2005). Some commonly used linkage criteria are shown in the Table 2.2.

Table 2.2 Linkage Methods with Figures (Slideshare, 2012)

Linkage Methods	**Figures**
Single Linkage	
Complete Linkage	
Average Linkage	
Centroid Linkage	
Ward Linkage	

2.2.2.1.2 K-Means Clustering. The K-means clustering divides a data set consisting of N data objects into K sets which is given as the input parameter. The aim is to maximize the in-cluster similarities (Forgy, 1965).

The K-means is the most commonly used clustering algorithm, since its implementation is easy. Large-scale data can be clustered quickly and efficiently. "K" represents the number of fixed clusters and it is needed before starting the algorithm.

With the recursive partitioning scheme, the K-means algorithm reduces the sum of the distances to the cluster to which each data belongs. The K-means algorithm tries to find the K pieces that are the smallest to make the quadrature error (Sculley, 2010). A K-Means clustering sample is shown in the Figure 2.7.

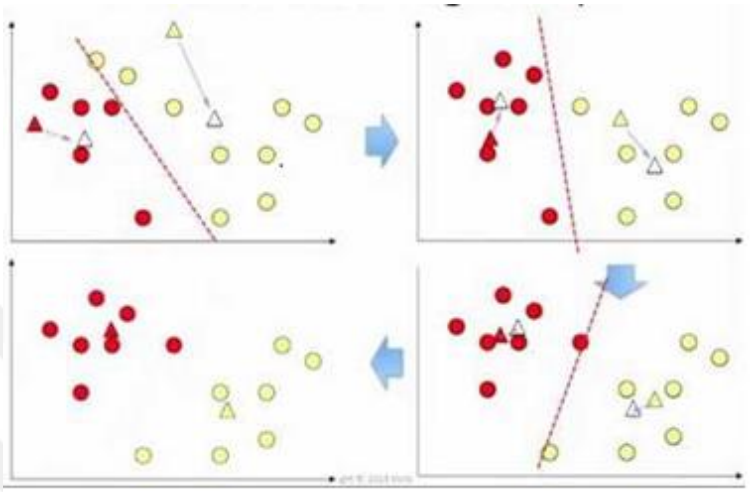


Figure 2.7 K-Means Clustering example

According to the working mechanism of the K-means algorithm, first, K objects are randomly selected to represent the center point or average of each bin. Other remaining objects are included in the closest similar clusters, taking into account the distances of the clusters to their mean values. Next, the average value of each cluster is calculated, new cluster centers are determined, and the center distances of the objects are examined again. The algorithm continues until there is no change in the cluster members (Celebi, Kingravi & Vela, 2012). How the K-means algorithm works is shown in the Figure 2.8.

The algorithm basically consists of 4 steps:

1. Determination of cluster centers
2. Clustering according to distances of data outside the center
3. Determination of new centers according to the clustering (or shifting old centers to new centers)
4. Repeat steps 2 and 3 until the stable state is reached.

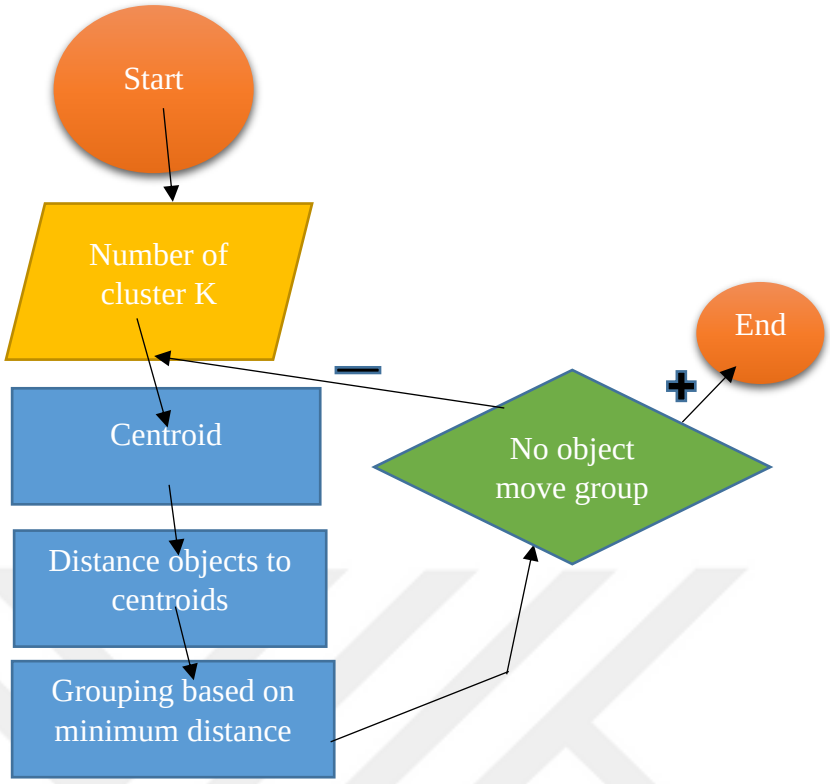


Figure 2.8 How the K-Mean Clustering algorithm works?

2.2.2.1.3 Self-Organizing Map (SOM) Clustering. Self-organizing maps are a special form of artificial neural networks and use unattended training during their training. At first, the system trains itself and competitive learning is used. In the second case of mapping, the network works to correct the incoming new arrivals properly (Haykin, 1999).

Fundamentally, the operations based on the reduction of output to a lesser extent in multi-dimensional inputs. For simplifying the problem, dimension reduction processes are done.

SOM, which may be an example of structurally feed forward networks, behaves similarly to the k-means algorithm for very small amounts of neurons. With the increase of the number of SOM, the difference also arises (Liu & Weisberg, 2011 ; Yin,Huang & Nii 2006). A SOM clustering sample is shown in the Figure 2.9.

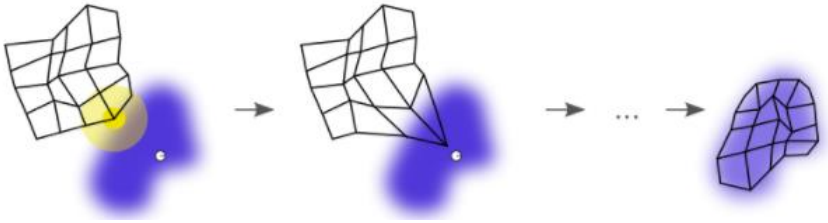


Figure 2.9 SOM (Self-organizing map) clustering example

2.2.2.1.4 Validation of Clustering. Validation of clustering results are very critical. Actually, the cluster validation methods are done for the partitioning that best match the microarray data. So it is a key tool in the interpretation of clustering results. In here, the most used three cluster validation methods were searched (Datta, 2003 ; Xu, Olman & Zu, 2002 ; Quackenbush, 2001).

Connectivity captures the degree to which genes are linked in a cluster by monitoring whether neighboring genes are put into the same cluster (Handl, Knowles & Kell, 2005). The Silhouette Index (SI) reflects the density and separation of clusters (Rousseeuw, 1987). The Jaccard Index (JI) uses the intersection ratio between two gene expressions set (Jaccard, 1912).The Rand Index is used to calculate the accuracy. Then we can measure of agreement between two clustering partitions (Rand, 1971).

Table 2.3 Validation of Clustering Methods

Validation Clustering Methods	**Formula**	**Result**
Connectivity	$Conn(P) = \sum_{i=1}^{m}\sum_{j=1}^{n} Xim_{i(j)}$	The value must be from 0 to Infinity and must be minimized.
The Silhouette Index	$s(P) = \frac{1}{m}\sum_{i=1}^{m}(b_i - a_i) \,/\, \max\{a_i, b_i\}$	The result value vary from -1 to 1 and higher value means better clustering results.
The Jaccard Index	$J(P_1, P_2) = \frac{a}{a+b+c}$	The Jaccard Index score ranges from 0 to 1 and higher value means better clustering results.
The Rand Index	$Rand(P_1, P_2) = \frac{a+b}{m(m-1)/2}$	The Rand Index ranges from 0 to 1, where a higher value indicates a higher accuracy.

P = {C1,C2,...,Ck} portion of Matrix (M)

ai = the average distance of gene I

bi = the minimum of the average distances of gene I

P1 and P2 = gene expression profiles

a = The number of gene expression profile pairs that belong to the same cluster in P1 as well as in P2

b = The number of gene expression profile pairs that belong to the same cluster in P1 but not in P2

c = The number of gene expression profile pairs that belong to the same cluster in P2 but not in P1

m = m is the number of genes

2.2.2.2 Integration Analysis

Integration analysis includes gene and sample annotation. This process also mention the combination of microarray data (Yin, Jianrong, Caro & Yves , 2007). For example protein-protein interaction network or drug target and protein network is

example of this process. Signaling pathway modeling for drug target detection in cancer is also example of this process.

2.2.2.2.1 Signaling Pathways. Cells are aware of what's happening around them, and they can reply in real time to a signal from the other cells and environment. That means cells can send and receive a lot of message via chemical signaling molecules. A cell signaling network or signaling pathway means that a group of proteins work together to control behavior as a response to chemical signals (Bu & Callaway, 2011).

Researchers have been investigating the behavior of signaling networks with new algorithmic ways. New algorithms needs to detect genes which affected by drugs. This can be done with biologically-derived data set which are specifically affected by only applied drug treatment. So drug treatment will be analyzed using gene-level drug data rather than metabolic patient data. Determination of cell signaling behavior is crucial for understanding the physiological response to a specific stimulus or drug treatment. Current approaches for large-scale data analysis do not effectively incorporate critical topological information provided by the signaling network (Isik, Ersahin, Atalay, Aykanat & Cetin, 2012).

Pathway-based analysis is a perspective that has emerged over the last few years to understand and examine gene expression profiling, which is an abundant amount of cellular signaling and metabolic levels. One of the most used signaling pathway is apoptosis mechanism in the cancer researches (Brune, Kenthen & Sandau, 1999 ; Huang et al.,2016).

2.2.2.2.1.1 Apoptosis. The cell death resulting from the activation of the system containing the self-destructing (suicide, suicide) program encoded in the genetic system. For this reason, the term "preprogrammed cell death" is also applied to the case of apoptosis. The most important purpose of the "suicide" program in most cells is to remove the cells that are damaged by impossible repairs from stress. Thus, future complications are avoided.

In the case of apoptosis, cell membrane does not deteriorate. The cell is divided into small pieces that can be phagocytized first. The dead cell, which is divided into small pieces and rapidly phagocytized, does not cause an inflammatory reaction in the surrounding tissues (necrosis). In necrosis, the cell membrane breaks down, the substances released from the cell cause an inflammatory response, and the necrotic cells are melted by the enzymes produced by the phagocytes (Douglas, 2011). The signaling pathway of Apoptosis is shown in the Figure 2.10.

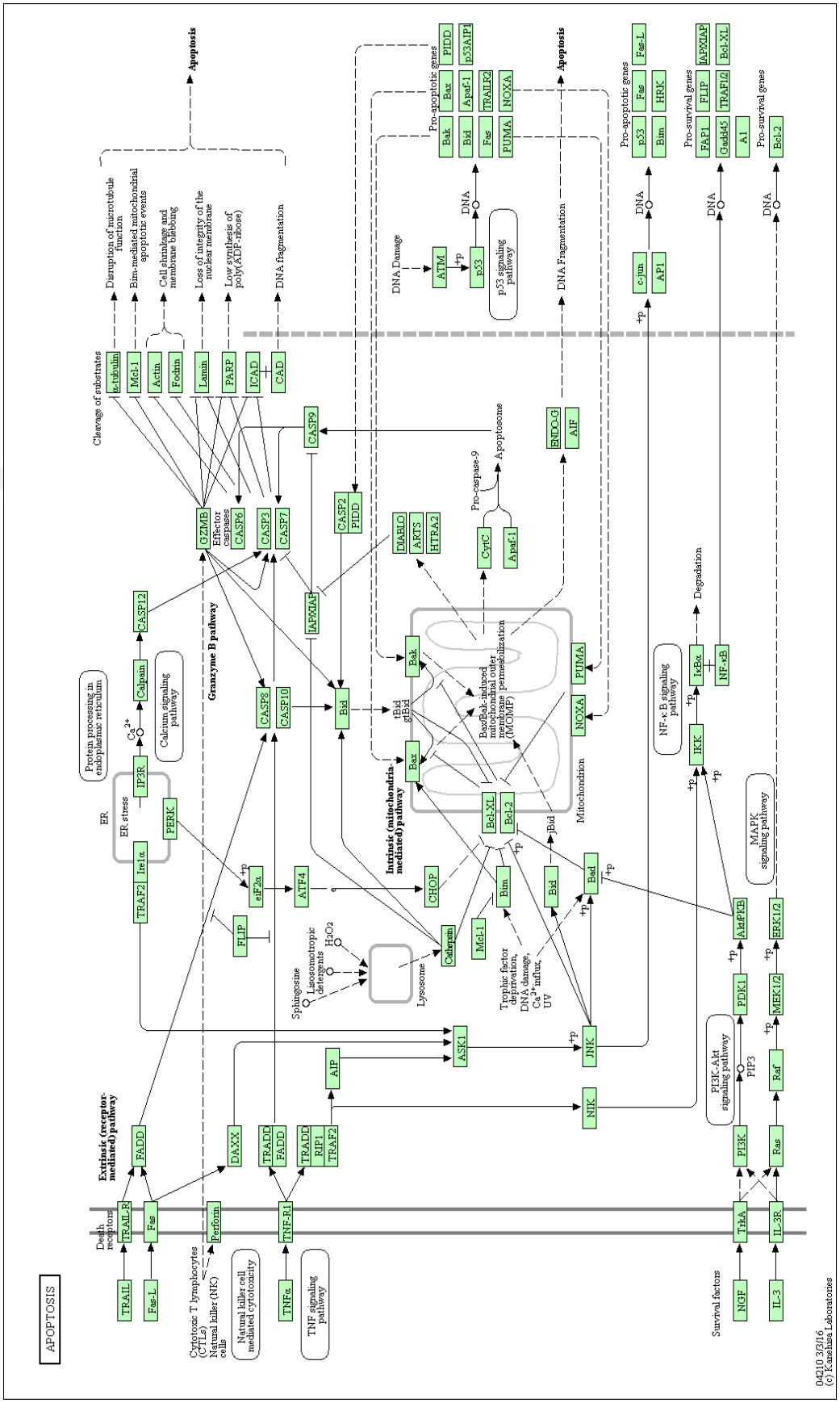


Figure 2.10 Apoptosis Mechanism (Genome,2016)

CHAPTER THREE
MATERIAL AND METHODS

3.1 DATA

3.1.1 Gene Expression Data

There are a lot of different methods to calculate cellular responses genomically. Microarray experiments aim to measure the mRNA levels of genes under certain conditions. After statistically different analyzes of the microarray experiments are performed, the expression levels of the genes are measured and the increase and decrease in the mRNA level compared to the control samples. The microarray experiments were used to observe the effects of 14 different drugs (Table 3.4) on lymphoma cancer cells (Bansal et al., 2014).

For each gene in our gene expression data set we have 3 separate samples for both drug-treated and control samples. Both the drug-treated and the control samples are represented as a single value by calculating the median within the values of 3 individual values. After the calculation, each of the gene symbols converts to the corresponding gene identifiers. At the last stage, the medicines and control samples for each gene are given as inputs to calculate the scores separately. Table 3.1 and Figure 3.1 show simulations on how drug treatments and control sample values are computed.

Table 3.1 Pseudo code for the calculating fold change in gene expression

```
Create Fold Change file both after drug and before drug.
Input : ProteinName stored all protein names with After Values consists AValue1,Avalue2
,Avalue3 and Before Values consists BValue1,BValue2,BValue3
Initiliaziation :
 for each vertex  name∈ ProteinName
            id=GetProteinId(name)
            Before_Drug(id)= Median(BValue1,BValue2,BValue3)
             After_Drug(id)  = Median (AValue1,AValue2,Value3)
Protein_ID_Unique = Protein_ID.Distinct()
for each vertex  Id ∈ Protein_ID_Unique
             Medyan_before = Median (Before_Drug(Id))
             Medyan_After = Median (After_Drug(Id))
Output :
     for each vertex x ∈ V do
          WriteFile(ProteinID(x),ProteinName(x),Medyan_First(x))
          WriteFile(ProteinID(x),ProteinName(x),Medyan_Last(x))
```

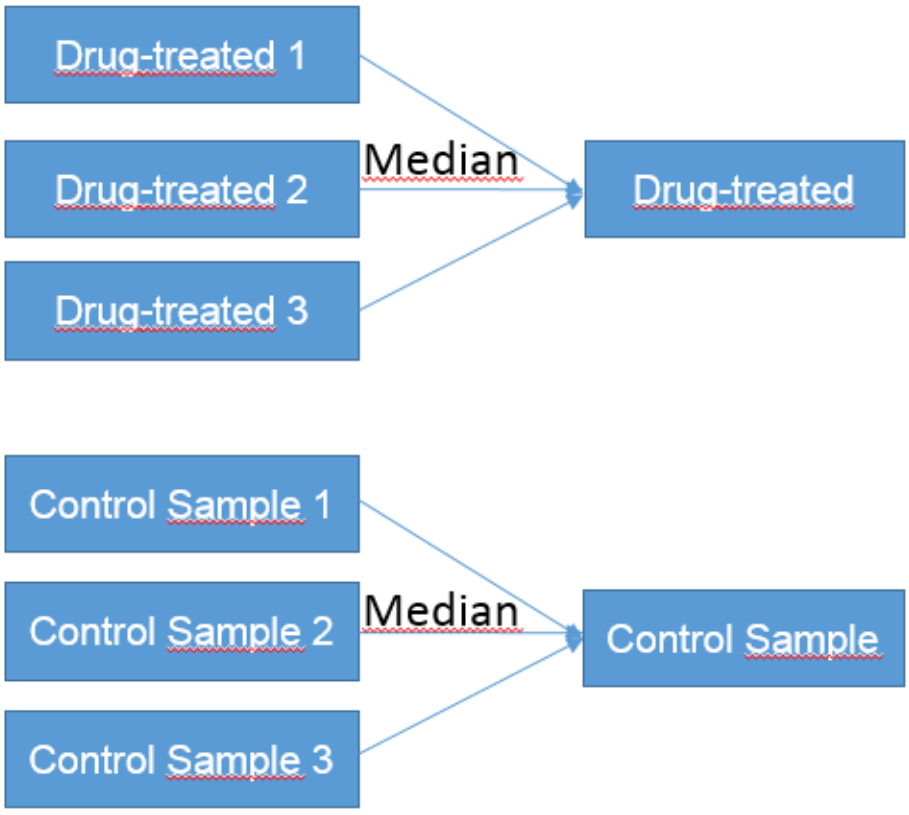


Figure 3.1 Calculating After Drug and Control Sample Value

3.1.2 Pathway Data

Pathway Commons is collection of publicly available pathway data from multiple organisms (Cerami et ai.,2011). Pathway Commons provides a web-based interface that enables biologists to browse and search a comprehensive collection of pathways from multiple sources represented in a common language, a download site that provides integrated bulk sets of pathway information in standard or convenient formats and a web service that software developers can use to conveniently query and access all data. Pathways include biochemical reactions, complex assembly, transport and catalysis events and physical interactions involving drug and proteins. Biochemical reactions, complex assembly, transport and catalysis events and physical interactions represents with edge. Drugs and proteins shows with node. Drugs can be only source node but proteins can be target and source nodes (Cerami et al., 2011). The relation between a source node and target node is shown in the Figure 3.2. All edge relations covered in KEGG pathways are shown in the Table 3.2. The main purpose of our study is to detect changes in genes after drug treatment. For this reason, we select only the "interacts-with" and "consumption-controlled-by" edge types from Pathway Commons.

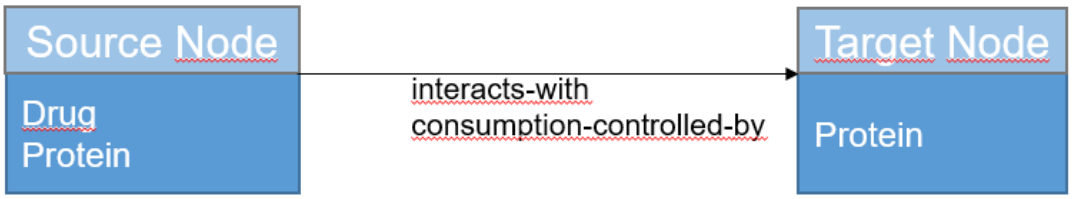


Figure 3.2 Source Node to Target Node

Table 3.2 Types of Binary Relation in KEGG pathways. Names, Descriptions and Inferred Binary Relations are showed in KEGG Pathways.In this project I use the bold rows.They are consumption-controled-by and interacts-with

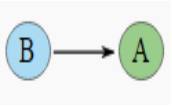
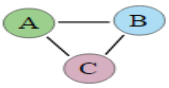
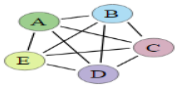
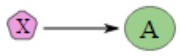
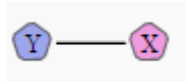
Name	**Description**	**Inferred Binary Relation(s)**
controls-state-change-of	First protein controls a reaction that changes the state of the second protein.	
controls-transport-of	First protein controls a reaction that changes the cellular location of the second protein.	
controls-phosphorylation-of	First protein controls a reaction that changes the phosphorylation status of the second protein.	
controls-expression-of	First protein controls a conversion that changes expression of the second protein.	
catalysis-precedes	First protein controls a reaction whose output molecule is input to another reaction controled by the second protein.	
in-complex-with	Proteins are members of the same complex.	
interacts-with	**Proteins are participants of the same Molecular Interaction.**	
neighbor-of	Proteins are participants or controlers of the same interaction.	
consumption-controled-by	**The small molecule is consumed by a reaction that is controled by a protein**	
controls-production-of	The protein controls a reaction of which the small molecule is an output.	
controls-transport-of-chemical	The protein controls a reaction that changes cellular location of the small molecule.	

Table 3.2 continues

chemical-affects	Small molecules are input to a biochemical reaction.	
reacts-with	Small molecules are input to a biochemical reaction.	
used-to-produce	A reaction consumes a small molecule to produce another small molecule.	

3.1.2.1 Gene Name to Identifier Conversion

Gene name to identifier conversion provides an efficient and reliable mechanism for conversion between identifier domains of interests. All tools in the DAVID Bioinformatics Resources aim to provide functional interpretation of large lists of genes derived from genomic studies (Dennis et al., 2003).

Gene name to identifier conversion was performed by using Pubmed-NCBI Database that contains correct identifiers for all Homo sapiens genes. Synonym is a symbol by which a gene has been alternatively known in the literature or databases, or which groups it into a known gene family. Synonyms are usually recorded along with the approved symbols as part of the gene entry to facilitate database searching. The HGNC Database, Ensembl Database, Entrez Gene Database, GeneCards Database and Online Mendelian Inheritance in MAN (OMIM) database are all contain both approved symbols and synonyms (Kristian, Ruth, Susan, Mathew & Elspeth, 2016). A sample for gene name to identifier conversion is shown in Table 3.3.

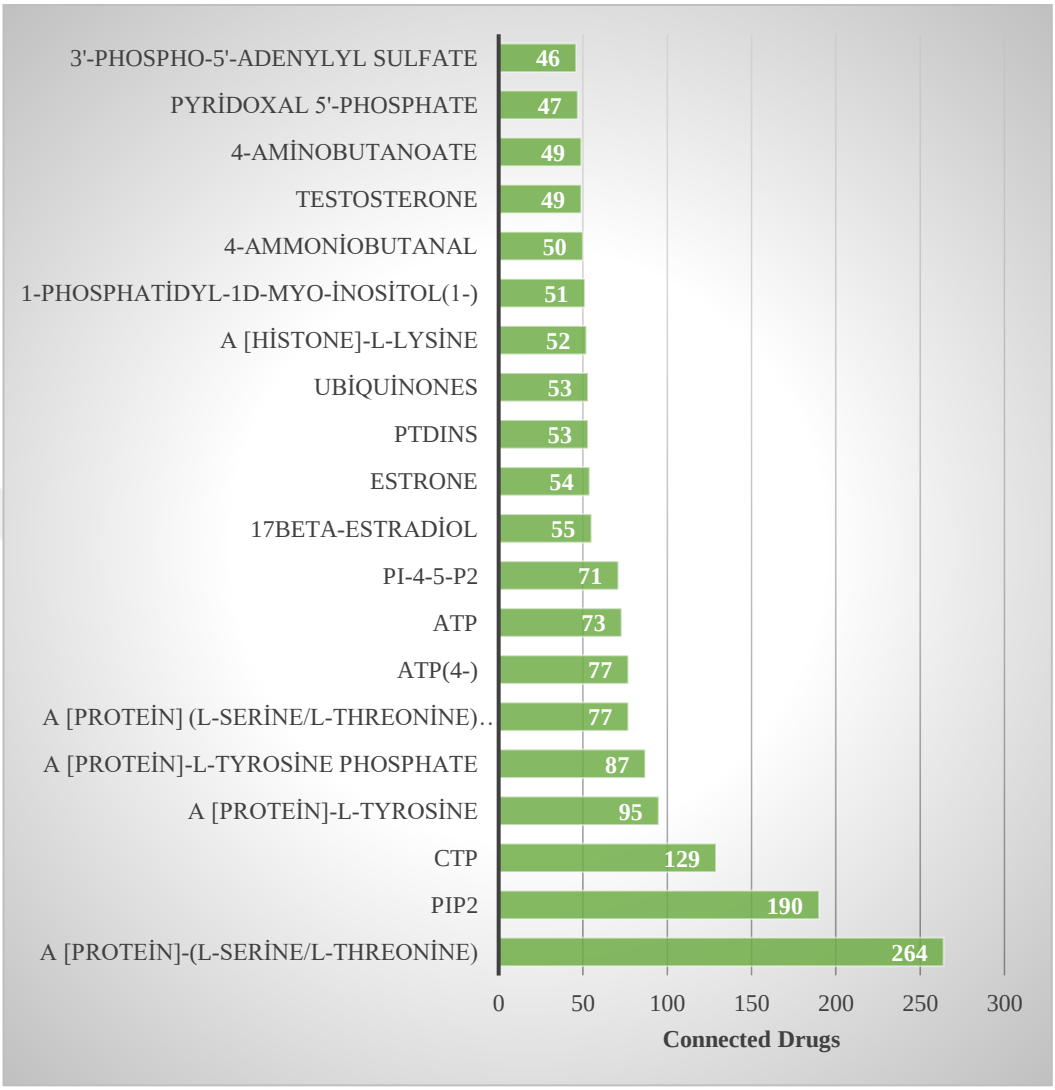
Table 3.3 Example for gene name to identifier conversion

Gene name - Identifier conversion		
Homosapiens Taxonomy ID:9606 Genbank common name: human		
GeneID	**Symbol**	**Synonyms**
1	A1BG	A1B|ABG|GAB|HYST2477
2	A2M	A2MD|CPAMD5|FWP007|S863
3	A2MP1	A2MP
9	NAT1	AAC1|MNAT|NAT-1|NATI
10	NAT2	AAC2|NAT-2|PNAT
11	NATP	AACP|NATP1
12	SERPINA3	AACT|ACT|GIG24|GIG25
13	AADAC	CES5A1|DAC
15	AANAT	DSPS|SNAT
16	AARS	CMT2N|EIEE29
17	AAVS1	AAV
18	ABAT	GABA-AT|GABAT|NPD009

3.1.3 Drug Targets

A drug target is a crucial part of any drug development program. A drug target can be a protein or enzyme. Drug targets function is changed, removed or corrupted after the drug treatment. Table 3.4 shows the top 20 drugs which connected with consumption-controlled-by edge type to target proteins.

Table 3.4 The details of drug name – drug target count Table. Total number of drug targets for most 20 connected drugs count are listed in the “connected drugs”, respectively



For example, the known targets of “Aclacinomycin A” drug are TOP1, TOP2A, and TOP2B proteins. Drug targets were derived from screens using cell culture or whole organisms and phenotypic or molecular readouts. Discovery of the direct target(s) of a drug is often the most challenging and time-consuming step of the drug development process (Warren, 2011).

Search Tool Interactions of Chemical (STITCH) is a searchable database that integrates information about interactions of proteins and chemicals. Chemicals are linked to other chemicals and proteins by evidence derived from experiments, databases and the literature. Some text mining and similarity methods are used for

detecting relations between the chemicals or proteins. STITCH has alot of interactions between small molecules and proteins within more than 1100 organisms (Sang, Hyun & Tae, 2011 ; Szklarczyk et al., 2016).

For this project, we have set drug targets for 14 different drugs from the STITCH database. These drugs have been treated with lymphoma cancer cells, the data set of each drug and the targets of these drugs are listed in Table 3.5.

Table 3.5 Drugs – Stitch Target Table. Drug names and drug targets from the STITCH database in the lymphoma tissue are listed in the table

Drugs	**Stitch Target**
Aclacinomycin A	TOP2A;TOP1;TOP2B
Blebbistatin	MYH2;MYH9;MYLK;MYH14;RAC3;RAC1
Camptothecin	TOP1;TP53;CASP3;BAX;PARP1;JUN;CHEK1;RAD51;ABCG2;ANXA5;CDK1;ABCB1;FASLG;MAPK8;RECQL;BCL2;E2F1;HIF1A;HSPA4;ATF3
Mitomycin C	TP53;TOP1;RAD51;BRCA2;BRCA1;NQO1;CASP3;CASP8;XRCC2;MX1;PPP1R15A;CD80;FGF2;POR;ABCB1;STX1A;RAD51C;FSCN1
Geldanamycin	HSP90AA1;ERBB2;AKT1;HIF1A;HSPA4;PTK2;CDC37;HSP90AB1;FKBP4;HSPA8;NR3C1;TP53;ZHX2;KDR;TEK;CFTR;UNC45B;IL6;DDX58;CSK;PAFAH1B1
Methotrexate	DHFR;TYMS;ABCC3;SLC19A1;ATIC;ABCC11;ABCG2;FPGS;GGH;CRP;QDPR;ABCC2;ABCB1;SLC22A6;ABCC4;MTHFR;SLC46A1;ABCC1;SLCO4C1;FOLR1;ALB;IL4;SLC22A8;SLCO1B1;SLCO1B3;SLCO1C1;AOX1;FASLG;SLC22A7;PCNA;SLC22A11;SLCO1A2;MAT1A
Monastrol	KIF11;RAP1A
H-7 Dihydrochloride	PRKACA; PKIA

Table 3.5 continues

Cycloheximide	CASP3;TP53;CASP8;IL6;NFKBIA;HIF1A;TNF;FOS;IL1B;ALB;GAPDH;ICAM1;INS;BTG2;ESR1;NGF;GREB1;IL1A;PTGS2;RPL3;FN1;IGF1;CCND1;CDKN1A;VEGFA;MAPK8;ATF4;CDH1;CSF2;RGS2;TFRC;CD4;FGF2;IL2;CYP1A1;CTGF;CCL5;APOLD1;EIF4EBP1;ALOX5;AR;POMC;PDGFB;GCG;PGR;CD44;LYZ;BDNF;FTH1;HMOX1;SLC2A1;DDIT3;SOCS3;CXCL10;KNG1;PLAT;CFTR;SGK1;BRCA1;GNRH1;SERPINE1;POLI;SLC3A2;PXN;NOS3;IGFBP1;IL10;IGFBP3;EDN1;XBP1;HTT;PENK;IL11;BMP2;CXCL9;C5AR1;LEP;RRN3;PTPN13;TGFBI;JMJD6;PTHLH;F2R;MET;EPO;IGFBP5
Etoposide	CASP3;TOP2B;TP53;CASP8;BAX;BCL2L1;CASP2;CDKN1A;TOP2A;FASLG;CYP3A4;PARP1;AKT1;BAK1;ABCG2;CASP9;PTEN;APAF1;CASP7;ABCC3;ABCB1;H2AFX;ABCC1;ABCC2;CASP6;MAPK8;FAS;DCK;HIPK2;RB1;MCL1;LMNB1;TOP1;CYCS;TNFSF10;BCL2L11;ATM;BCR;TOP3A;AFF1;TNFRSF10B;CDK2;HRAS;TNFRSF10A;JUN;ERCC1;ABCC6;GSTP1;HSPA4;IGF1;BIRC5;HIF1A;FOS;BDNF;TXNIP;MMP1;GDF15;PTGS2;JAK2;EGR1;MMP2;UGT1A1;PEBP
Trichostatin A	HDAC1;HDAC2;HDAC9;HDAC8;HDAC7;HDAC6;HDAC3; HIST1H4C;HIST1H4F;HIST1H4B;HIST2H3PS2;HIST2H4A; HIST2H4B;H3F3A;HIST1H4E;HIST1H4A;HIST4H4;HIST1H4K;HIST1H4L;HIST1H4J;HIST1H4I;HIST1H4D;HIST1H4H;H3F3B;HDAC4;LCOR;HDAC10;ESR1;CCND1;HDAC11;HDAC5;PGR;TNFRSF10B;CCND2;AR;B2M;CIITA;MLH1;PPARG;ABCC10;ERBB2;VEGFA;CDKN2C;FOXP3
Vincristine	TP53;CASP3;CYP3A4;CYP3A5;BUB1B;PTGS2;MIXL1

Table 3.5 continues

Doxorubicin hydrochloride	ABCB1;TP53;CASP3;ABCC1;TOP2A;NOS3;BIRC5;CASP9;ATM;AKT1;PARP1;CASP8;PTEN;VEGFA;CDK1;SOD2;CHEK1;JUN;PCNA;HIF1A;PTGS2;CASP7;THBS1;ABCG2;MAPK8;MYCN;NOS1;RELA;GSR;PLAU;HIPK2;BIK;GDF15;LGALS3;CLU;MAPK14;BIRC3;ERBB2;TOP2B;FASLG;BAX;CDKN1A;GADD45A;RAB6C;TNFRSF10A;RRM2B;ABCC3;TNFRSF10B;RALBP1;MTOR;DNMT1;MDM4;KAT2B;FOS;CSF2;CBR3;POR;TNF;BCL6;NPPA;CFLAR;ABCC2;MAP3K5;KDR;ABCB11;CYP3A4;ABCC6;NQO1;CDC25A;CBR1;IGFBP3;PRODH2;TNNT2;SLC22A16;JAK2;AR;FTL;FTH1;HMOX1;EGR1;FAS;CAPN6;IL10;TRAF1;BECN1;WT1;NOS2;BCL2L1;CYP2B6;CD274;ATF3;TWIST1;H2AFX;AKR1A1;KLF4;PDGFB;CD2AP;SPHK1;SLC6A6;NCL;ABCC10;MDM2;EGFR;JAK1;BHLHE40;BCL2;AKR1C3;
Rapamycin	MTOR;FKBP1A;RPS6KB1;EIF4EBP1;RPTOR;RPS6;FKBP3;EIF4E;RHEB;AKT1;IRS1;CCND1;AKT2;HIF1A;VEGFA;CDK2;IL2;CDKN1B;EEF2;JUN;CCND3;IL10;RB1;RYR1;CYP3A4;ABCB1;PDCD4;RPS6KB2;EIF4G1;FKBP2;FKBP4;RPS6KA1;PTEN;FKBP5;RICTOR;IL2RA;CDK4;MAPK8;PPIA;FKBP1B;CD28;INS;EIF4B;IGF1;ULK1;HTT;DDIT4;SMG1;FOXP3;CD4;EGFR;PPP2R4;STAT3;TGFB1;CASP3;HSP90AA1;PPARG;SQSTM1;LEP;IRS2;GSK3B;NOS3;PRKDC;ATIC;BAX;FBXW7;MYC;GZMB;ULK2;RAC1;RAC3;BIRC5;PCNA;ERBB3;IL6;IL12B;IL3;STAT1;CASP8;CD86;CDK1;IL7;EIF2AK4;RRN3;EIF4EBP2;MMP9;CSN1S1;NGF;FN1;CCND2;S100A4;AKT1S1;F3;CD44;KCNA1;BDNF;CCR7;CXCL12;CCR5;POLI;HMOX1;TBX21;TLR4;MAP3K5;PDGFRB;CCL2;CYP3A5;TGFBI;IGFBP3;NPM1;VEGFC;IGFBP1;SMAD2;PROM1;SOCS3;SERPINE1;SMUG1;SLC2A1;EEF1A1;PLAT;CD36;LMNA;GFAP;RPL30;SELE;LDLR;ODC1;LDLR

3.2 Pre-Processing of Pathways

3.2.1 BFS (Breadth-First Search)

Breadth First Traversal (or Search) for a graph is similar to Breadth First Traversal of a tree. The only catch here is, unlike trees, graphs may contain cycles. To avoid processing a node more than once, we use a Boolean visited array. In this problem all vertex are in alphabetical order so we don't need to have any Boolean visited array control. For simplicity, it is assumed that all vertices are reachable from the starting vertex ("Breadth First Traversal for a Graph," n.d.). A BFS example for Aclacinomycin A drug before the algorithm is shown in the Figure 3.3.

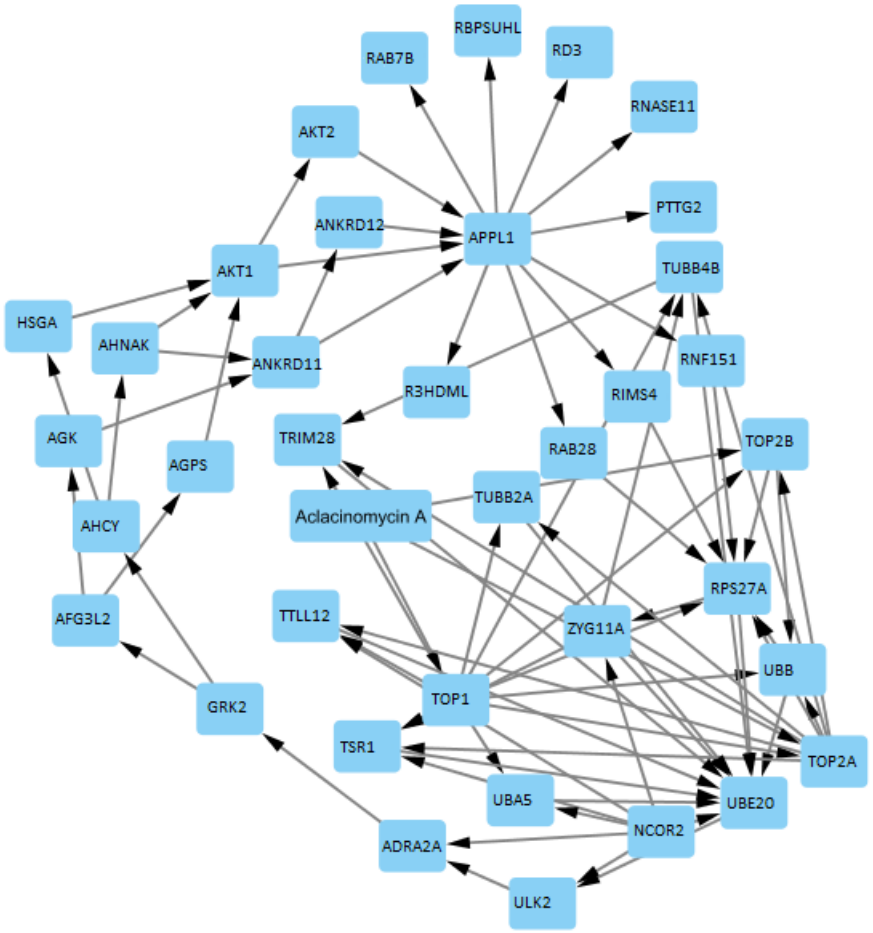


Figure 3.3 BFS Example for AclacinomycinA drug

3.2.2 Levelize a KEGG Pathway with BFS

Levelize a Kegg Pathway starts with finding out start point, this is the specified drug. Starting point means also level 1.Level 1 contains only drugs. Level 2 contains proteins which target of specified drug. Then levelization process continues till no protein left. An example levelization for “Aclacinomycin A” is shown in Figure 3.4. Briefly, each drug has its own network. The network is made up of protein networks which are larger than drug targets while drug targets are lower than the children (Szklarczyk et al., 2016 ; "Breadth First Traversal for a Graph," n.d.). The algorithm is shown in the Table 3.6.

Table 3.6 Pseudo code for levelize a KEGG pathway with BFS Algorithm

```
Input :Directed graph G stored in-adjacency and out-adjacency list format.
outAdj(x):out-adjacency list of node x.
 Initialization  :
    for each vertex x ∈ V do
      if in-degree(x) = 0 then
        color(x)=Black
        d(x)=0
        ENQUEUE(0,x)
      else
        color(x)=White
Levelization     :
    While  Q <> ∅
      x  =  DEQUEUE(Q)
      for  each  vertex y ∈ outAdj(x) do
          if color(y) = White then
          color(y) = Black
          d(y) = d(x)+1
          Vd(y)= Vd(y) U{y}
           ENQUEUE(Q,y)
return {V0,V1,V2,..Vl-1}
```

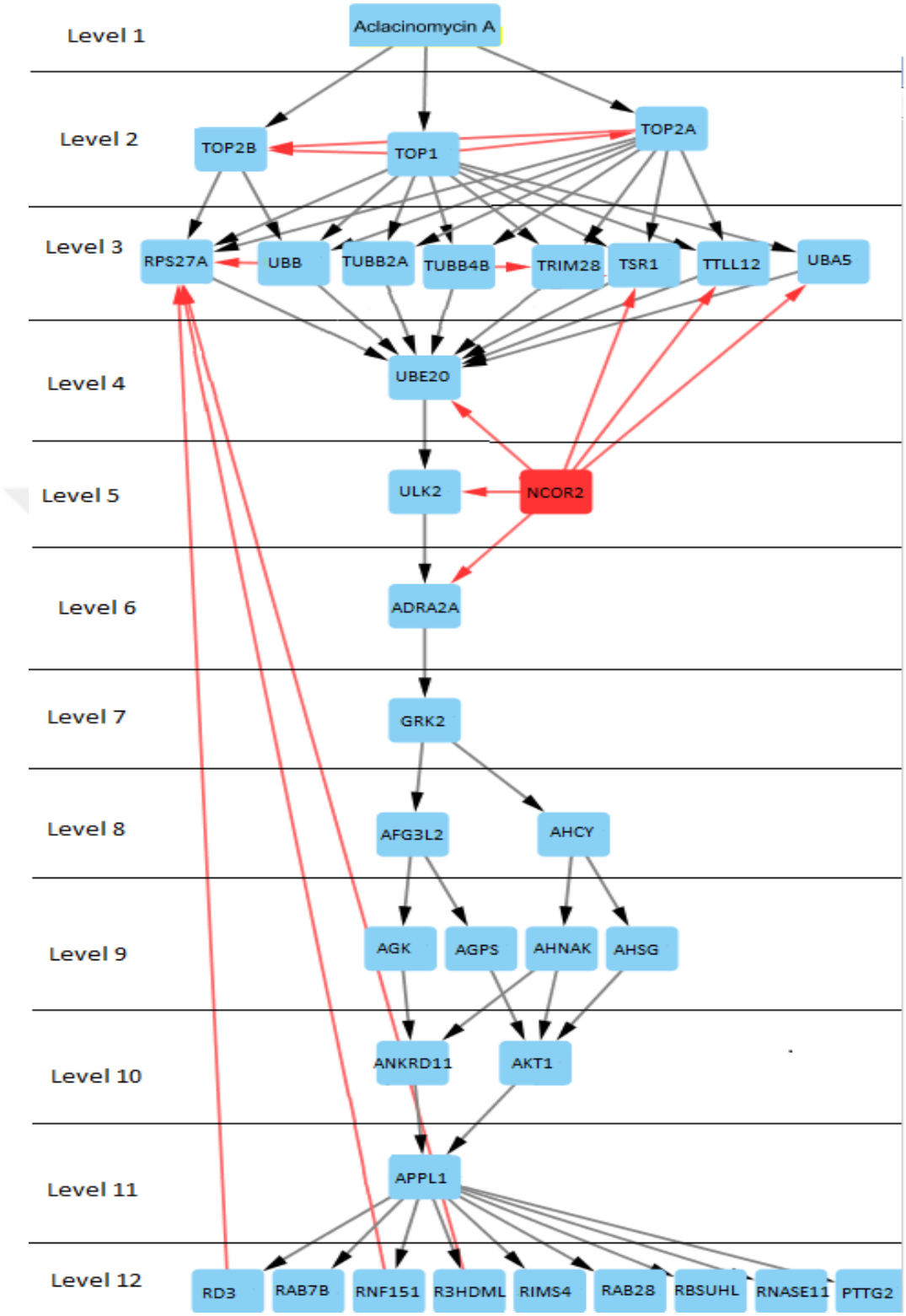


Figure 3.4 The levelization network for AclacinomycinA drug. The red edges and nodes were removed for abstaining from the cycles in the original pathway

3.2.3 Generate Drug Network

After BFS levelization , i used a recursive procedure for deletion of unused edges and nodes. So , the cycles were removed in the BFS tree (Szklarczyk et al., 2016 ; "Breadth First Traversal for a Graph," n.d. ; Cline et al. ,2007). This process is shown in the Table 3.7. In Figure 3.4, red nodes and edges were removed for abstaining from the cycles in the final drug tree. Final drug network for Aclacinomycin A drug is shown in the Figure 3.5. After the levelization and cycle deletion stages, the network information (total number of edges, nodes, levels, drug targets) for every drug in this project is listed in Table 3.8.

Table 3.7 Pseudo code for Generate Drug Network up to water-flow method

```
Create Drug Network for  getting drug network up to water-flow method
Input :Directed graph G stored in-adjacency and out-adjacency list format.
outAdj(x):out-adjacency list of node x.
T{p} :Set of target nodes showing process in G levelization information V0,V1,V2,..Vl-1 obtained
by running Network Levels-Obtained by BFS
Initialization :
  For each level = 0,1,2,..,l-1 do
       For each vertex x { Vi do
            For each vertex y { outAdj{x} doy
            if  NodefromLevel(y)>level
           Delete(x→y)
For each vertex x { V1 do
        For each vertex y { outAdj{x} do
       RecursiveTravel(y)
RecursiveTravel  :
  if outAdjExists(x)
       For each vertex y { outAdj{x} do
          RecursiveTravel(y)
          WriteFile(x→y)
DrugNetworkOrganize :
   SortAscending(FirstVertex)
         For each vertex x { Vi do
              SortAscending(SecondVertex)
```

Table 3.8 The details of every drug network in this project. Total number of edges, nodes and drug targets are listed in the "# of Edges", "# of Nodes" and "# of Drug Targets" columns.

Drug Name	# of Edges	# of Nodes	# of Levels	# of Drug Targets
Aclacinomycin A	13979	4791	12	3
Mitomycin C	251979	12955	8	18
Rapamycin	81717	13003	7	128
Doxorubicin hydrochloride	118752	12875	7	108
H-7 Dihydrochloride	59644	4430	11	2
Geldanamycin	116033	12826	9	21
Methotrexate	274265	13011	8	33
Vincristine	263037	11558	11	7
Blebbistatin	55838	5690	11	6
Monastrol	94305	8039	11	2
Camptothecin	148146	12991	7	20
Trichostatin A	139873	11581	10	44
Etoposide	131129	13034	8	63
Cycloheximide	116875	12797	10	86

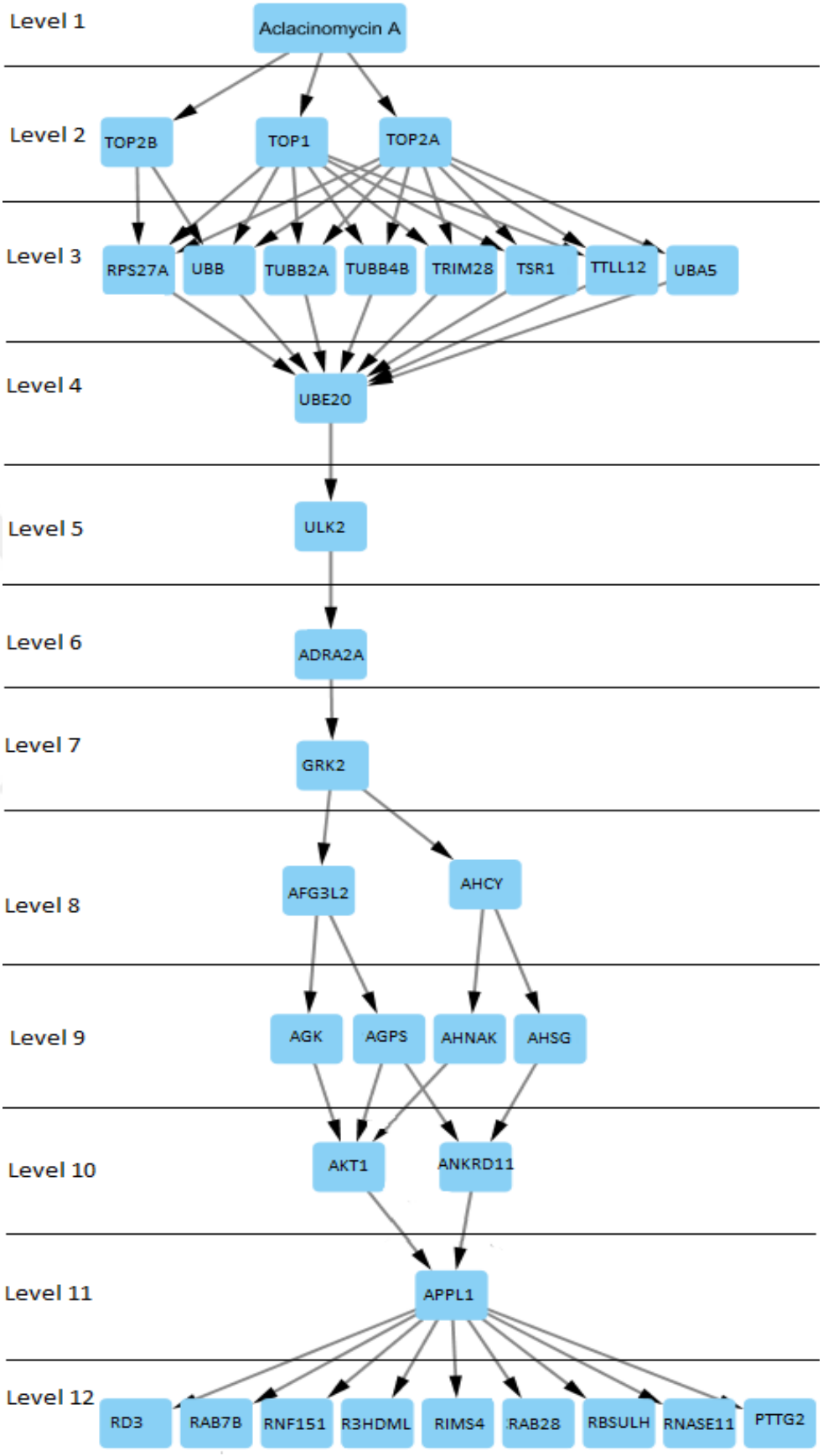


Figure 3.5 Drug Network for AclacinomycinA drug

3.3 Score Flow Algorithm

The top-level of the vertex is drug. That means , level 1 is drug. Level 2 refers to target proteins that are related to drug. The proteins initial values are loaded from drugs micro array file. The edge value is the last value of the parent node divided by the number of edges of the parent node. Then acording to BFS algorithm children nodes connect with parent nodes with edge from the upper level. Parent nodes scores pass through the children nodes acording to arithmetic average value. The final values of the proteins are calculated from the values from the beginning added to all the values from all the edges which is connected to the protein ("Breadth First Traversal for a Graph," n.d. ; Sang, Hyun & Tae, 2011 ; Szklarczyk et al., 2016). The pseudo-code of this algorithm is given in Table 3.9.

Table 3.9 Pseudo code for the Score Flow Algorithm

```
Score: indicates initial score of each node provided by microarray file
outScore: contains out-score of each node
marray: indicates self-score of each node provided by microarray file
outAdj(x): out-adjacency list of node x.
ECount(x): Number of edges from x node
All(x): List of all proteins in the network
Levelization information V0,...,Vl-1 is obtained by running the BFS algorithm
Initialization:
        For each vertex x in All (x) do
    If marray(x) contains then
                Score(x) = marray (x)
      outscore(x) = marray (x)
    else
                Score(x)=0
                outscore(x) = 0
Score Computation:
   For each level = 0,...,l-1 do
      For each vertex x in Vi do
        For each vertex y in outAdj{x} do y
                                outscore(y) = outscore(y)+ outscore(x)/ECount(x))
```

3.4 Statistical Significance of Output Scores Algorithm

First, the original score for each gene is calculated by dividing the final score by the initial score. This formula showed in the formula 4.1.

$$OriginalScore = FinalScore \div InitialScore \tag{4.1}$$

Then, the input scores of all the genes are taken and they are assigned randomly to all the genes. The final gene scores are calculated according to the score flow algorithm by using randomly derived initial scores. The random final score of each gene is calculated by dividing the final score obtained from the randomly generated score by the randomly generated initial score. This formula showed in the formula 4.2.

$$RandomScore = RandomFinalScore \div InitialRandomScore \tag{4.2}$$

If the random gene score is between 0.9 and 1.1 times the original score, the error for that gene is increased by one. The Formula of the error calculation is showed in the formula 4.3.

$$\begin{aligned} &OriginalScore(gene) \times 0.9 \le RandomScore(gene) \le OriginalScore(gene) \times 1.1 \\ &Error(gene) = Error(gene) + 1 \end{aligned} \tag{4.3}$$

This calculation is repeated for 1000 times. Finally, for every gene an error is computed via *p*-value. The *p*-value shows the probability of having the final score of gene with a randomized initial data. If the *p*-value (i.e., error) of a gene *x* is less than 0.05 by using random initial inputs, then the actual output score of the gene *x* is assumed to be statistically significant. Table 3.10 shows how the p-values are computed.

Table 3.10 Pseudo code for the calculating p-value of output score of each gene

All(x): List of all proteins in the network
FinalScore: contains final score of each node
InitialScore: contains initial score of each node
OriginalScore : contains (FinalScore / InitialScore)of each node
RanFinalScore: contains random final score of each node changes in every iteration obtained by random-input running.
RanInitialScore: contains random initial score of each node changes in every iteration.
RandomScore : contains (RanFinalScore / RanInitialScore) of each node
Error : contains error of each node
Initialization:
For each vertex x in All (x) do
OriginalScore (x) = FinalScore(x) / InitialScore(x)
Random-Input Running:
For each iteration = 0,...,1000 do
For each vertex x in All (x) do
RandomScore (x) = RanFinalScore (x) / RanInitialScore (x)
if Original Score (x)*0.9 < Random Score (x) <1.1*Original Score (x)
Error(x)=Error(x)+1

3.5 The Jaccard Index

First, the number of gene is calculated in the list of all statistically significant proteins for the two drug networks. Then the number of common gene is calculated in the list of all statistically significant proteins for the two drug networks. Then the number of union gene is calculated in the list of all statistically significant proteins for the two drug networks. Union gene set formula is showed in the formula 4.4.

$$UnionGeneSet = GeneSet1 + GeneSet2 - InterSection(GeneSet1, GeneSet2) \quad (4.4)$$

The Jaccard index is found by dividing the number of elements in the intersection gene set by the number of elements in the union gene set. This formula showed in the formula 4.5.

$$The\,Jaccard\,Index = Union\,Gene\,Set\,Count \div Inter\,Section\,Gene\,Set\,Count \quad (4.5)$$

The reason for calculating the Jaccard index between two drugs is to calculate the similarity rate of these two drugs. The Jaccard Index score can range from 0 to 1 and higher score means better similarity results.

Table 3.11 Pseudo code for the calculating the jaccard index between drug1 and drug2

```
All (Drug1) : List of all statistically significant proteins in the Drug1 network
All (Drug2) : List of all statistically significant proteins in the Drug2 network
Ct(Drug1) : Contains number of proteins in the Drug1 network
Ct(Drug2) : Contains number of proteins in the Drug2 network
Ct_Union(Drug1,Drug2) : Contains number of proteins in the Union Drug1 and
Drug2 network
Ct_Intersection(Drug1,Drug2) : Contains number of proteins in the Intersection
Drug1 and Drug2 network.
JI(Drug1,Drug2) : Contains intersection ratio between Drug1 and Drug2
Initialization :
   For each vertex x in  All (Drug1)  do
      Ct(Drug1) =   Ct (Drug1) +1
      For each vertex y in All(Drug2) do
          If  x == y
            Ct_Intersection(Drug1,Drug2) = Ct_Intersection(Drug1,Drug2) +1
   For each vertex y in All(Drug2) do
       Ct (Drug2) =  Ct (Drug2) +1
Output :
   Ct_Union (Drug1,Drug2)=Ct(Drug1)+Ct(Drug2)-Ct_Intersection(Drug1,Drug2)
   JI(Drug1,Drug2) = Ct_ Association (Drug1,Drug2) / Ct_Union(Drug1,Drug2)
```

3.6 Implementation

3.6.1 Gene Expression Data with Gene Identifier

Gene expression data extraction process is the conversion process which we used to calculate gene expression data to provide as the initial input of the algorithm. Hence, for each of 14 drugs, a 24-hour drug-treated data were given on columns Y, Z and AA; and 24-hour control sample (DMSO) data on columns AB, AC and AD. We then calculate the median values separately for drug-treated values and control samples. Gene expression file before processing is shown in Table 3.12.

Table 3.12 Gene Expression File before processing

Y	Z	AA	AB	AC	AD
TH001_AP_100427_01C_C07	TH001_AP_100427_01C_C08	TH001_AP_100427_01C_C09	TH001_AP_100427_01C_A06	TH001_AP_100427_01C_A12	TH001_AP_100427_01C_D04
Aclacinomycin A	Aclacinomycin A	Aclacinomycin A	DMSO	DMSO	DMSO
24	24	24	24	24	24
IC20	IC20	IC20	0.01	0.01	0.01
5.21	5.19	5.31	4.98	4.77	5.3
4.82	4.7	4.59	4.6	4.59	4.53
4.87	5.15	5.1	5.01	4.64	5.07
5.48	5.11	5.46	5.42	5.11	5.46
4.17	3.87	4.32	4.1	4	4.09
4.25	4.3	4.19	4.2	4.21	4.12
4.19	4.3	4.39	4.22	4.14	4.44
7.18	7.25	7.11	7.12	6.52	7.28
4.22	4.48	4.45	4.54	4.66	4.62
4.4	4.35	4.25	4.02	4.13	4.12
4.82	5.23	4.69	4.91	4.81	5.2
6.33	6.81	6.49	6.65	6.18	6.58

The last step of the gene expression data is coded in C# language in .Net IDE and the final gene expression file is generated. Gene names are converted to Gene Identifier with the help of homo_sapiens.gene_info file from Pubmed-NCBI

(https://www.ncbi.nlm.nih.gov/pubmed/). Then drug-treated and control samples for each gene are given as inputs to calculate the scores separately.

3.6.2 Pathway Data up to Drug Target, Score Calculation and Statistical Significance Codes

Pathway Commons contains free pathway database and KEGG pathways can be obtained from www.pathwaycommons.org website. A pathway includes two elements: a node and a connection edge. The source node represents a protein or drug; the target node represents a protein alone. The connection represents the biological relationship or type of reaction between the source and target nodes. We focused on the "interacts-with" and "consumption-controlled-by" connections on the KEGG pathways, because they are the only bridges that includes drug-protein interaction information. The connection types and their total counts in Pathway Commons specific to KEGG pathways are shown in Table 3.13.

Table 3.13 Pathway Commons Connection Type Coverage

Connection Type	**Count**
consumption-controlled-by	20721
used-to-produce	11212
reacts-with	1632
chemical-affects	62806
interacts-with	1912848
neighbor-of	2443172
in-complex-with	110622
controls-state-change-of	76856
catalysis-precedes	71313
controls-production-of	21440
controls-transport-of-chemical	3000
controls-transport-of	4293
controls-phosphorylation-of	11890
controls-expression-of	149738
Total	4901543

The known drug targets information were obtained from the STITCH database. We only used "interacts-with" connections, because they focuses on protein-protein interaction relationship. Pseudo codes for Pathway up to drug target, Score Calculation and Random-Input Running for detecting the statistically significant genes are shown in Table 3.14. This algorithm was coded in Eclipse environment with the Java programming language.

Table 3.14 Pathway Data up to Drug Target, Score Calculation and Random-Input Running for detecting the statistically significant genes pseudo codes

```
Hashtable <String, String[]> graph : Save the whole relations between genes up to Kegg Pathways
Hashtable <String, Double> output_score : Save the final score for a gene
Hashtable <String, Double> score  : Save the first score for a gene
Hashtable <String, String> adjacency_matrix : Save the relations between the proteins
Hashtable <String, String[]> marray : Save the microarray file values in a gene
Hashtable <Integer, Vector<String>> bfs_level: Save the proteins in the level
Hashtable <String, Double> rate : Save the percentage between final and initial score in a gene
Hashtable <String, Integer> error_count : Save the error of a gene,Initial value is 0 for all genes
All(drug): List of all drugs and the drug targets.
All_Genes(drug) : List of all genes up to specified drug
Initialization :
graph = Get_All_Kegg_Pathway() ; // All Kegg Pathways store in the graph hashtable.
Pathway Data up to Drug Target :
For each drug in All (drug) do // Applied for 14 drugs
    marray = ReadArrayFile (drug) ; //Get microarray file to marray hashtable
    bfs_level, adjacency_matrixPerformBFS(drug_target) ; //Sort entire data by drug targets
    score = ConstructArrayScore(); //Save marray hashtable in score hashtable
 Score Flow Calculation and calculate statistical sigficant genes in the drug network
 For each iteration = 0,...,1001 do
    InitializeMatrix();//Obtain output_score hashtable up to score and  adjacency_matrix hashtable
    ScoreComputation()//Run Score Flow Algorithm up to water-flow model
    If(iteration == 0 ) then   SaveScoreFlowResults(); // Save the score flow algorithm results
      For each gene in  All_Genes(drug) do //All genes in the specified drug
        rate.put(gene, output_score.get(gene)/score.get(gene));//contains rate of each node
    Else //Score Flow results up to randomize generated data
      For each gene in  All_Genes(drug) do //All genes in the specified drug
        result=output_score.get(gene)/score.get(gene);//Calculate rate up to random input running
        if((rate(gene)*0.9)<result && result<(1.1*rate(gene))
          error_count.put(gene,error_count(gene)+1);//Increase 1 the error of the gene
    score = ConstructRandomArrayScore(); //Obtain score data from randomize the initial values
 For each gene in  All_Genes(drug) //All genes in the specified drug
    If(error_count(gene)<50)    //If error count is less than p-value
      SaveTheGene(level,error_Count,ProteinID);//Save the statistically significant genes
```

3.6.3 The Jaccard Index between the drugs and Most Common Used Proteins

Pseudo codes for the Jaccard index between drugs and most common used proteins are shown in Table 3.15. This algorithm was coded in .Net environment with C# programming language.

Table 3.15 The Jaccard Index between the drugs and Most Common Used Proteins pseudo codes

```
Hashtable protein_table <String, String[]>  :Save the  statistically significant genes in the drug
Hashtable drug_table <String, String[]>     :Save the drugs that statistically significant genes use
All(drug)                                   :List of all drugs and the drug targets.
All_Genes(drug)                             :List of all genes up to specified drug
All_Genes()                                 :List of all statistically significant genes
All_Drugs(protein)                          : List of all drugs in a statistically significant gene
Kesisim                                 :Save the intersection count between the drug-drug pair
Birlesim                                :Save the union count between the drug-drug pair
Benzerlik                               :Save the jaccard index percentage between the drug-drug pair
Initialization :
For each drug in All (drug) do // Gets the 14 drugs
    drug_table,protein_table =ReadFile(drug) // Set the proteins to the specified drug and protein table
Jaccard Index between the drugs :
For each drug1 in All (drug) do // Gets the 14 drugs
    For each drug2 in All (drug) do // Gets the 14 drugs
        kesisim = 0;  birlesim = 0;  benzerlik = 0; //reset the values
        For each gene1 in  All_Genes(drug1) //Gets the gene from the specified drug
            For each gene2 in  All_Genes(drug2) //Gets the gene from the specified drug
                If gene1 == gene2 //If the gene is common in the drug pair
                    kesisim++; //Increase 1 the intersection
                    exit for; //Exit from the for
        birlesim = All_Genes(drug1) +All_Genes(drug2) –kesisim //calculate the intersection
        benzerlik = (kesisim / birlesim) * 100;  // Calculate the jaccard index
        SaveFile(Drug1,Drug2,benzerlik) //Save Jaccard Index
Most Common Used Proteins :
For each gene in  All_Genes()  //Gets the all gene from the specified drug
    For each drug in All_Drugs(protein)  //Gets the drug which used the specified gene
        SaveFile(gene,drug) // Save the most common used protein
```

CHAPTER FOUR
EXPERIMENTAL RESULTS

4.1 Score Calculation

We used the score flow algorithm for each 14 drugs. One detailed instance for Aclacinomycin A drug is showed in the Figure 4.1 that displays the score computation from first to fifth levels to Aclacinomycin A drug network. Aclacinomycin A has three drug targets. They are TOP1, TOP2A, TOP2B which are in the second level of the tree. The drug's score is zero at first. So the drug edges set to zero which connected from first to second level. TOP2B protein is in the second level and input score is 8.48, the incoming edges are all set to zero. The output score for that gene is equal to the input score. This output-score is divided between its children (32 children); each child of TOP2B will get a score of 0.33 from TOP2B's outgoing edge. The algorithm continues till the deepest level of the network. After running the score flow algorithm for each drug, the last output scores of all genes are saved.

The highest influenced of a drug treatment proteins in the drug network was found up to difference between microarray score and the last score calculated by the score flow algorithm. The proteins with the one of the most difference in their scores are selected as the highest influenced ones and showed in Table 4.1. An example for Aclacinomycin A drug between first five levels score flow showed in the Figure 4.1

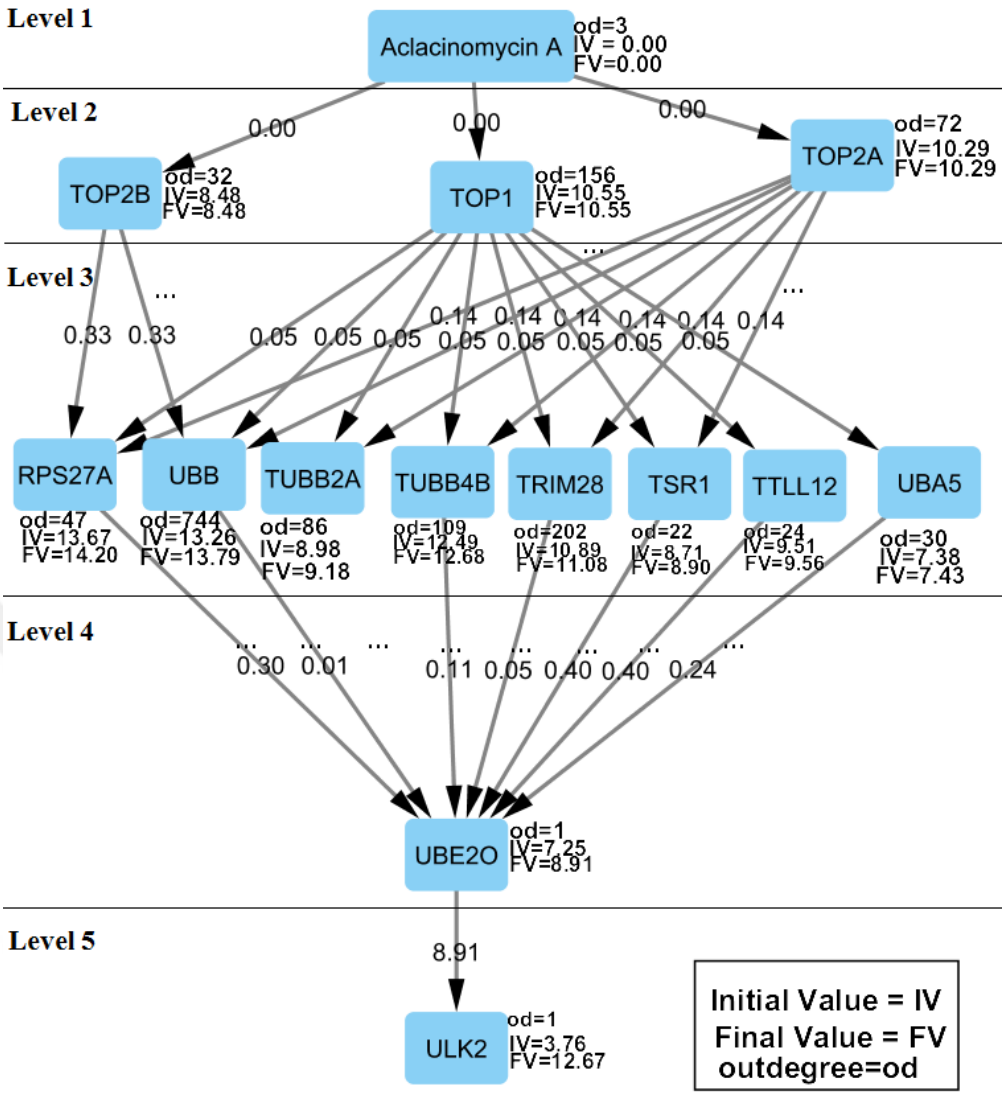


Figure 4.1 Score calculations for the “Aclacinomycin A” drug from first level to fifth level

The highest influenced proteins are not drug targets; because they are on lower levels, most usually in the third level, of the BFS tree. That conclusion is found in many latest studies (Isik, Baldow, Cannistraci & Schroeder,2015 ; Iskar et al.,2013). Those proteins which is in the faraway parts of the networks have changed importantly because of the drug treatments. Some of those proteins are found also in the different drugs. For instance, COX7A2 is the most effected protein in the Aclacinomycin A, H-7 Dihydrochloride, Methotrexate, and Mitomycin C drugs network. NFE2L1 protein is also found in the Methotrexate and Mitomycin C drug networks. USP15 protein is identified the most effected protein in the Camptothecin and Etoposide drug networks.

We performed a publications search about the previous studies about the most effected proteins monitored in the various drug treatments. A search found that the dys-regulation NFE2L1 protein can cause tumor (Oh, Rigas, Cho & Chan, 2012).The other research also showed that NFE2L1 protein is "related to the cell survival under stress condition" (Biswas, Kwong, Park, Nagra, & Chan, 2013). A previous study showed that "USP15 protein regulates the TGF-β pathway and USP15 has an important role in glioblastoma cancer" (Eichhorn et al. , 2012). Those proteins are one of the most influential because they are present in most of our drug treatment datasets and have the same roles for cancer cell treatment activities. Shortly, the literature also shows our results which is obtained from the score flow algorithm that can assist a better understanding about molecular responses of a cell in the metabolic level after a drug treatment.

Table 4.1 The highest influenced 3 proteins of all the 14 drugs. For each drug, three proteins with the most affected score change (showed as "Difference" column) were selected as the most affected proteins

Drug Name	**Protein Name**	**Initial Score**	**Final Score**	**Difference**	**Level**
Aclacinomycin A	ZRANB2	9.45	71.99	62.55	3
Aclacinomycin A	COX7A2	11.33	73.02	61.69	6
Aclacinomycin A	ECH1	9.50	70.26	60.76	6
Blebbistatin	TAF1	5.44	133.08	127.64	3
Blebbistatin	USP7	8.30	96.49	88.19	3
Blebbistatin	YME1L1	10.07	88.60	78.53	3
Camptothecin	USP15	5.70	225.42	219.72	3
Camptothecin	TEX10	6.97	155.05	148.08	3
Camptothecin	TFPI2	3.65	124.86	121.21	4
Cycloheximide	YEATS4	9.28	227.75	218.47	3
Cycloheximide	ZBTB7C	6.10	207.20	201.10	3
Cycloheximide	WDR61	8.43	171.72	163.29	3
Doxorubicin hydrochloride	ZCCHC11	5.15	258.38	253.23	3
Doxorubicin hydrochloride	VSIG8	3.70	190.53	186.83	3
Doxorubicin hydrochloride	ZNFX1	9.96	173.33	163.37	3
Etoposide	USP15	6.04	345.46	339.42	3
Etoposide	WDR26	8.96	180.13	171.17	3
Etoposide	SMURF2	4.46	153.33	148.87	3

Table 4.1 continues

Geldanamycin	TK1	6.39	218.23	211.84	3
Geldanamycin	ZNF346	6.03	156.38	150.36	3
Geldanamycin	TALDO1	8.69	152.95	144.26	3
H-7 Dihydrochloride	COX7A2	11.39	120.69	109.30	6
H-7 Dihydrochloride	ZNF30	5.80	83.15	77.35	4
H-7 Dihydrochloride	ZNRF4	4.15	81.37	77.22	4
Methotrexate	YARS	9.80	142.33	132.53	3
Methotrexate	NFE2L1	7.16	139.13	131.97	3
Methotrexate	COX7A2	11.02	115.07	104.05	4
Mitomycin C	ZMPSTE24	9.97	109.39	99.42	3
Mitomycin C	COX7A2	11.07	105.53	94.46	4
Mitomycin C	NFE2L1	6.94	100.78	93.84	3
Monastrol	RXFP1	3.82	133.42	129.61	4
Monastrol	SIAH2	11.40	122.16	110.76	4
Monastrol	TMEM223	8.45	115.43	106.98	4
Rapamycin	RCC1L	6.01	172.69	166.68	3
Rapamycin	TRIP6	6.75	150.61	143.86	3
Rapamycin	ZFP36	10.53	147.42	136.89	3
Trichostatin A	YTHDF3	7.86	236.42	228.57	3
Trichostatin A	YARS2	8.58	182.77	174.19	3
Trichostatin A	UBE2NL	4.21	168.68	164.48	3
Vincristine	PSMB10	0.00	131.70	131.70	4
Vincristine	SBDS	10.92	129.93	119.01	4
Vincristine	MRPL52	9.60	125.57	115.98	4

4.2 Statistical Significance of Output Scores

We tested the statistical significance of output score of each gene for each 14 drugs separately according to the same score flow algorithm with randomly derived initial scores. This calculation is repeated for 1000 times. So the genes whose p-value are less than 0.05 were assumed to be statistically effected due the drug treatment.

The detailed examples for all drugs are given from Figure 4.2 to Figure 4.15, which show the statistically significant genes with their significance level. According to those

figures most of the proteins are in the level 3 or level 4 but for Aclacinomycin A drug most of the proteins are in the level 6.

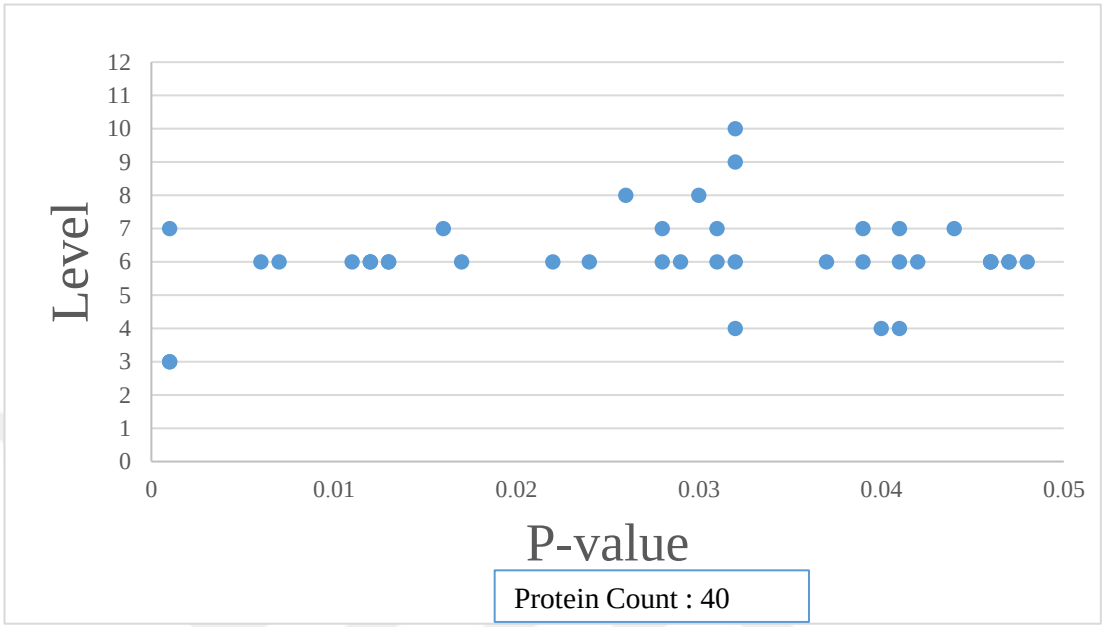


Figure 4.2 The genes whose p-value are less than 0.05 with level information for Aclacinomycin A drug

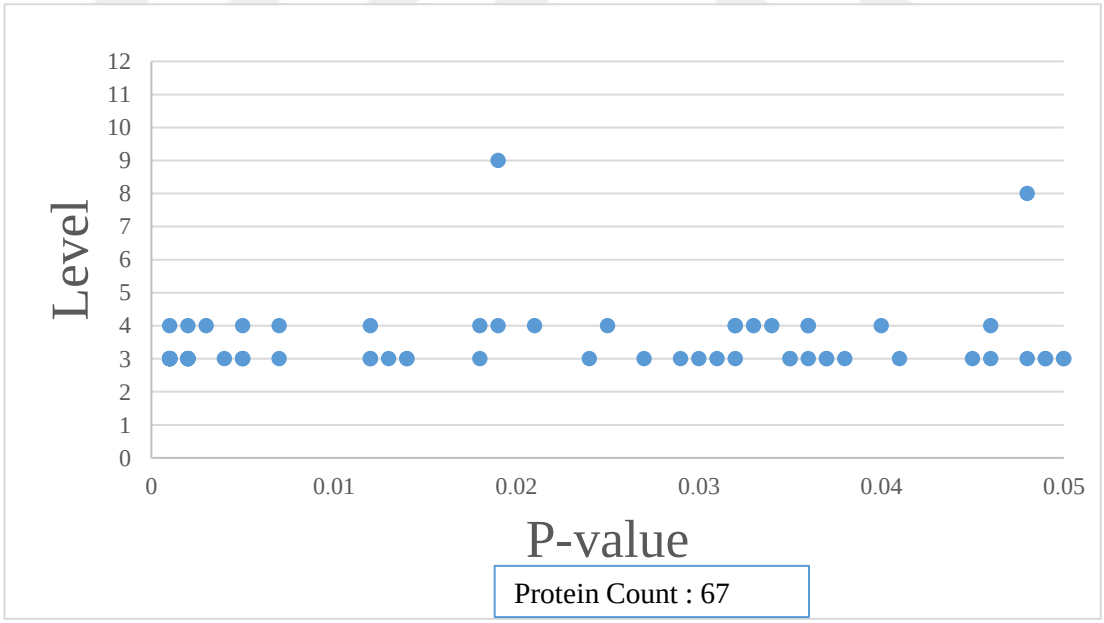


Figure 4.3 The genes whose p-value are less than 0.05 with level information for Blebbistatin drug

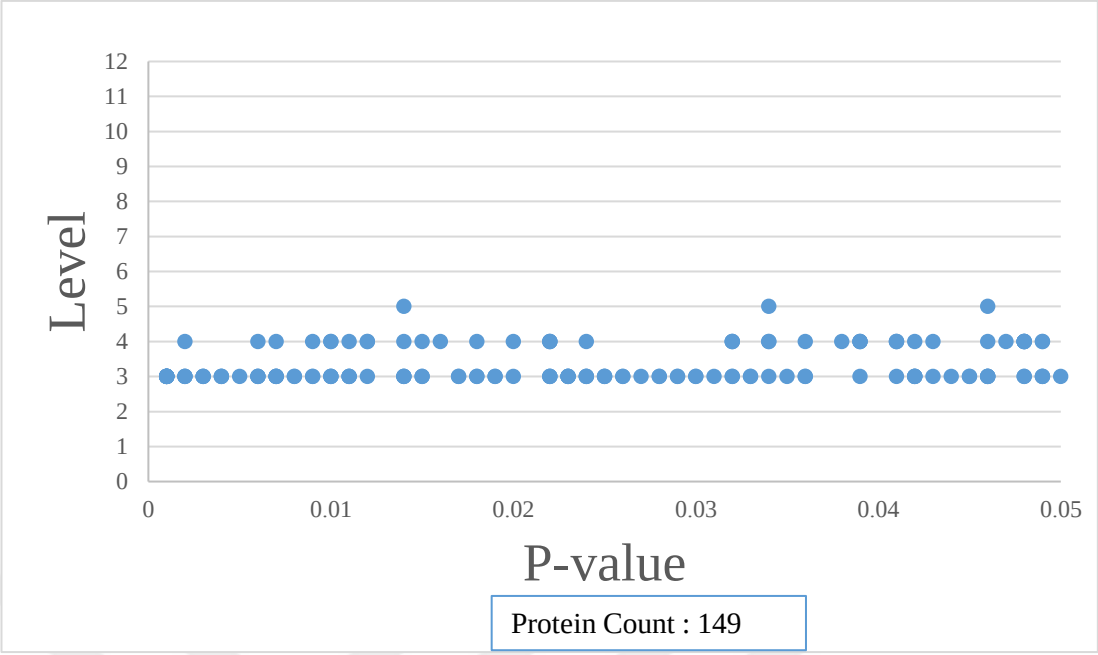


Figure 4.4 The genes whose p-value are less than 0.05 with level information for Camptothecin drug

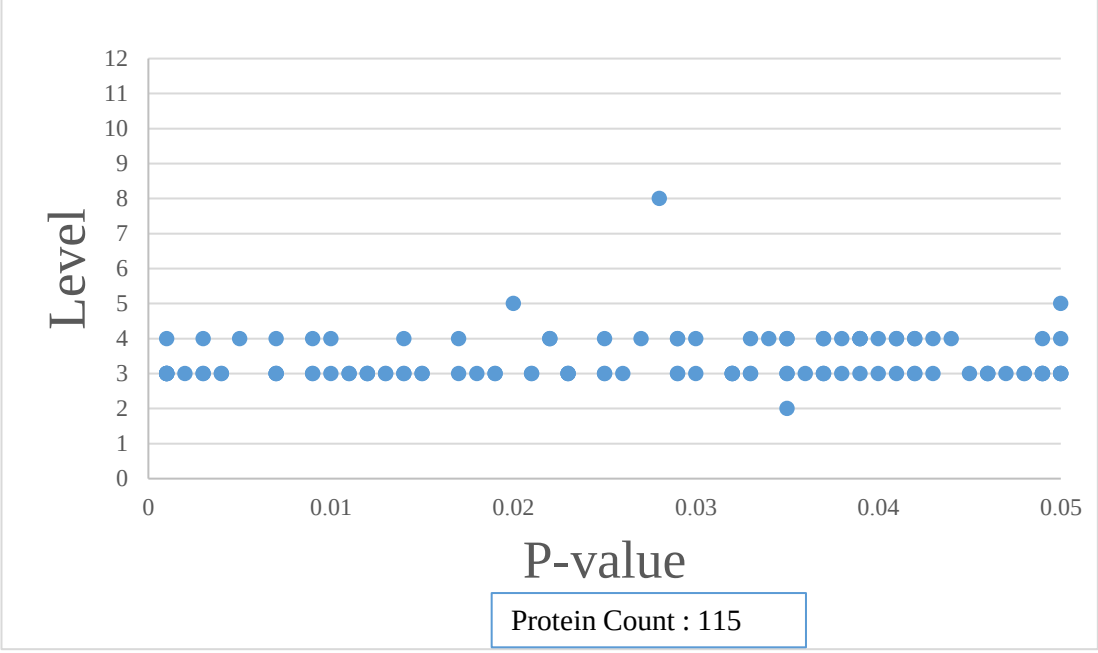


Figure 4.5 The genes whose p-value are less than 0.05 with level information for Cycloheximide drug

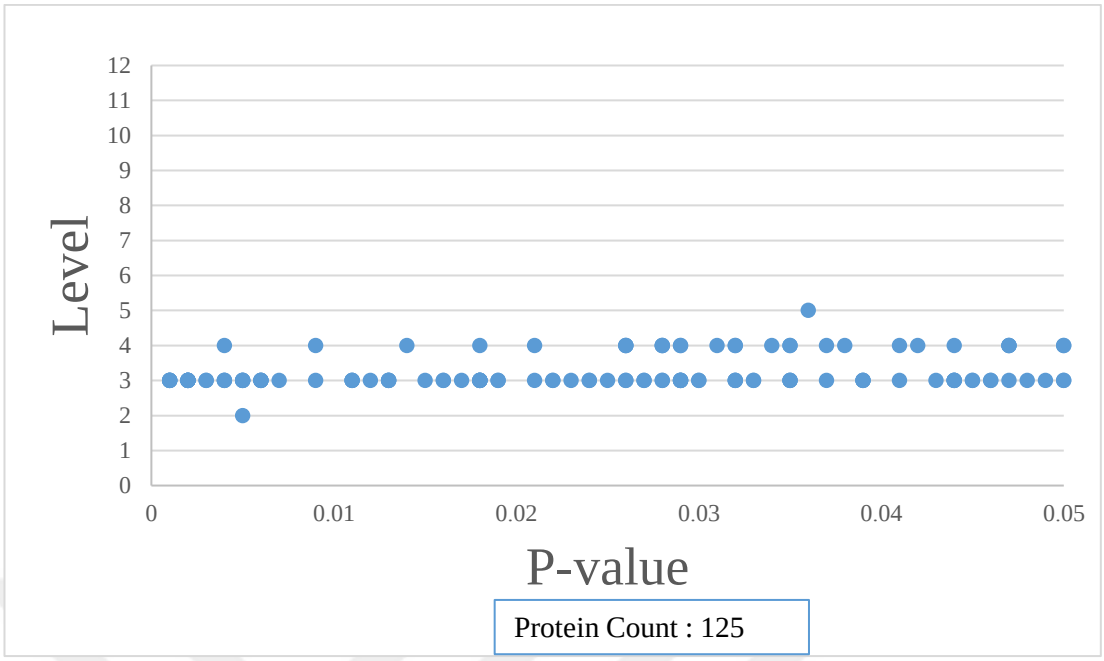


Figure 4.6 The genes whose p-value are less than 0.05 with level information for Doxorubicin hydrochloride drug

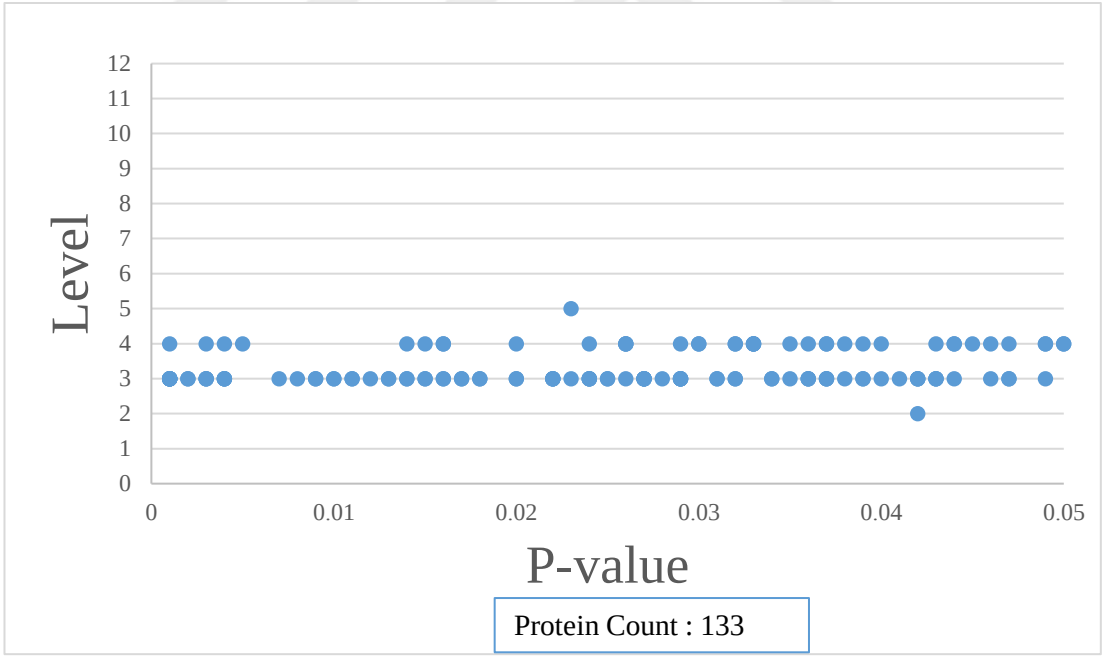


Figure 4.7 The genes whose p-value are less than 0.05 with level information for Etoposide drug

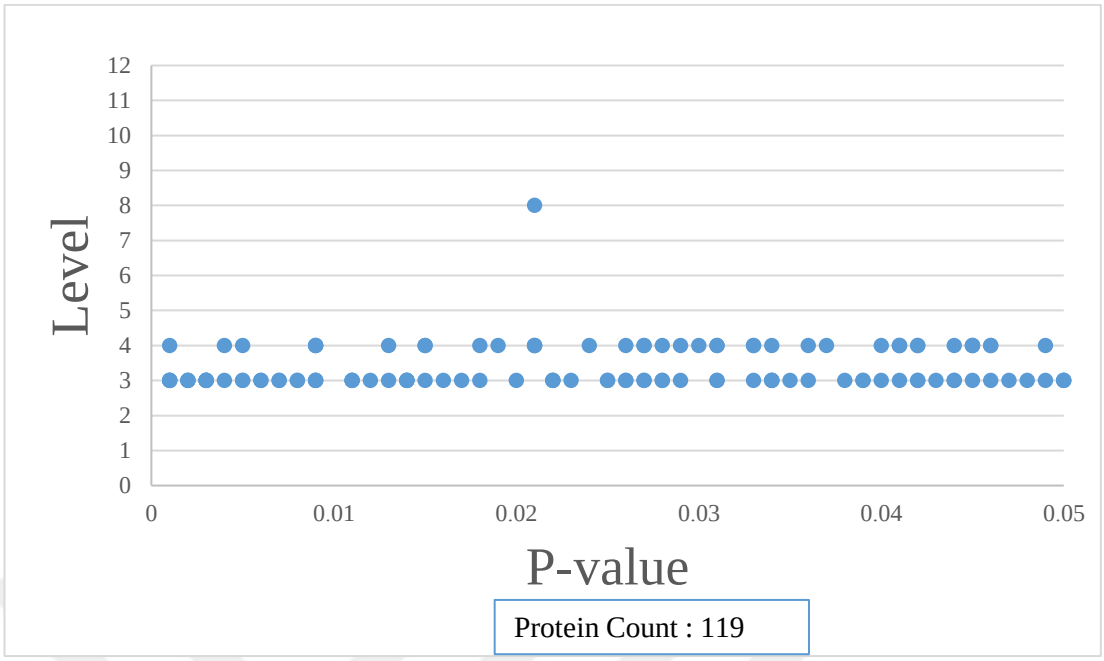


Figure 4.8 The genes whose p-value are less than 0.05 with level information for Geldanamycin drug

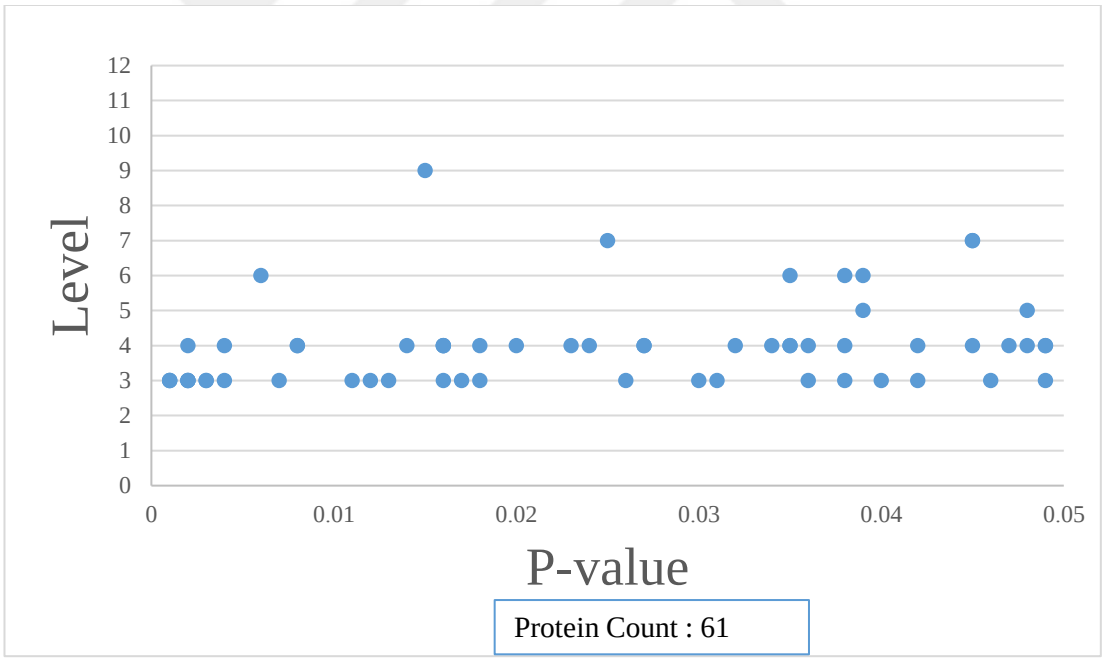


Figure 4.9 The genes whose p-value are less than 0.05 with level information for H-7 Dihydrochloride drug

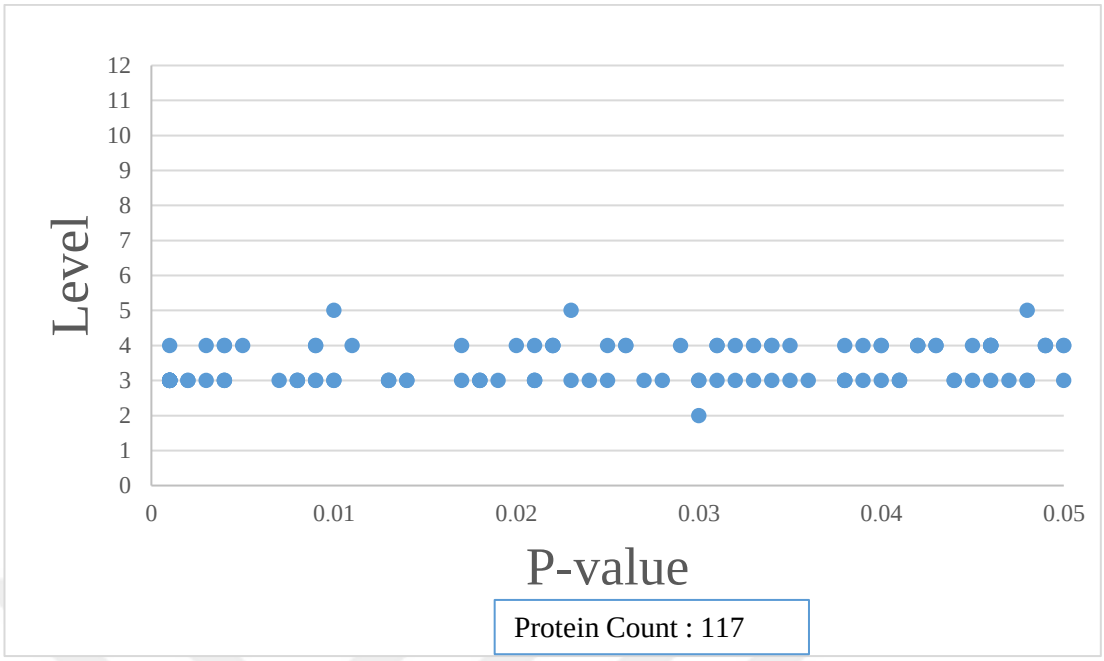


Figure 4.10 The genes whose p-value are less than 0.05 with level information for Methotrexate drug

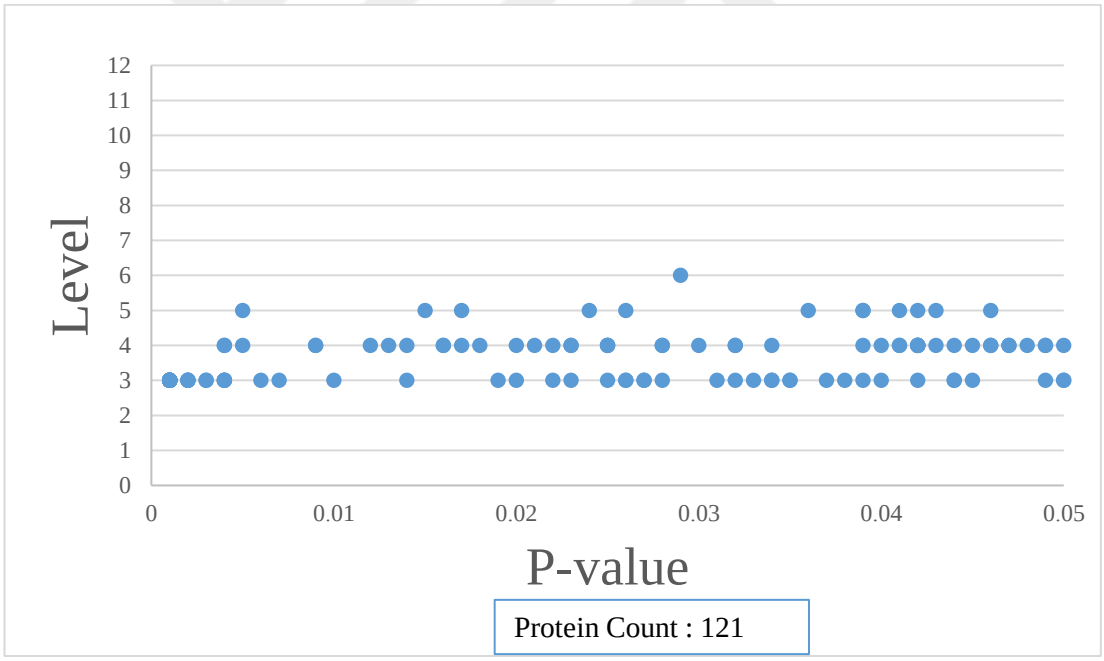


Figure 4.11 The genes whose p-value are less than 0.05 with level information for Mitomycin C drug

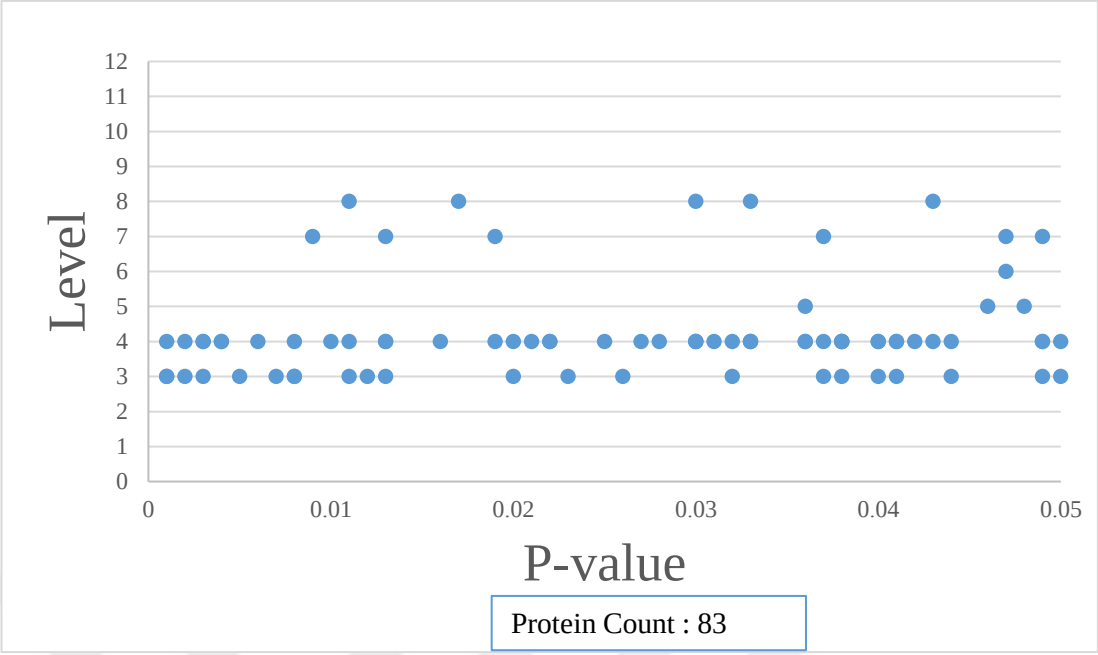


Figure 4.12 The genes whose p-value are less than 0.05 with level information for Monastrol drug

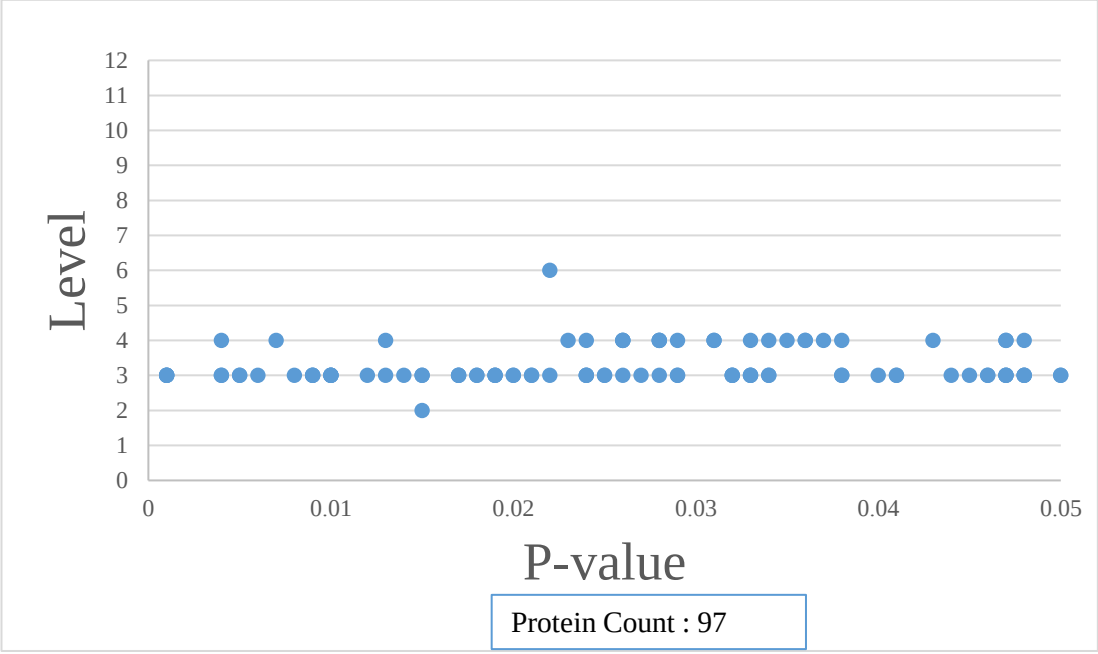


Figure 4.13 The genes whose p-value are less than 0.05 with level information for Rapamycin drug

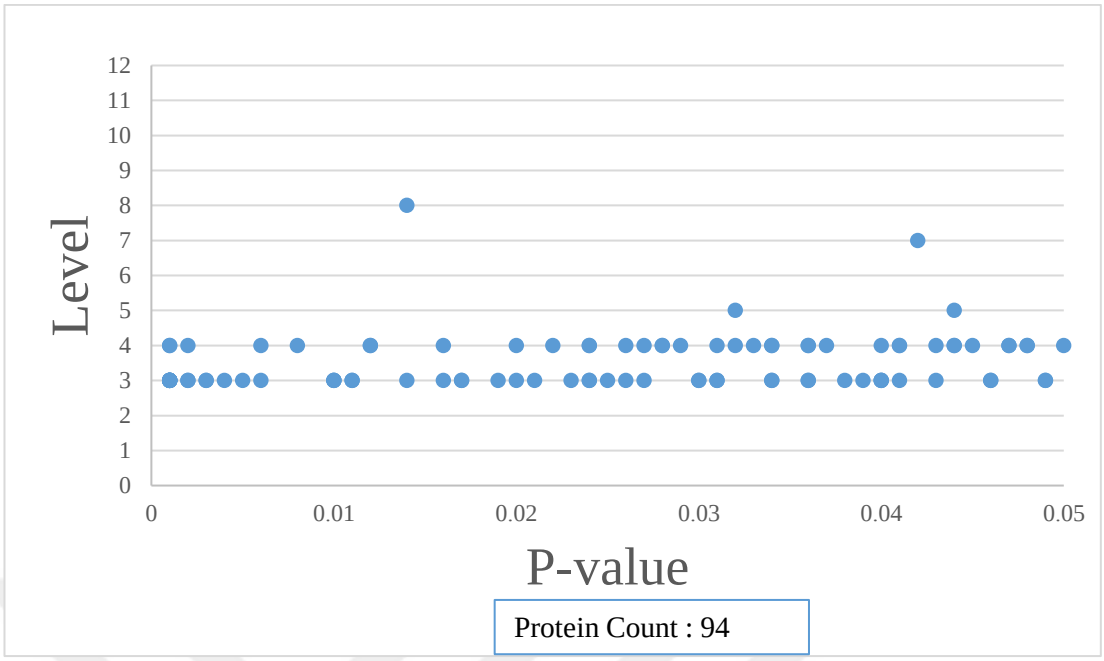


Figure 4.14 The genes whose p-value are less than 0.05 with level information for Trichostatin A drug

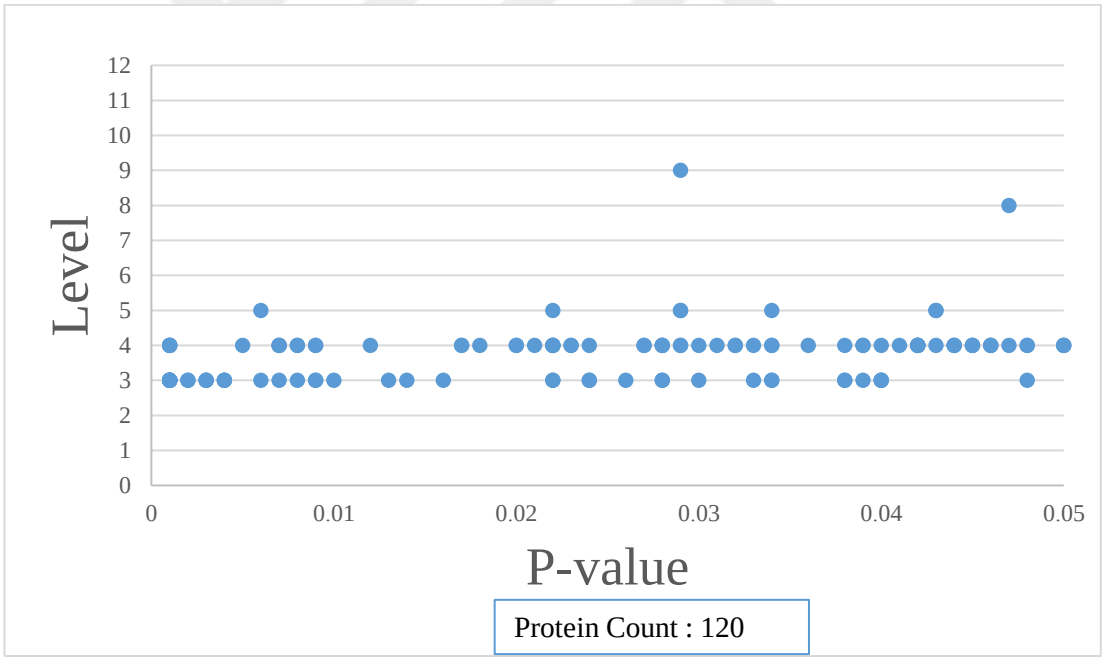


Figure 4.15 The genes whose p-value are less than 0.05 with level information for Vincristine drug

4.3 The Jaccard Index between the drugs

The Jaccard index (JI) method was used to calculate similarities between the affected proteins of two drugs. The JI shows the ratio of common genes, whose *p*-value are less than 0.05, between two drugs. The JI values of each drug pair in descending order are shown in the Table 4.2.

Most commonly observed proteins are also listed in the Table 4.3. TMEM50A protein is the most common one for all drugs based on its *p*-value. WIF1, TIFAB, and ZNF454 are the other common proteins almost found in the most of drugs. Levels of TMEM50A protein in the drug networks are showed in the Table 4.4. Levels of WIF1 protein in the drug networks are showed in the Table 4.5. Levels of TIFAB protein in the drug networks are showed in the Table 4.6. Levels of ZNF454 protein in the drug networks are showed in the Table 4.7.

We performed a literature search about the biological functions of the common proteins which monitored for several drug treatments. A prior research showed that "TMEM50A protein appears to be highly upregulated in late stage cervical cancer in comparison to normal cells" (Attwood et al., 2016). Another study also suggested that WIF1 protein regulates cancer stemness and senescence, which can lead major implications in the field of cancer biology (Ramachandran et al., 2014). Another study research showed that The Cancer Genome Atlas (TCGA) confirms that TIFAB protein is one of the genes that is deleted in nearly all reported cases of the aggressive subtypes of del(5q) myelodysplastic syndrome which is also considered a slow-growing (chronic) blood cancer (Dutt et al., 2010; Varney et al., 2015). A recent study found that ZNF454 protein leads many tumor and cancer stem in the cell organization (Sottoriva et al., 2015).

We found out that drug similarities predict the number of shared significant genes across drug pairs statistically. We performed a literature search about similarities between drug pairs with the jaccard index of the statistically significant proteins. Here, we crosschecked our results with MeSHDD tool. This tool is a framework in

calculating drug repositioning using literature-derived drug similarity. MeSHDD also uses drug cluster-based repositioning for drug similarities (Brown & Patel, 2017). You can find MeSHDD tool via http://apps.chiragjpgroup.org/MeSHDD/. According to this tool, the most similar drug to Etoposide is Doxorubicin hydrochloride; and the distance between these two drugs is given as 0.5786. When we subtract 0.5786 value from 1, we get 42.14% as a result. We found the a similar outcome with our algorithm with 42.54% result. Likewise, this tool also supports the similarity between Cycloheximide and Doxorubicin hydrochloride; Mitomycin C and Vincristine similarities. Another study also suggested that Camptothecin and Etoposide have the same reaction in tumor cells during their treatment (Sha et al., 2012). A recent study also suggested that Camptothecin and Doxorubicin hydrochloride are also conducive in fighting with breast cancer (Kathryn et al., 2011).

Table 4.2 The Jaccard Index scores of each drug pair

Drug 1	**Drug 2**	**Jaccard Index**
Doxorubicin hydrochloride	Etoposide	42.54%
Camptothecin	Doxorubicin hydrochloride	36.18%
Camptothecin	Etoposide	33.49%
Cycloheximide	Doxorubicin hydrochloride	31.15%
Mitomycin C	Vincristine	30.27%
Cycloheximide	Etoposide	28.50%
Camptothecin	Methotrexate	25.84%
Cycloheximide	Rapamycin	25.44%
Monastrol	Vincristine	25.31%
H-7 Dihydrochloride	Monastrol	25.22%
Camptothecin	Geldanamycin	25%
Doxorubicin hydrochloride	Rapamycin	24.72%
Cycloheximide	Geldanamycin	24.47%
Etoposide	Rapamycin	24.32%
Camptothecin	Cycloheximide	24.29%
Cycloheximide	Methotrexate	24.06%
Etoposide	Geldanamycin	22.93%

Table 4.2 continues

Doxorubicin hydrochloride	Methotrexate	22.84%
Methotrexate	Mitomycin C	22.68%
H-7 Dihydrochloride	Vincristine	22.30%
Geldanamycin	Rapamycin	22.03%
Doxorubicin hydrochloride	Geldanamycin	22%
Etoposide	Methotrexate	21.95%
Camptothecin	Mitomycin C	21.92%
Vincristine	Trichostatin A	19.55%
Blebbistatin	Mitomycin C	18.99%
Camptothecin	Rapamycin	18.54%
Cycloheximide	Trichostatin A	18.08%
Blebbistatin	Vincristi ne	17.61%
Mitomycin C	Trichostatin A	17.49%
Mitomycin C	Monastrol	17.24%
Methotrexate	Trichostatin A	17.22%
Geldanamycin	Trichostatin A	16.39%
Blebbistatin	H-7 Dihydrochloride	16.36%
Blebbistatin	Monastrol	16.28%
Geldanamycin	Methotrexate	16.26%
Etoposide	Trichostatin A	15.82%
Camptothecin	Trichostatin A	15.38%
Cycloheximide	Mitomycin C	15.12%
Etoposide	Mitomycin C	14.93%
Doxorubicin hydrochloride	Mitomycin C	14.42%
Trichostatin A	Aclacinomycin A	13.56%
Methotrexate	Vincristine	13.40%
Geldanamycin	Mitomycin C	13.21%
Blebbistatin	Methotrexate	12.88%
Camptothecin	Vincristine	12.71%
Methotrexate	Rapamycin	12.63%

Table 4.2 continues

Blebbistatin	Trichostatin A	12.59%
H-7 Dihydrochloride	Mitomycin C	12.35%
Doxorubicin hydrochloride	Trichostatin A	12.31%
Vincristine	Aclacinomycin A	11.89%
Cycloheximide	Monastrol	11.86%
Etoposide	Monastrol	11.34%
Doxorubicin hydrochloride	Monastrol	11.23%
Monastrol	Rapamycin	11.11%
Rapamycin	Trichostatin A	11.05%
Geldanamycin	Monastrol	10.99%
Etoposide	Vincristine	10.96%
Mitomycin C	Rapamycin	10.66%
Camptothecin	Monastrol	10.10%
Geldanamycin	Vincristine	9.63%
Blebbistatin	Geldanamycin	9.41%
Doxorubicin hydrochloride	Vincristine	9.38%
Monastrol	Trichostatin A	9.26%
Blebbistatin	Camptothecin	9.23%
Blebbistatin	Doxorubicin hydrochloride	9.09%
Blebbistatin	Cycloheximide	8.98%
H-7 Dihydrochloride	Rapamycin	8.97%
Cycloheximide	Vincristine	8.80%
Blebbistatin	Etoposide	8.70%
Methotrexate	Monastrol	8.70%
Vincristine	Rapamycin	8.50%
H-7 Dihydrochloride	Trichostatin A	7.64%
Doxorubicin hydrochloride	H-7 Dihydrochloride	7.51%
Cycloheximide	H-7 Dihydrochloride	7.32%
Blebbistatin	Rapamycin	7.19%
Etoposide	H-7 Dihydrochloride	6.59%

Table 4.2 continues

H-7 Dihydrochloride	Aclacinomycin A	6.32%
Cycloheximide	Aclacinomycin A	6.16%
Camptothecin	H-7 Dihydrochloride	6.15%
Mitomycin C	Aclacinomycin A	5.92%
Rapamycin	Aclacinomycin A	5.38%
H-7 Dihydrochloride	Methotrexate	5.33%
Blebbistatin	Aclacinomycin A	4.90%
Geldanamycin	H-7 Dihydrochloride	4.65%
Geldanamycin	Aclacinomycin A	4.61%
Methotrexate	Aclacinomycin A	3.97%
Etoposide	Aclacinomycin A	3.59%
Monastrol	Aclacinomycin A	3.36%
Camptothecin	Aclacinomycin A	2.76%
Doxorubicin hydrochloride	Aclacinomycin A	1.85%

Table 4.3 Most common proteins that were found to be significant at least more than seven drugs

Protein Name	**# of Drugs Found in Common**
TMEM50A	14
WIF1	13
TIFAB	12
ZNF454	11
IL12A	10
MS4A4A	10
IL17D	10
ELOA2	9
NRB1	9
IGDCC3	9
MXRA5	9
KLRC1	9
CD1B	8
PTGR1	8
ZNF568	8
S100B	8
SEC16B	8
SERPINB8	8
CBLC	8
FAAP20	8
TNN	8
SLC4A4	8
HSPB9	8
GPRC5B	8
ZNRF4	8
UGT2B4	8

Table 4.4 Levels of TMEM50A protein in drug networks

Drug Name	**Level in which protein is found**
Aclacinomycin A	10
Blebbistatin	9
Camptothecin	5
Cycloheximide	8
Doxorubicin hydrochloride	5
Etoposide	5
Geldanamycin	8
H-7 Dihydrochloride	9
Methotrexate	5
Mitomycin C	6
Monastrol	8
Rapamycin	6
Trichostatin A	8
Vincristine	9

Table 4.5 Levels of WIF1 protein in drug networks

Drug Name	**Level in which protein is found**
Aclacinomycin A	Not Found
Blebbistatin	4
Camptothecin	3
Cycloheximide	4
Doxorubicin hydrochloride	3
Etoposide	3
Geldanamycin	3
H-7 Dihydrochloride	4
Methotrexate	4
Mitomycin C	4
Monastrol	4
Rapamycin	3
Trichostatin A	3
Vincristine	4

Table 4.6 Levels of TIFAB protein in drug networks

Drug Name	**Level in which protein is found**
Aclacinomycin A	Not Found
Blebbistatin	4
Camptothecin	4
Cycloheximide	4
Doxorubicin hydrochloride	4
Etoposide	4
Geldanamycin	4
H-7 Dihydrochloride	Not Found
Methotrexate	4
Mitomycin C	4
Monastrol	5
Rapamycin	4
Trichostatin A	4
Vincristine	4

Table 4.7 Levels of ZNF454 protein in drug networks

Drug Name	**Level in which protein is found**
Aclacinomycin A	3
Blebbistatin	3
Camptothecin	3
Cycloheximide	3
Doxorubicin hydrochloride	3
Etoposide	3
Geldanamycin	3
H-7 Dihydrochloride	Not Found
Methotrexate	3
Mitomycin C	3
Monastrol	Not Found
Rapamycin	3
Trichostatin A	3
Vincristine	Not Found

CHAPTER FIVE
CONCLUSION AND FUTURE WORKS

More information about cell signaling is needed to better understand the analysis of microarray experiments. For this reason, it became more interesting to gain a new perspective on the signaling pathway and interaction network analysis issues in the traditional analysis of gene expression data (Cline et al., 2007; Garcia, Espinal & Hernandez, 2015; Kotelnikova, Pyatnitskiy, Paleeva, Kremenetskaya & Vinogradov, 2016). Novel methods have started to detect one of the most important proteins and cellular processes in the signaling pathways. Those results have not been discovered by traditional methods. New methods need to apply to discover such results (Isik, Baldow, Cannistraci & Schroeder, 2015; Kaushik, Ali & Gupta, 2017; Mueller, Tew, Vasieva, Clegg & Canty, 2016; Tarca et al., 2009; Zhou et al., 2016).

Biological pathways are collection units of proteins that perform a specific metabolic task cooperatively. They represent the attitude of gene groups in response to an external signal like a drug treatment (Pas, Hemert, Hulsegge, Rebel & Smits, 2008). Pathway-based analysis is a quite new idea in understanding gene expressions in large quantities in signal and metabolic levels. Cell signaling networks and gene expression profiling can help to understand the metabolic effects of drug development in cells with graphical topologies such as tree networks in the metabolic level (Zhang, Gao, Liu, Zhao & Che, 2009).

In our studies, we observed molecular effects of a score flow algorithm in lymphoma cancer cells integrated into KEGG signaling networks after 14 different drug applications. Our results have shown that certain proteins are mostly affected by these drugs. We have also observed that the vast majority of these proteins are also effective in cancer related cell responses. In addition, statistically significant proteins are generally located at the center of the signaling network, which is why the other proteins are linked to each other at a greater number of bonds at later levels.

Some proteins are also found in different medicines. COX7A2 is identified as the highly affected protein in the drug network of Aclacinomycin A, H-7 Dihydrochloride, Methotrexate, and Mitomycin C drugs. NFE2L1 is identified in the drug networks of Methotrexate and Mitomycin C. USP15 is also highly influenced protein in the drug networks of Camptothecin and Etoposide. TMEM50A protein is the most common for all drugs based on its *p*-value. WIF1, TIFAB, and ZNF454 are also common proteins practically found in most of the drugs based on their *p*-value. The most similar drug pair is Etoposide and Doxorubicin hydrochloride in the our data set. The other similar drugs are Camptothecin and Doxorubicin hydrochloride; Camptothecin and Etoposide; Cycloheximide and Doxorubicin hydrochloride; Mitomycin C and Vincristine.

The literature verifies our results obtained from the score flow algorithm that can assist to better understand the molecular responses of a cell in the signaling level after a drug treatment. With these particular experimental conditions, we showed the significance of the methods recommended in understanding of cellular responses. We have shown that a computational method can be very important in understanding of cellular responses in lymphoma cancer cells. Lastly, literature review supports the results obtained from the score flow algorithm on responses at the signaling level of a cell after drug treatment. Such methods can be crucial in finding new drug combinations at the molecular signal level.

REFERENCES

Ahnert, S., Fink, T., & Zinovyev,A.(2008).How much non-coding DNA do eukaryotes require? . *Journal of Theoretical Biology*, *252*(4):587–592, 2008.

Alberts, B., Johnson, A., Lewis, J.,Raff, M.,Roberts, K.,Walter, P. (2002*). Molecular Biology of the Cell. 4th edition*.New York: Garland Science

Attwood,M.,Arunkumar,K.,Pivotti,V.,Yazdi,S.,Almen,M.S.,Schiöth H.B.,et al. (2016) .Topology based identification and comprehensive classification of four-transmembrane helix containing proteins (4TMs) in the human genome.*Bio.Med. Central Genomics 17*:268,2592-7

Baldi,P., Hatfield,G.W.(2002).DNA microarrays and gene expression: from experiments to data analysis and modeling. *Cambridge Univ Pr, 2002.*

Bansal, M., Yang, J., Karan, C., Menden, M.P., Costello, J.C., Tang, H., et al. (2014). A community computational challenge to predict the activity of pairs of compounds. *Nat Biotechnol. 32(*12), 1213-22.

Bellaachia, A., Portnoy, D., Chen, Y., Elkahloun, A.G. et al.(2002).E-CAST: A Data Mining Algorithm for Gene Expression Data, Workshop on Data Mining in Bioinformatics, *National Institute of Health (NIH) 2*, 49-54, 2002.

Berg,J.M.,Tymoczko,J.L.,Stryer L.(2002) *Biochemistry* (5th edition) (chapter 3) New York: W H Freeman; 2002.

Biswas, M., Kwong, E. K., Park, E., Nagra, P., & Chan, J. Y. (2013). Glycogen synthase kinase 3 regulates expression of nuclear factor-erythroid-2 related transcription factor-1 (Nrf1) and inhibits pro-survival function of Nrf1. *Exp Cell Res., 319*(13), 1922-31.

Breadth First Traversal for a Graph, Retrieved October 2011 from http://www.geeksforgeeks.org/breadth-first-traversal-for-a-graph

Brown, A., Patel, C. (2017) MeSHDD: Literature-based drug-drug similarity for drug repositioning. *Am Med Inform Assoc. 2017;24*(3):614-618

Brown C.S.,Goodwin P.C,& Sorger,P.K. (2001) Image metrics in the statistical analysis of DNA microarray data. *Proceedings of the National Academy of Sciences of the United States of America : 98*(16):8944-9

Brune, B., Kenthen, A.,& Sandau,K.B. (1999). "Nitric oxide (NO): an effector of apoptosis". *Cell Death and Differentiation. 6* (10): *969–975.*

Bu, Z.,& Callaway D.J. (2011). "Proteins MOVE! Protein dynamics and long-range allostery in cell signaling". *Advances in Protein Chemistry and Structural Biology. 83*: 163–221.

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2012). *A comparative study of efficient initialization methods for the k-means clustering algorithm.Expert Systems with Applications. 40* (2013): 200–210.

Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., et al. (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res. 39*,D685–D690.

Chen-An, T., Yi-Ju, C., & James, J.C.(2003). Testing for differentially expressed genes with microarray data.*Nucleic Acids Res. 2003 May 1; 31*(9): e52.

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., et al. (2007). Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc. 2*(10), 2366-82.

Cooper, G.M. (2000) *The Cell: A Molecular Approach. 2nd edition*.Sunderland (MA): Sinauer Associates

Datta S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data, *Bioinformatics, 19*(4) , 459-466.

Defays,D.(1977).An efficient algorithm for a complete-link method.*The Computer Journal, British Computer Society. 20* (4): 364–366.

Dennis, G., Sherman, B.T., Hosack, D.A.,Yang, J., Gao, W., Lane, H.C., et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol. 2003;4*(5):P3.

Douglas,R.G. (2011). *Means to an End: Apoptosis and other Cell Death Mechanisms*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Dutt, S., Narla, A., Lin, K., Mullally, A., Abayasekara, N., Megerdichian, C.,et al.(2010).Haploinsufficiency for ribosomal protein genes causesselective activation of. p53 in human erythroid progenitor cells. *Blood. 2011 Mar 3;117(9):*2567-76.

Eichhorn, P. J., Rodon, L., Gonzalez-Junca, A., Dirac, A., Gili, M., & Martinez-Saez, E., et al. (2012). USP15 stabilizes TGF-β receptor I and promotes oncogenesis through the activation of TGF-β signaling in glioblastoma. *Nat Med., 18(3),* 429-35.

Forgy,E.W.(1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics. 21*: 768–769.

Garcia-Campos, M.A., Espinal-Enriquez, J.,& Hernandez-Lemus, E. (2015). Pathway Analysis: State of the Art. *Front Physiol. 6,* 383.

Genome Apoptosis - Kegg Pathway, March 2016 from http://www.genome.jp/kegg/pathway/hsa/hsa04210.html

Handl,J.,Knowles,J.,& Kell,D.B.(2005). Computational cluster validation in post-genomic data analysis, *Bioinformatics, 21* (2005), 3201–3212

Haykin,S.(1999). *9. Self-organizing maps.* Neural networks - A comprehensive foundation (2nd ed.). Prentice-Hall.

Huang, S.,Chong, N.,Lewis, N.E.,Jia,W.,Xie, G.,Lana, X.G.,et al.(2016).Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis.*Genome Med.* 2016; 8*:34*. 13073-016-0289-9

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., Speed, T.P.,et al. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res. 2003 Feb 15;31*(4):e15.

Isik, Z., Baldow, C., Cannistraci, C.V.,& Schroeder, M. (2015). Drug target prioritization with perturbed genes and network information, *Sci Rep. 5*, 17417.

Isik, Z., Ersahin, T., Atalay, V., Aykanat, C. & Cetin-Atalay, R.,et al. (2012). A signal transduction score flow algorithm for cyclic cellular pathway analysis, which combines transcriptome and ChIP-seq data, *Mol Biosyst, 8*, 3224-31.

Iskar, M., Zeller, G., Blattmann, P., Campillos, M., Kuhn, M., & Kaminska, K. H., et al. (2013).Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. *Mol Syst Biol., 9*, 662.

Jaccard,P. (1912) *The distribution of flora in the alpine zone.* New Phytologist, *11* (1912), 37–50

Jain, A.N.,Tokuyasu T.A.,Snijders A.M.,Segraves R., Albertson D.G.,Pinkel D., & et al.(2002).Fully automatic quantification of microarray image data. *Genome Research 12*(2):325-32.

Jelili,O.,Itunuoluw,I.,Funke, O.,Olufemi, A., Efos,U.,Faridah,A.et al.(2016). Clustering Algorithms: Their Application to Gene Expression Data. *Bioinform Biol Insights. 2016; 10*. 237–253.

Kaushik, A., Ali, S., & Gupta, D. (2017). Altered Pathway Analyzer: A gene expression dataset analysis tool for identification and prioritization of differentially regulated and network rewired pathways. *Sci Rep. 7*, 40450.

Kathryn, M. C., Sunny, K., Stefano, M., Douglas, R. V., Aaron, C. A., Samir, M. et. al. (2015) .Synergistic Antitumor Activity of Camptothecin - Doxorubicin Combinations and their Conjugates with Hyaluronic Acid. *J Control Release. 2015; 210: 198–207.*

Kerr, M.K., & Churchill, G.A. (2007). Statistical design and the analysis of gene expression microarray data. *Genet Res, 89*(5-6): 509-14.

Kotelnikova, E.A., Pyatnitskiy, M., Paleeva, A., Kremenetskaya, O., & Vinogradov, D. (2016). Practical aspects of NGS-based pathways analysis for personalized cancer science and medicine. *Oncotarget. 7*(32), 52493-516.

Kristian, A.G.,Ruth, LS.,Susan T., Mathew, W.W., Elspeth, A. B. (2016) A review of the new HGNC gene family resource *Human Genomics 201610*:6

Lander, E.,Linton, L.,Birren, B.,Nusbaum, C.,Zody,M.,Baldwin,J.,et al.(2001) *Initial sequencing and analysis of the human genome.* Nature, 409(6822):860–921

Li, J., & Ong, H.L. (2004).Feature Space Transformation for Better Understanding Biological and Medical Classifications. *Journal of Research and Practice in Information Technology, 36,* 3, August 2004

Liu, Y., & R.H. Weisberg (2011) *A review of self-organizing map applications in meteorology and oceanography.* Intech Open Science: 253-272.

Liu,W.M.,Mei,R.,Di,X.,Ryder, T.B.,Hubbell, E.,Dee,S., et al.(2002) Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics. 2002 Dec;18*(12):1593-9.

Mariani, T.J., Budhraja, V.,Mecham, B.H., Gu, C.C., Watson, M.A., Sadovsky, Y., et al. (2003). *A variable fold change threshold determines significance for expression microarrays.* FASEB J. *17* (2): 321–323.

Meelis,K.,& Jaak, V. (2008).Fast approximate hierarchical clustering using similarity heuristics . *BioData Min. 2008*; 1: 9

Mendel,G.(1865) *Experiments in plant hybridization* (R. A. Fisher , Trans.) Arkana-Verl Publishers. (Original work published 1865).

Mueller, A.J., Tew, S.R., Vasieva, O., Clegg, P.D., & Canty-Laird, E.G. (2016). A systems biology approach to defining regulatory mechanisms for cartilage and tendon cell phenotypes. *Sci Rep. 6*, 33956.6

Oh, D. H., Rigas, D., Cho, A., & Chan, J. Y. (2012). Deficiency in the nuclear-related factor erythroid2 transcription factor (Nrf1) leads to genetic instability. *FEBSJ. 279*(22), 4121-30.

Pingzhao, H., Celia, M.T.G.,& Joseph, B.Using the ratio of means as the effect size measure in combining results of microarray experiments (2009) . *BMC Syst Biol.* 2009; 3: 106.

Pray, L. (2008) *Discovery of DNA structure and function: Watson and Crick. Nature Education 1*(1):100

Pulverer,W., Noehammer,C.,Vierlinger K.,& Weinhaeusel, A., (2012).*Updates in the Understanding and Management of Thyroid Cancer ,* (Chapter 5 : Principles and. Application of Microarray Technology in Thyroid Cancer Research)

Quackenbush,J.(2001) Computational analysis of microarray data.*Nature Reviews Genetics, 2*(6) ,418-427.

Ramachandran, I., Ganapathy, V., Gillies, E., Fonseca, I., Sureban, S.M., Houchen, C.W., Reis, A.,et al.(2014).Wnt inhibitory factor 1 suppresses cancer stemness and induces cellular senescence.*Cell Death & Disease : 5:219*.e1246

Rand,W.M.(1971) *Objective criteria for the evaluation of clustering methods*, Journal of the American Statistical Association, *66* (1971), 846–850.

Rickman,D.S.,Herbert C.J.,& Aggerbeck,P.L.(2003).Optimizing Spotting Solutions for Increased Reproducibility of cDNA Microarrays.*Nucleic Acids Research:31* (18), e109

Rokach, L.,& Oded M.(2005).*Clustering methods*.Data mining and knowledge discovery handbook. Springer US. 321-352.

Rousseeuw,P.J.(1987) *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational Applied Mathematics, 20 (1987), 53–65.

Sang, Y.L., Hyun, U. K.,& Tae, Y.K.,(2011) *Method for predicting drug targets and screening for drugs for pathogenic microorganisms using essential metabolites* , Korea Advanced Institute of Science and Technology 373-1

Sculley, D. (2010). Web-scale k-means clustering. *Proceedings of the 19th international conference on World wide web.* ACM, 1177–1178.

Sha, S., Jin, H., Li, X., Yang, J., Ai, R., Lu, J.,et al.(2012) Comparison of caspase-3 activation in tumor cells upon treatment of chemotherapeutic drugs using capillary electrophoresis. *Protein Cell. 2012 May; 3* (5):392-9

Shapiro, G.P., & Tamayo, P. (2003) Microarray Data Mining: Facing the Challenges. *SIGKDD Explorations, 5, 3,* 1-5, 2003.

Sibson,R.(1973) SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal, British Computer Society*:*16* (1):30–34.

Slideshare Marketing Research An Applied Orientation , November 2012 from http://www.slideshare.net/uzairjavedsiddiqui/malhotra20

Smolkin,M.,& Debashis, G. (2003) . Cluster stability scores for microarray data in cancer studies . *BMC Bioinformatics 2003* 10.1186/1471-2105-4-36

Sottoriva,A.,Kang,H.,Ma,Z.,Graham,T.A.,Salomon,M.P.,Zhao J.,et al.(2015) A Big Bang model of human colorectal tumor growth.*Nat Genet. 2015 Mar; 47*(3): 209–216.

Storey, J.D. ,& Tibshirani, R.(2003) *Statistical significance for genomewide studies.* Proc Natl Acad Sci U S A, 2003. 100 (16): 9440-9445.

Székely, G. J.,& Rizzo, M. L. (2005). Hierarchical clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method. *Journal of Classification. 22* (2): 151–183.

Szklarczyk, D., Santos, A., von Mering, C., Jensen, L.J., Bork, P., Kuhn, M., et al. (2016). STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res. 4,* 44(D1), D380-4.

Tarca, A.L., Draghici, S., Khatri, P., Hassan, S.S., Mittal, P., Kim, J.S., et al. (2009). A novel signaling pathway impact analysis. *Bioinformatics. 25,* 75-82.

Te Pas, M.F.W., Hemert, S., Hulsegge, I., Rebel, J.M.J., Smits, M.A.,et al.(2008). Pathway Analysis Tool for Analyzing Microarray Data of Species with Low Physiological Information. *Adv Bioinformatics.* 719468.

Tusher, V.G., Tibshirani ,R.,& Chu,G.(2001)*Significance analysis of microarrays applied to the ionizing radiation response*. Proc Natl Acad Sci U S A, 2001. *98*(9): 5116-21.

Wang,X.,Ghosh, S.,& Guo,S.W.(2001). Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research 29*(15): E75-5.

Varney, M.E., Niederkorn, M., Konno, H., Matsumura,T., Gohda, J.,Yoshida, N.,et al. (2015)Loss of Tifab, a del(5q) MDS gene, alters hematopoiesis through derepression of Toll-like receptor–TRAF6 signaling.*The Journal of Experimental Medicine :212* (11): 1967

Warren, J.(2011). *Drug discovery: lessons from evolution.* Br J Clin Pharmacol. *71*(4), 497–503.

Wong K (2016) *Computational Biology and Bioinformatics: Gene Regulation.* Hong Kong, CRC Press Taylor&Francis Group

Xu,Y.,Olman V.,Zu D. (2002) Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees, *Bioinformatics, 18*(4), 536-545.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., Speed, T.P.,et al. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.*Nucleic Acids Res. 2002 Feb 15;30*(4):e15.

Yin,L.,Huang, C.H.,&Ni J.(2006).Clustering of gene expression data: performance and similarity analysis.*BMC Bioinformatics. 2006; 7*(Suppl 4): S19.

Yin,T.,,Jianrong, L., Caro,F.,Yves,& A. L. (2007) .Information theory applied to the sparse gene ontology annotation network to predict novel gene function.*Bioinformatics. 2007 Jul 1; 23*(13): i529–i538.

Zhang, D.Y., Ye, F., Gao, L., Liu, X., Zhao, X., Che, Y., et al. (2009). Proteomics, pathway array and signaling network-based medicine in cancer, *Cell Div. 4*: 20.

Zhou, J.X., Isik, Z., Xiao, C., Rubin, I., Kauffman, S.A., Schroeder, M. et al. (2016). Systematic drug perturbations on cancer cells reveal diverse exit paths from proliferative state. *Oncotarget. 7*(7), 7415–25.

Zien, A.,Schoelkopf, B. ,Tsuda, K. ,Vert. J.P. (2004) *A primer on molecular biology. Kernel methods in computational biology*, 3, 2004.MIT Press.