2013-07-29

Network Driven Bio-Data Integration and Mining for Bio-Medical Predictions

Qabaja, Ala

UNIVERSITY OF CALGARY

Network Driven Bio-Data Integration and Mining for Bio-Medical Predictions

By

Ala Qabaja

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILMENT OF TH REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

CALGARY, ALBERTA

July, 2013

Abstract

Drug repositioning is increasingly attracting much attention from pharmaceutical community to tackle the problem of long term development in drug discovery. The complex nature of human diseases, for example cancer, poses major challenges in pharmaceutical industry nowadays. With the increasing amount of research conducted to understand associations between drugs and diseases, a new direction of research has come to light. Thanks to the development of high-throughput technologies to generate tremendous amount of data and to the web-based systems to store and organize the generated data, drug repositioning has become cost and time effective.

Since biomolecular interactions and omics-data integration has had success in drug development, we have been motivated to develop a new paradigm that integrates data from three major sources to predict novel therapeutic drug indications. Microarray data, biomedical text mining data and biomolecular interactions are all integrated to predict ranked lists of genes based on their relevance to a particular drug's or disease's molecular action. These ranked lists of genes are used as raw input for building a disease-drug connectivity map based on enrichment statistical measure. This integrative paradigm was able to report a sensitivity improvement of 18% and 26% in comparison with using text-mining and microarray data, respectively, independently. In addition, this paradigm was able to predict many clinically validated disease-drug associations that could not be captured with using microarray or text mining data independently.

The robustness of the integrative paradigm has been further investigated to predict functional miRNA-disease associations. In here, disease-gene associations from microarray experiments and text mining together with miRNA-gene associations from computational predictions and protein networks have been integrated to build miRNA-disease associations. The findings of the proposed model were validated against gold standard datasets using ROC analysis and results were promising (AUC = 0.81). The proposed integrated approach allowed us to reconstruct functional associations between miRNAs and human diseases and to uncover functional roles of newly discovered miRNAs

Acknowledgements

I would like first to express my sincere gratitude and thanks to the best friend and the best supervisor Prof. Reda Alhajj. I would like to thank you for introducing me to the field of data mining, databases and bioinformatics. Without your academic and financial support, I wouldn't have been able to finish this thesis.

I would like to thank Dr. Moussavi and Dr. Rokne for spending time reviewing this thesis.

I would like to thank Dr. Lisa Hiaghm for teaching me the concepts of algorithms and their design and analysis, Dr. Moussavi for teaching me the concept of software requirements gathering and engineering, Dr. Gordon Chua for teaching me the applications of computer algorithms on real biological problems. Without knowing these concepts, it would have been really hard for me to build this system.

I would like to thank all great friends who had their fingerprints helping me to build this manuscript. Also I would like to thank those who were caring about me and about my family. Thanks to Ali Rahmani, Atieh Sarraf, Fatemeh Keshavarsh, Shermin Bazazian, Thaer Shunnar, Omar Zarour, Omar Addam, Mohammed Shalalfeh, Abdullah Al-Shiekh, Ala Hamdan, Sara Aghakhani, Sarah Fakhoury, Derar Al-Asi, Abdelrahman Qiswani, Ala Kassab, Maisa Barbarawi, Nour El-Masri and Tamer Jarada.

Special thanks to Dr. Jalal Kawash for his kind assistance

Finally my deepest love and thanks go to my family. Thanks to my greatest father, (Mohammed), most amazing mother (Mazuza), sisters (Nadin, Suzan, Bayan and Doa) and brothers (Abood and Mahmoud).

To my first and only love, Suha Janajreh

To my newborn boy, Jad Qabaja

My whole life is nothing without you

Table of Contents

List of tables

List of figures

Abbreviations

DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
mRNA	Messenger Ribonucleic Acid
cDNA	Complementary Deoxyribonucleic Acid
SVD	Singular Value Decomposition
UFF	Unsupervised Feature Filtering
PCA	Principle Component Analysis
SAM	Significant Analysis of Microarrays
FDR	False Discovery Rate
ULR	Univariate Linear Regression
MLR	Multivariate Regression Model
TM	Text-Mining
GSEA	Gene Set Enrichment Analysis
RGES	Randomly Generated Enrichment Score
ROC	Receiver Operator Characteristic

Chapter 1: The Need for Bio-Data Integration and Mining

Many traditional methods have been proposed in order to diagnose and treat diseases. Despite of all researches and methods proposed in order to optimize their treatment, Cancer and other incurable diseases in all of their forms are worldwide and increasing rapidly. Mainly there are five standard methods for cancer and other incurable diseases treatment; surgery, chemotherapy, radiation therapy, immunotherapy and biologic therapy. Even though new concepts, new methods, and new approaches have been introduced for treatment, the mainstay of incurable diseases chemotherapy remains with systemic anti-causative agents that interfere with cellular DNA functionality. Therefore, it was a crucial step to study and analyze the gene-gene, gene-drug and gene-disease interactions in order to better understand the toxic effect of a drug on cancer cells and analyze the ability of that drug to treat diseases. The introduction of new technologies and new computational algorithms combined with our understanding of human genome has dramatically and superiorly improved the research in human diseases especially cancer.

1.1. Introduction

Many years ago the successfulness of a specific drug was based on its ability to generate desired changes in the physiological states of animals without paying any attention to the biochemical reactions this drug might introduce. Similarly, diseases were all diagnosed by looking at specific symptoms that disease might trigger. Later on, the development of many methods that enable the isolation and the study of individual cells and molecules have shifted the understanding of drugs/diseases from being at physiological level to more accurate molecular level.

Recently and after the introduction of the genome sequencing project that provides complete list of genes and gene's products, drug discovery and disease diagnosis processes have been revolutionized. More specifically, this project has enabled a better understanding for drugs and diseases mode of actions by discovering the genes or proteins that play a major in their molecular action. These genes have become tempting targets for many pharmaceutical companies. In addition, this tempting area of science has become one of the most well studied and interesting research areas in many labs worldwide.

1.2. Problem Definition and Motivation

The rapidly evolving researches in the biomedical fields have made a huge amount of biological data hidden in the web in many formats that can be classified into three major categories: (i) Microarray gene expression profiles. (ii) Text published papers. (iii) Gene-Gene, Gene-Disease, and Gene-Drug interaction databases.

The availability of such rich amount of biological data has shed the light on utilizing computer science techniques and algorithms to mine data and infer biomedical knowledge. High-throughput microarray technology, biological databases and text-mining data integration process is a very powerful technique that can provide a closer insight to many biological problems including drug repositioning or predicting disease-miRNA interactions.

Drug repositioning (also known as drug re-profiling or re-tasking) can be simply defined as a technique that seeks investigation of new therapeutic applications for already approved drugs or drug candidates that have not succeeded in advanced clinical trials for reasons other than safety [1]. The process of drug repositioning offers several advantages over the traditional drug development including; reduced development costs and shorter paths to approval [2]. The costly and laborious traditional paradigm takes about 15 years and almost $1 billion to test, to validate and to launch a new drug to the market [3].

Accurate prediction for Disease-miRNA interaction is also very vital in medical research. MiRNAs are new key players in the disease paradigm demonstrating roles in several human diseases. The functional association between miRNAs and diseases remains largely unclear and far from complete. With the advent of high-throughput functional genomics techniques, it is now possible to infer functional association between diseases and biological molecules by integrating disparate biological information.

Out of the different methods proposed to integrate biological data, two are known to be most popularly used in this field; **(i)** build a list of differentially expressed gene using microarray data and then use the text-mining techniques to prioritize this list of genes in relevance to a particular disease **(ii)** build a set of relationships between different biological entities (genes, diseases or drugs) using text-mining techniques and validate these associations by resorting to microarray data. According to our knowledge, there is was not any method that has been proposed to integrate the ranked list of genes obtained from high-throughput technology with the ranked list of genes obtained from text-mining for the purpose of knowledge discovery. More specifically there was not any method that was able to provide a paradigm that integrates text-mining, microarray data and biological networks for the purpose of drug or disease understanding. Therefore, I have been motivated to build a framework that integrates these three data sources into a single paradigm for the hope of getting more accurate results than using any of them independently.

1.3. Contribution

In this thesis I tackle the great demand in integrating biological data from different sources to elicit better knowledge regarding drug discovery. I have two major contributions in this thesis by providing two different integrative paradigms for drug-repositioning and disease-miRNA prediction. These two paradigms utilize the power of text-mining methods in discovering hidden or indirect relationships, the power of microarrays in providing a global view of drugs/diseases molecular effects and the power of gene network in understanding the functional and behavioral correlation between genes.

For drug repositioning, an unsupervised statistical procedure (Gene Set Enrichment Analysis) was employed to predict drug-disease associations. Unlike other methods that always use this technique on microarray data, the proposed framework integrates data from three sources namely; microarray, text-mining and biological networks and employed this technique on the newly generated data. In brief, a ranked list of genes for every single drug and every single disease was built by using microarray expression data. Later on, another ranked list of genes for every single drug and every single disease was built by using text-mining together with gene network biology. Noting that text-mining based ranked list of genes has never been done before. Finally the ranked lists for each entity have been integrated into one and representative ranked list that have been used to build the drug-disease connectivity map using enrichment analysis statistical measure. A flow diagram for the proposed methodology is provided in **chapter 3, Figure 3.2**.

For disease-miRNA interactions prediction, two disease-gene interaction networks and a miRNA-gene interaction network were combined and used as input to a regularized logistic regression model. The disease-gene interaction networks have been built utilizing both microarray and text-mining data. The miRNA-gene interaction network has been built using microarray data and biological network. Therefore three sources of biological data were integrated into a single paradigm to predict disease-miRNA interactions. A flow diagram for the proposed methodology is provided in **chapter 7, Figure 7.1**.

The two integrative approaches mentioned here are novel and are implemented for the first time in computational biology field. Results from both approaches showed that integrating data from different sources can improve the accuracy of prediction.

1.4. Organization of the Thesis

Since this thesis targets researchers with computer science and biology background, it has been designed in a way to be self-contained and easy to understand without any need for a strong background in the computational biology field. Therefore, the rest of this thesis is organized as follows.

In chapter two I discuss in details all the necessary biological terms and concepts that have been used in this thesis. Furthermore, I discuss in details and provide the reasoning for using all computational algorithms implemented in this thesis. More precisely, I discuss all data mining tools and techniques that have been used for biological predictions namely; differentially expressed genes, clustering, classification, regression modeling, text mining and gene set enrichment measure.

Since the main focus of this thesis is on drug repositioning, I discuss all methods that have been used for drug repositioning in **chapter three**. These methods can be categorized into structure based, microarray based, text-mining based and integration based approaches.

In chapter four, I discuss the data collection procedure and sources, data normalization algorithms and data preprocessing steps.

In chapter five, I discuss all statistical measures and techniques that have been used in order to build drug-disease connectivity map. More precisely, I discuss the process of generating ranked list of genes both from microarray and text-mining and the process of utilizing gene set enrichment analysis to obtain drug-disease connectivity map.

In **chapter six**, I discuss in details the results that were obtained from drug repositioning approach. In addition, I provide a detailed biological analysis for some of the predicted associations. Finally I provide some limitations of my work and potential for future work direction.

In **chapter seven**, I explain in details the procedure for disease-miRNA interaction prediction. More precisely, I provide a descriptive introduction for disease-miRNA problem, and then follow by describing the methodology in details. Finally I describe the obtained results with computational and biological discussion.

In **chapter eight**, I provide a conclusion and future remarks for the integrative paradigm proposed in this thesis.

Chapter 2: Computational Biology, a Tempting Field for Green Drug Development

2.1. Biological Introduction

Although there are over 30 million types of organisms ranging in complexity from bacteria to human, they all use the same basic materials and mechanisms to be able to survive and adapt on this planet. More specifically, all living organisms on this planet are composed of one or more cells that compose the basic unit of structure and function in an organism. In order for a cell to function properly, the organelles within each cell have to collaborate, communicate and interact within very complex processes. This complicated system is basically driven genetically by three major macromolecules deoxyribonucleic acid, ribonucleic acid and protein.

2.1.1. Deoxyribonucleic Acid (DNA)

DNA is the basic and the most important genetic material within each cell in all organisms. These very long chains of codes are stored in an organelle called nucleus. Four basic units or nucleotides are combined together in a very long and repeated sequence to build out the DNA molecules and these are; Adenine, Guanine, Cytosine and Thymine. DNA is double-stranded by which, each nucleotide must be bound to a complementary nucleotide to formulate what is called base pairs. The length of DNA strand varies from one organism to another. For example, there are 3 billion base pairs in the human genome and 12 million base pairs in yeast. Noteworthy that a genome consists of the entire set of chromosomes that in particular compromises a series of DNA molecules. Each of these DNA molecules contains many genes [4]. In details, the gene is the hereditary unit found on a chromosome where a chromosome is a linear DNA molecule. A genome is a term that is used to represent all genes regarding a particular organism. In other words, biological information contained in a genome is encoded by its DNA and is divided into discrete units called genes [4].

2.1.2. Ribonucleic Acid (RNA)

RNA is very similar to DNA in that, it is a chain of nucleotides with a particular direction. However, it is typically found as single-stranded molecules and replaces uracil with thymine nucleotide. The most important RNA molecule is the messenger RNA (mRNA) that is created from genes that code for proteins during transcription. This mRNA is used to carry the genetic information encoded in the DNA to the ribosome, the protein assembly machinery. The mRNA is then used by the ribosome as a template to synthesize the protein that is encoded by the gene.

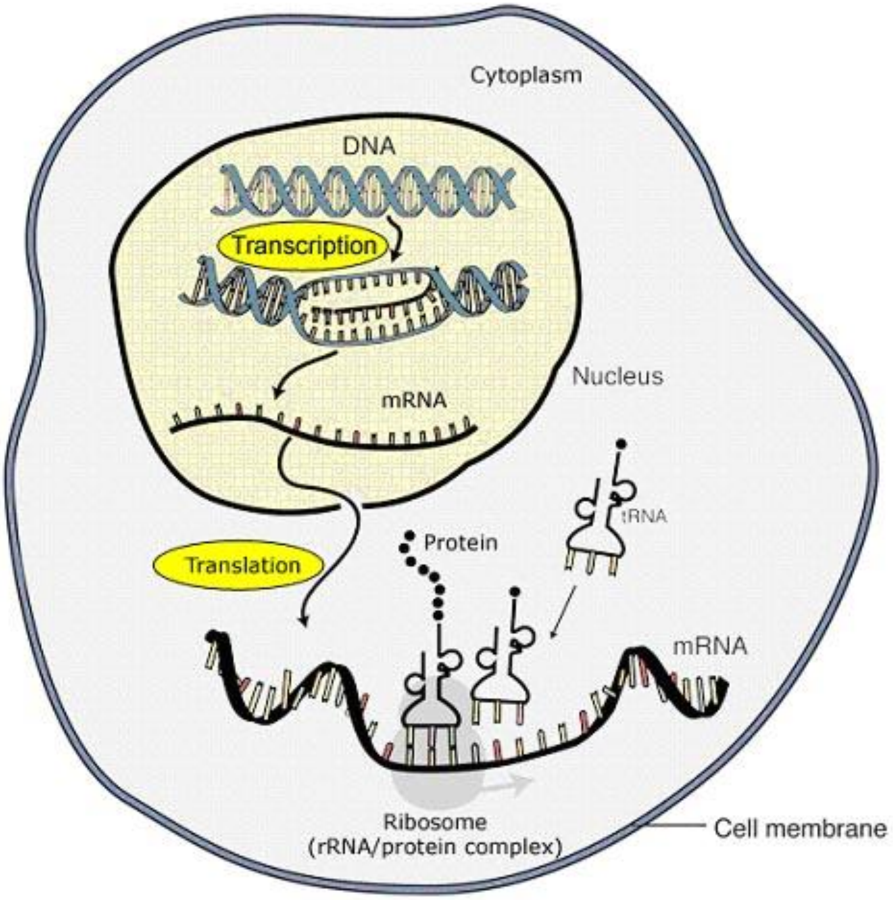
2.1.3. Protein

Proteins are the basic building blocks of nearly all molecular machinery of an organism. Proteins are made up from long chains of 20 distinct amino acids. It is the duty of genes in the DNA to specify the order of amino acids in a particular protein and thus defining the protein shape and function.

2.1.4. Genetic Circuits

Genes encoded in DNA are used as templates to synthesize mRNA through the process of transcription. These mRNA molecules work as templates that carry necessary information for the purpose of protein synthesis in a process called translation as shown in **Figure 2.1**. These genes and proteins are interacting with each other in a complex network that precisely controls the amount of production of a gene product (protein) and it can also modify the product after it is made. In fact, genes include not only coding sequences that specify structure and function of proteins, but also regulatory sequences that control the rate of transcription for that gene. These regulatory sequences are very sensitive to a protein category called transcription factors (TF). If any of these TF binds to the regulatory sequences a consequent action like activation or repression of transcription process might result in. This is because TF can either attract or rebel DNA polymerase II that is in charge of activating the transcription process.

Figure 2.1 an overview of transcription and translation



This figure shows the whole process of protein manufacturing starting from DNA to ribosome. The process starts in the nucleus when DNA strand works as a template to build mRNA molecule in a process called transcription. The mRNA molecule will move out the nucleus to cytoplasm where it binds to protein assembly machinery called the ribosome. In here, mRNA will work as a template that carries instructions to guide ribosome in building proteins with the right amino acids ordering in a process called translation. **Adopted from National Human Genome Research Institute.**

severe to moderate changes in one or even hundreds of genes might result in from changes in any of the following: (i) DNA chemical structure or TF binding sites (ii) external factors that might affect the activity of TF via activation or inhibition such as drugs, chemicals, temperature and light [5] (iii) or mutations that might lead to change in DNA or TF chemical structure or other proteins manufacturing process. This is due to the fact that genes are regulating each other in a complex network; thereby a small change in a single gene expression might result in systematic change in the expression of many other genes.

2.2. Microarrays Data Mining

For the purpose of understanding the whole molecular effect of a drug or a disease it was necessary to take a global view of biological processes that require simultaneous monitoring for cellular gene expression. DNA microarrays technology provides a simple and natural yet systematic and comprehensive vehicle for exploring the genome. Indeed simultaneous examination of thousands of proteins and genes in a single experiment has led to a renewed interest in discovering novel biomarkers for cancer [6]. Adding to its capability of performing parallel analyses as opposed to serial analysis run by older technologies, these technologies provided the ability to discover cancer biomarkers with all of its forms DNA, mRNA, and proteins. Microarray technology is carried out by hybridizing complementary DNA (cDNA) to an immobilized DNA template. DNA microarrays can be used to measure changes in expression levels between control and sample groups. In other words, it can measure the level of expression for every single gene between two biological states leaving a big room for medical, statistical, computational and mathematical analysis. Microarray technology has been widely used to provide insights into cancer classification including leukemia [7]**,** breast cancer [8] and colon cancer [9]. In addition, it has been also used to study drug's mechanism of action on many diseases including cancer cell lines [10] [2]**.**

Even though microarray expression data has a great importance in class discovery and knowledge retrieval, it is considered as one of the really huge and highly dimensionalized data. Even the expression data from a single microarray experiment requires computational tools for analysis. A common task for analyzing microarray data is to determine which genes are differentially expressed across two tissue samples or samples obtained under two experimental conditions (drug treated versus control or disease sample versus healthy sample). Two color microarrays are typically with cDNA prepared from the two samples to be compared, whereas one color microarrays provide intensity data for each probe indicating relative levels of hybridization with labeled target. Noteworthy that the pre-processed microarray data is often noisy and not normally distributed [11]**.** Therefore, several statistical and data mining methods have been proposed to normalize, analyze and infer knowledge from it.

2.2.1. Determining Differentially Expressed Genes

Determining differentially expressed genes is one of the most important statistical measures for microarray data analysis. Differentially expressed genes are a set of genes that are known to have a significant shift in their expression when comparing samples from two different biological states. These set of genes can give better insight into the molecular changes involved in tumor progression or even drug mode of action.

Many statistical, machine learning and data mining methods have been proposed to identify the set of differentially expressed genes. Data mining methods are mostly dependent on defining the entropy level for every single gene and then extract the most informative set of genes. For example Singular Value Decomposition (SVD) [12], Unsupervised Feature Filtering (UFF) [13] and Principle Component Analysis (PCA) have been all used to extract the most informative genes out of a set of genes. Statistical method tends to define a differential expression score for every single gene individually. Since in this thesis, the purpose of finding differentially expressed genes was to define similarity between biological entities based on ranked lists of genes, a statistical method that can assign a score for every single gene was used. This assigned score describes the differential expression between two biological states.

2.2.1.1. Singular Value Decomposition

Let X denote microarray expression profiles matrix of size N x M such that N represents the number of genes and M represents the number of samples. So the x_{ij} entry represents the expression level of the i_{th} gene in the j_{th} sample. Therefore, each gene g_i would be represented by an M-dimensional vector and each sample s_j would be represented by an N-*dimensional* vector. The equation for singular value decomposition would be given as $X=USV^T$ such that U is an M x M matrix, S is an N x N diagonal matrix and V^T is an N x N matrix. The columns of U are called the left singular vectors and form an orthonormal basis for sample expression profiles. The rows of V^T contain elements of the right singular vectors and form an orthonormal basis for the gene transcriptional responses. Finally the elements of S are nonzero on the diagonal and are called the singular values. The order of singular vectors is determined by high-to-low sorting of singular values, with the highest singular values in the upper left index of the S matrix. SVD is usually computed by first computing V^T and S according to the following equation $X^TX=VS^TSV^T$ and computing U according to the following equation XV=US.

The diagonal entries of S are s_1 ………. s_n such that s_i represents the singular value for gene g_i. According to Alter et al. [14] the relative importance p_i for gene g_i is computed according the following:

$$pi = \frac{si^2}{\sum_{j=1:j\neq i}^{n} sj^2} \qquad \textbf{Equation 2.1}$$

And the Shanon entropy of the data for the matrix X_{MxN} is calculated as:

$$E(X_{MxN}) = -\frac{1}{\log N}\sum_{i=1}^{N} pi \log pi \qquad \textbf{Equation 2.2}$$

Finally the proposed method of Varshavsky et al. [13] suggested defining the contribution of every single gene using the leave-one-out strategy. Therefore, the contribution of gene g_i, $Cont_i$, is computed according the following equation:

$$Cont_i = E(X_{MxN}) - E(X_{Mx(N-1)}) \qquad \textbf{Equation 2.3}$$

Such that $X_{Mx(N-1)}$ is the expression matrix excluding the expression vector of gene g_i. Thus assume having the *Cont* for all genes. Let *c* be the average of all these scores and let *d* be the standard deviation for these scores. Then the level of contribution of gene g_i is decided upon according to the following:

1- If $Cont_i > c+d$, then gene g_i has a high contribution.

2- If $c+d > Cont_i > c-d$, then gene g_i has an average contribution.

3- If $Cont_i<c-d$, then gene g_i has a low contribution.

2.2.1.2. Statistical Methods

Many statistical measures have been proposed for the purpose of inferring the most concise and informative set of genes. The output from these measures is a score that is assigned for every single gene indicating its relative change, or differential expression, between two states. The more positive is the score, the more likelihood that the corresponding gene has been overexpressed in comparison with the control sample. The more negative is the score, the more likelihood that the corresponding gene has been down-regulated in comparison with the control sample. Some researchers consider taking a cut-off for the set of over-expressed and down-regulated genes for further analysis [15] [10] [2]. This set of overexpressed together with the set of down-regulated genes are called the signature for that particular drug or disease. On the other hand, researchers might consider ranking genes from the most positively expressed to the most negatively expressed in relevant to a particular biological entity and run rank based comparison analysis between these entities [10] [15].

Statistical methods for inferring the set of differentially expressed genes range in their complexity from just considering the fold change to more complex methods that consider different parameters for the purpose of improving accuracy. For instance, **DeRisi J et al.** [16] Identified differentially expressed genes using 3-fold for log ratio of expression levels. Later on, it has been shown by [17] that considering log ratio cut-off is not sufficient to extract the informative genes. This is because in microarray experiments most genes will show expression ratio close to 1 thus it is considered to be un-appropriate.

Other statistical approaches have been proposed to model some distributional properties of gene expression. **Long AD et al.** [18] used the traditional T-test based on Bayesian estimate of variance among replicates with normally distributed expression measurement. **Dudoit S et al.** [19] used the non-parametric t-test with an adjusted p-value. Finally and the most popularly used method, Significant Analysis of Microarray (SAM), has been proposed [20].

SAM uses the concept of permutation of repeated measurement to estimate the False Discovery Rate (FDR). The major advantage of SAM over other methods is that it added a positive constant to minimize the coefficient of variation and thus alleviating the problem of giving low variances among replicates with low expressed genes. That is because low variances will result in high differentially expressed value in case of using traditional t-test. SAM assigns a score that is based on changes related to standard deviation of repeated measurements for that gene. This score represents the importance of expression change for that gene related to the sample of interest. High positive score indicates that the gene has been up-regulated after treatment with that drug or upon comparison between healthy and diseased sample. High negative score indicates that the gene has been down-regulated after treatment with that drug or upon comparison between healthy and diseased sample. An overview for these and other statistical measurements can be found in [21].

After obtaining the ranked list of genes many computational algorithms can be used to further analyze and understand the molecular nature of a drug or a disease including clustering, classification, gene-set enrichment analysis and regression modeling.

2.2.2. Clustering

Clustering can be simply defined as the process of grouping data objects belonging to the same class into one cluster and separate them from other data objects belonging to different classes. Clustering analysis has been widely used in many applications such as biology, business intelligence, image pattern recognition, web based systems …etc. Cluster analysis is a very valuable mining tool that enables understanding data distribution, data characteristics and considering set of clusters for further analysis. Furthermore, clustering serves as a preprocessing technique for other data mining tools, such as feature selection and classification [22]. This is because a cluster is a collection of data objects that are similar to one another within the cluster and dissimilar to objects in other clusters, thus a cluster can be considered as an implicit class for these objects [22]. Generally speaking, clustering methods can be classified into four major categories; partitioning methods, hierarchical methods, density based methods and Grid based methods. Out of these, k-mean clustering and hierarchical clustering methods are the most popularly used in the field of bioinformatics [23]. Therefore, in the following section I discuss proximity measurements to define similarities between objects, k-mean clustering, and hierarchical clustering. Finally I shed the light on the reason of choosing hierarchical clustering over k-mean clustering in this study.

2.2.2.1. Proximity Measurements for Data Objects

Since all data objects can be represented by numerical vectors, proximity measurement measures the distance between data objects to be clustered. Thus most clustering algorithms start by finding distances between objects using a proximity function implemented on the corresponding vectors. Assume having two data objects *X* (*x1……xp*) and *Y* (*y1……yp*) where *p* represents the number of features, then the distance is measured on the corresponding vectors $\vec{X}$ $\vec{Y}$ according to one of the following distance measures; Euclidean distance, Manhattan distance, Mahalanobis distance, Pearson correlation or any other measure.

Euclidean distance tends to measure the difference between two objects rather than measuring the actual patterns between different objects. Thus it is preferable to standardize each vector by zero mean and variance before starting the clustering process. The Euclidean distance between $\vec{X}$ $\vec{Y}$ can be measured according to the following formula:

$$Euclidean(\overrightarrow{X}\overrightarrow{Y}) = \sqrt{\sum_{i=1}^{p}(xi - yi)^2} \qquad \textbf{Equation 2.4}$$

To address the issue of measuring the similarities between the shapes of two expression patterns, Pearson correlation has been proposed. Pearson correlation between $\vec{X}$ $\vec{Y}$ can be computed according to the following formula:

$$Pearson(\overrightarrow{X}\overrightarrow{Y}) = \frac{\sum_{i=1}^{p}(xi - \overline{X})(yi - \overline{Y})}{\sqrt{\sum_{i=1}^{p}(xi - \overline{X})^2}\sqrt{\sum_{i=1}^{p}(yi - \overline{Y})^2}} \qquad \textbf{Equation 2.5}$$

Pearson’s correlation coefficient has been widely used and has proven to be an effective similarity measure for various bioinformatics researches [24] [25]. But it has been shown later that Pearson correlation might be biased towards some fake similarities and yield false positives [26] . More specifically, if two patterns have a common peak at a single feature, then the correlation might be dominated by this feature, although the patterns at the remaining features are dissimilar. This has shed the light to an improved measure called Jackknife correlation [26]. Jakknife correlation considers measuring the Pearson correlation in a leave-one-out fashion, leaves one feature out each time, then takes the minimum among those to be Jakknife similarity. Mathematically, Jakknife correlation between $\vec{X}$ $\vec{Y}$ is computed according to the following formula:

$$Jakknife(\vec{X}\vec{Y}) = \min_{i:i=1-p}(Pearson(\vec{X}\vec{Y})^{i}) \qquad \textbf{Equation 2.6}$$

Where $Pearson(\vec{X}\vec{Y})^{i}$ is Pearson correlation between object $\vec{X}$ and $\vec{Y}$ excluding *feature i.* Jakknife similarity measure was used in this thesis to run clustering algorithm.

2.2.2.2. K-Means Clustering

Given a dataset *D* with *n* objects, *k*-mean clustering needs a predefined number of clusters, *k*, to form a partitioning algorithm that splits objects into *k* <=*n* partitions. The output of this algorithm will be a set of clusters *C1,……., Ck* such that *Ci* ⊆ *D* and *Ci* ∩ *Cj*=Φ for (1<= *i, j* <=*k*). Noteworthy that the major objective of k-mean algorithm is to increase the intra-cluster similarity (minimize distance between objects within the same cluster) and decrease inter-cluster similarity (maximize distance between objects in different clusters). K-mean clustering algorithm steps are described in the following formulation:

Input:

k: the number of clusters, *D:* dataset containing *n* objects

Output:

set of clusters *C1, ……., Ck* such that $Ci \subseteq D$ and $Ci \cap Cj = \Phi$ for (1<= *i, j* <=*k*)

Algorithm:

1- Start by heuristically picking k different centroids where each centroid would be considered as a specific cluster representative.
2- Compute the distance between objects in D and all the selected centroids.
3- Assign each object to the closest centroid.
4- Re-compute the new centroid by taking the mean of all objects belonging to the same cluster.
5- Repeat step 2 through 4 until no change happens in the centroid values.

Finally the quality of cluster *Ci* can be measured by the within cluster variation, that is the sum of squared difference between centroid *ci* and all other objects belonging to cluster *Ci*. Thus the total error for a specific run *A* for a clustering algorithm would be computed according to the following formula:

$$Error(A) = \sum_{i=1}^{k} \sum_{j=1}^{b} dist(oj, ci)^2$$ **Equation 2.7**

Where *b* represents the number of objects in cluster *Ci* and *ci* represents the centroid of cluster *Ci*.

2.2.2.3. Hierarchical Clustering

In contrast to k-mean clustering, hierarchical clustering partitions data into groups at different levels leading to hierarchy or a tree. This organization is very useful when clustering is used for data summarization and visualization. Generally speaking, hierarchical clustering method can be either in bottom-up or top-down fashion and therefore it can be categorized as agglomerative or divisive approach, respectively. In agglomerative hierarchical clustering, every single object would be represented by a cluster at the beginning. After computing a distance matrix between these clusters, the single-object clusters merge in a hierarchical way until they form a single cluster that contains all objects. In each time a merge occurs between two clusters, they would be replaced by a representative cluster that is the mean (or any other appropriate measure min, max, average … etc.) of the original two clusters. On the other hand, the divisive hierarchical clustering starts by having a single cluster that contains all objects and then divide it into smaller subclusters in a recursive way. The algorithm will terminate when subclusters at the lowest level are coherent enough (that is objects are very similar to each other) or when these subclusters contain at most a single object.

The distance between clusters is computed in each run to find closest clusters and merge them accordingly. Let *Ci* and *Cj* be two clusters that have means *mi* and *mj*, respectively, and have number of objects *ni* and *nj*, respectively. Then the distance between cluster *Ci* and *Cj* can be computed according to one of the following measures:

Minimum distance: $MinDist(Ci, Cj) = \min_{oi \in Ci, oj \in Cj} |oi - oj|$

Maximum distance: $MaxDist(Ci, Cj) = \max_{oi \in Ci, oj \in Cj} |oi - oj|$

Mean distance: $MeanDist(Ci, Cj) = |mi - mj|$

Average distance: $AverageDist(Ci, Cj) = \frac{\sum_{oi \in Ci, oj \in Cj} |oi - oj|}{ni + nj}$

As been mentioned in [22], the minimum and the maximum measures represent extremes when measuring the distance between clusters. More precisely, the minimum considers the two closest points and the maximum considers the two distant points in two clusters to evaluate the distance. On the other hand, the average distance can be considered as a compromise between these two measures and it is advantageous in a way that it is not sensitive to outliers and noise as the minimum and maximum measures might be.

In microarray analysis the aim of clustering process is to find genes that are coexpressed together or genes that share the same expression patterns after drug treatment in a process called gene based clustering. In this case genes are considered as the objects and samples are considered as the features. Another important issue is to find cell lines or samples that share same expression patterns after being treated with drugs where this process called sample based clustering. In this case samples under different drug treatment are considered objects and genes are considered features. Hierarchical clustering [27] , K-means clustering [28] and many other algorithms that have been reviewed in [29] were all used to analyze biological data.

Since the purpose of clustering in this thesis was to visualize the drug-disease connectivity map that resulted from the integrative paradigm, hierarchical clustering with the average distance measure, to find distance between clusters, was used in this thesis.

2.2.3. Classification

Classification is the process of predicting labels or numeric values for data of interest. It first starts by a learning step where a classification model is constructed and then a classification step to predict class labels for given data. In the learning step, the classification algorithm builds a classifier by analyzing and learning a pattern in the training data that is made up of set of tuples and associated class label.

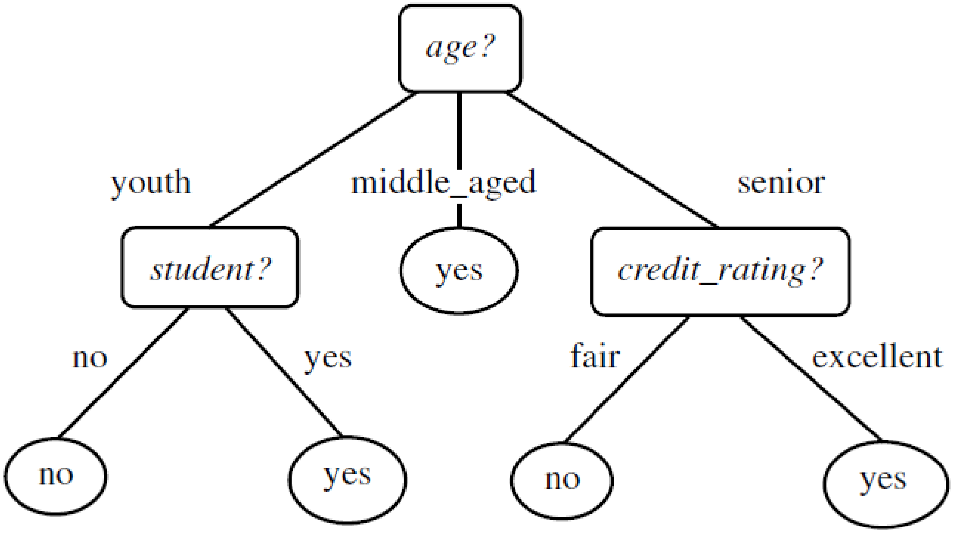
Different classification methods from statistical and machine learning area have been applied to improve inferring biomedical and pharmacogenomics knowledge. These algorithms basically start by extracting the most informative genes in a process called feature selection and thus improving cancer classification. After selecting the informative genes, classification algorithms can play a major role in revealing the most informative biomarkers that can even sub-classify different types of tumors or different compound treatments. In this section I discuss some of the most commonly used classification algorithms; Decision tree classifier, Naïve Bayes classifier and Regression models.

2.2.3.1. Decision Tree classifier

Decision Tree classifier is represented by flow chart like tree structure where each internal node denote a test on attribute, branch represents outcome of the test and leaf node holds class label. Given a new data object, we can track its attribute on the tree until we make a decision about its class label. An example of decision tree classifier is shown in **figure 2.2**.

The main issue with decision tree classifier is with selecting a tree induction algorithm that is time efficient and accurate using the training tuples. Many algorithms have been proposed to induce decision tree; Iterative Dichotomiser (ID3), C4.5 and classification and regression trees (CART). These algorithms follow a top-down recursive divide and conquer technique to induce the decision tree. More specifically, these algorithms apply attribute selection method to find the best splitting criterion or the attribute to split at specific level in the tree. This splitting attribute is the one that provides the best way to separate or partition the tuples in *D* into individual classes.

Figure 2.2: Example of Decision Tree Classifier



This figure shows an example of decision tree classifier where the classification attribute is buys computer. Having any new customer with specific attributes, we can track the tree and predict a yes/no answer for the concept buys-computer. Adopted from [22]

Let *D* be set of data and let us suppose that the class label attribute has *m* distinct classes *Ci* for i=1, ……… m (in the previous example *m*=2, that is yes or no). Let $C_{i,D}$ be set of tuples of class *Ci* in *D* and let the attribute *A* has *n* distinct values *aj* for j=1,……….. n*.* Then according to Information Gain algorithm the attribute with maximum Information Gain is selected. Information Gain is computed in two steps. First it computes the expected information needed to classify a tuple according to the following

$$Info(D) = -\sum_{i=1}^{m} \frac{|C_{i,d}|}{|D|} \log\left(\frac{|C_{i,d}|}{|D|}\right)$$

Equation 2.8

Then it computes the information needed using Attribute *A* as splitting attribute according to the following

$$Info_A(D) = \sum_{j=1}^{n} \frac{|D_j|}{|D|} * Info(D_j$$ **Equation 2.9**

Finally the information gain obtained with using attribute *A* is computed according to the following

$$Gain(A) = Info(D) - Info_A(D)$$ **Equation 2.10**

The only problem with using this score is that some attribute might have n-distinct values for n-tuples. Then the information gain prefers to split using this attribute and thus splitting each tuple independently. This will result in a maximal *Gain (A)* but this is bias and does not represent what we seek for. Therefore, Gain Ratio algorithm has been used to create a new score that tackle this problem. Gain Ration is computed in two steps. First it computes the information to be generated by splitting dataset *D* into *n* partitions according to the *n* distinct values for attribute *A* according to the following

$$SplitInfo_A(D) = -\sum_{j=1}^{n} \frac{|D_j|}{|D|} * \log\left(\frac{|D_j|}{|D|}\right)$$ **Equation 2.11**

Then it selects the attribute with the maximum Gain Ratio that is computed according to the following

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)}$$ **Equation 2.12**

2.2.3.2. Naïve Bayesian Classification

Bayes classifier is a statistical classifier that can predict the probability that a given tuple belongs to a particular class. The term Naïve came from the fact that, this algorithm does not consider the dependency between different attributes, if there is any. Suppose having *m* different classes C_j for j=1….. m. Let $P(Ci|X)$ denotes the probability that tuple X belongs to class C_i, then Naïve Bayesian classifier predicts that tuple X belongs to class C_i if and only if $P(Ci|X) > P(Cj|X)$ for j =1……..m such that $i \neq j$ Where $P(Ci|X)$ is computed according to the following

$$P(Ci|X) = \frac{P(X|Ci) * P(Ci)}{P(X)} \qquad \textbf{Equation 2.13}$$

Where

$$P(X|Ci) = \prod_{j=1}^{n} P(xj|Ci) = P(x1|Ci) * P(x2|Ci) .... * P(xn|Ci) \qquad \textbf{Equation 2.14}$$

And

$$P(Ci) = \frac{|C_{i,D}|}{|D|} \qquad \textbf{Equation 2.15}$$

Note that *P(X)* is constant for all classes and it would not affect the final decision, therefore the problem is reduced to a maximization problem for the term $P(X|Ci) * P(Ci)$. Finally it is really important to note that $P(xj|Ci)$ represents the probability of having value

x_j for attribute A_j in class C_i in dataset D, and this will result in two cases depending on the data type of attribute A_j. If A_j is categorical (discrete value) then $P(xj|Ci)$ represents the number of tuples of class Ci in D having x_j for attribute A_j divided by the number of tuples in class C_i in D. On the other hand, if A_j is continuous, then it must have Gaussian distribution with mean M and standard deviation S that is

$$G(x, M, S) = \frac{1}{\sqrt{2\pi S}} * e^{\frac{-(x-M)^2}{2*S^2}} \qquad \textbf{Equation 2.16}$$

Then

$$P(xj|Ci) = G(xj, M_{Ci}, S_{Ci}) \qquad \textbf{Equation 2.17}$$

2.2.3.3. Regression Modeling

Regression modeling is one of the most powerful machine learning techniques that is been used for predictions. In this section I discuss Univariate Linear Regression, Multivariate Linear Regression and Regularized Logistic Regression. Finally I highlight the reasons for using Regularized Logistic Regression as a prediction algorithm in this thesis.

2.2.3.3.1. Univariate Linear Regression (ULR)

Linear regression in general is used to predict some real valued output for prediction attribute given the right value for this attribute in the training tuples. In ULR, there is only a single input feature, let us say x, a prediction attribute, let us say y and m training samples. Let $(x^{(i)}, y^{(i)})$ denote the i^{th} training example. Then ULR tries to find a linear line in the two dimensional plane that best describes the data. More specifically it defines the following linear function to fit the training examples

$$H(x) = \beta_0 + \beta_1 x \qquad \textbf{Equation 2.18}$$

where β_0 and β_1 are model parameters. The main issue in ULR is to choose β_0 and β_1 so that *H(x)* is close to *y* for training examples (*x, y*). In other words, the problem can be described as a minimization problem that minimizes the distance between the prediction hypothesis and other *y's* in the training samples. Therefore the cost function *J* (β_0, β_1) that needs to be minimized and is defined as follows

$$J(\beta 0, \beta 1) = \frac{1}{2m} \sum_{i=1}^{m} (H(x^{(i)}) - y^{(i)})^2 \qquad \textbf{Equation 2.19}$$

Gradient Descent optimization algorithm can solve this problem very efficiently according to the following procedure:

1- Start with any random value for β_0 and β_1

2- Keep changing β_0 and β_1 until reaching the minimum score.

This is to say, gradient descent start with a specific point on the function *J* (β_0, β_1) and keep on going down gradually until reaching the minimum. Note that, in some applications starting at different points will result in many local minimums, but the good thing about the cost function of regression models is that it is always convex in shape, thus result in only a global minimum. The optimization procedure utilizing gradient descent is summarized in the following algorithm:

Repeat until convergence for *j=0* and *j=1* {

$$\beta o := \beta 0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (H(x^{(i)}) - y^{(i)})$$

$$\beta 1 := \beta 1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (H(x^{(i)}) - y^{(i)}) * x^{(i)}$$

}

The new assignments for β_0 and β_1 have been obtained after taking the first derivative for the cost function J (β_0, β_1) on β_0 and β_1, respectively. Note that α is called the learning step or the step size that varies from being too small (moving small steps towards minimum) to being too large (moving large steps that may fail to converge or even diverge).

2.2.3.3.2. Multivariate Linear Regression (MLR)

MLR applies the same concept and technique that ULR does. The only difference is with having *n* variables (Multivariate) instead of a single variable. the regression model will be summarized according to the following

$$H(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots\ldots\ldots \beta_n x_n \qquad \textbf{Equation 2.20}$$

Accordingly, the gradient descent would be represented with the following algorithm

Repeat until convergence for j=1……………..n {

$$\beta j := \beta j - \alpha \frac{1}{m} \sum_{i=1}^{m} (H(x^{(i)}) - y^{(i)}) * x_j^{(i)}$$

}.

Note that β can also be found by using linear algebra (or normal equation) technique. That is to say, the vector $\beta \in \mathbb{R}^{n+1}$ can be found according to the following expression $\beta = (X^TX)^{-1}X^Ty$ where X represents the data matrix, y represents the prediction attribute.

2.2.3.3.3. Logistic Regression

In linear regression, after finding the parameters vector β, one can implement the following formula to find a prediction for new feature x, $H(x)= \beta^TX$. Again the predicted value here is a real (continuous) score for specific prediction attribute. Sometimes, as in this thesis, the interest would be in finding a probability that a certain data object belongs to specific class C_i. Here where logistic regression becomes in use. That is to say, instead of predicting a continuous score, logistic regression finds the probability that a certain data object belongs to the positive class. In order to satisfy this need, logistic regression defines a new prediction hypothesis that is $H(x)= g(\beta^TX)$ where $g(z)=1/(1+e^{-z})$ is called the sigmoid (or logistic) function. Therefore the new prediction hypothesis $H(x)$ would be

$$H(x) = \frac{1}{1+e^{-\beta X}} \qquad \textbf{Equation 2.21}$$

Thus the new prediction score ranges from 0 to 1. A nice property about this function is that whenever $\beta^TX>0$ then $H(x)$ would be more than or equals to 0.5 and thus one can predict x to be in the positive class. On the other hand, if $\beta^TX<0$ then $H(x)$ would be less than or equal to 0.5 and one can predict that data object x to be in the negative class. Another interesting property about logistic regression is that, unlike linear regression, it can define non-linear decision boundaries enabling for defining more accurate decision boundaries between the two classes.

Finally it is worth mentioning that logistic regression has a new definition for the cost function. This is because considering the same cost function (the square difference between *H(x)* and *y*) defined earlier for linear regression would not result in a convex shape (because of having a different definition for *H(x)*) and thus it is not guaranteed that the algorithm would converge to global minimum. Therefore, the cost function for logistic regression is defined as follows

$$\underset{j=1\ldots n}{J(\beta j)} = \frac{1}{m}\sum_{i=1}^{m} Cost(H(x^{(i)}), y^{(i)}) \qquad \textbf{Equation 2.22}$$

where

$$Cost(H(x), y) = \begin{cases} -\log(H(x)) & if\ y = 1 \\ -\log(1 - H(x)) & if\ y = 0 \end{cases} \qquad \textbf{Equation 2.23}$$

Thus the cost function can be rewritten as

$$J(\beta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log H(x^{(i)}) + (1 - y^{(i)}) \log(1 - H(x^{(i)})\right] \qquad \textbf{Equation 2.24}$$

Again, parameters *β* can be computed using gradient descent algorithm by keep on updating β_j for j=1……n simultaneously.

2.2.3.3.4. Regularized Regression

For some application for regression models, especially when data contains large number of features, the resulted model might have the problem of over-fitting. Over-fitting problem can be described as fitting the training set very well to the level of producing high variance and making it hard for the model to generalize for new examples. On the other extreme, sometimes the model will result in under-fitting where the model is not robust enough to fit the training data very well. **Figure 2.3** shows an example for a model with under-fitting, just right model and a model with over-fitting for the same training samples.

Since the whole model is built based on parameter values in vector *β,* a penalty can be added to avoid having large scores of *β* and consequently avoiding the problem of over-fitting. Therefore, this penalty (regularization parameter) would control the trade-off between fitting data very well and keeping parameters small. So the cost function for regularized logistic regression can be re-written as

$$J(\beta) = -\left[\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log H(x^{(i)}) + (1-y^{(i)}) \log(1-H(x^{(i)}) + \frac{\lambda\beta j}{m}\right] \qquad \textbf{Equation 2.25}$$

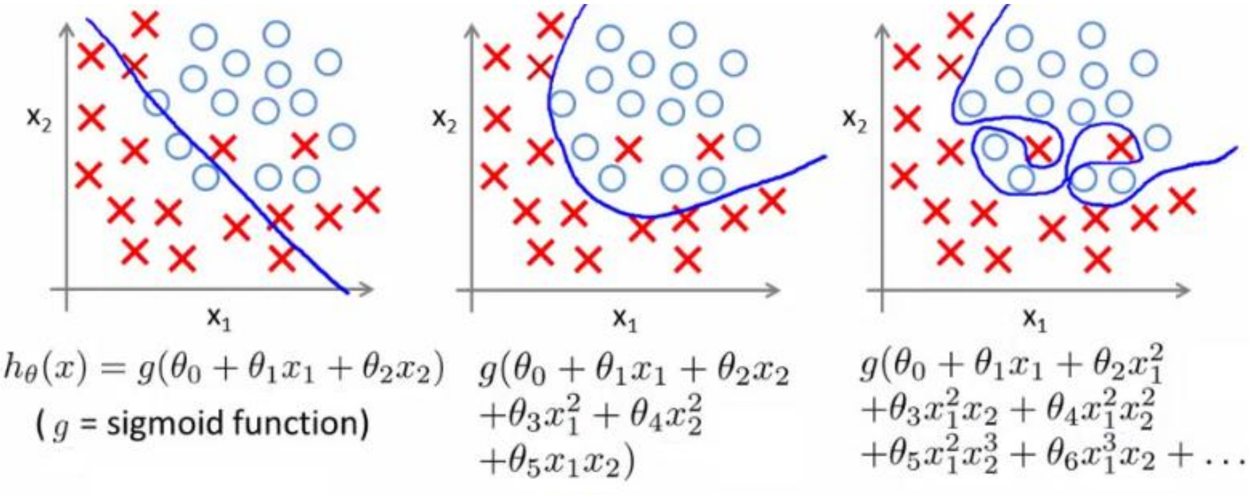
Accordingly, the gradient descent algorithm for this function can be re-written as

Repeat until convergence for j=1……………..n {

$$\beta_0 := \beta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)} - y^{(i)}) * x_o{}^{(i)}$$

$$\beta_j := \beta_j - \alpha\left[\frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)} x_j{}^{(i)}) + \frac{\lambda\beta_j}{m}\right] \}$$

Figure 2.3 Under-fitting, Just-Right and Over-fitting


x_2
x_1
$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$
(g = sigmoid function)
x_2
x_1
$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2$
$+\theta_3 x_1^2 + \theta_4 x_2^2$
$+\theta_5 x_1 x_2)$
x_2
x_1
$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2$
$+\theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2$
$+\theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \dots$

This figure shows three examples of three different regression models applied on the same data set. These models represent under-fitting, Just-Right and Over-fitting from left to right respectively. Adopted from [105]

In this thesis, logistic regression model was used for various reasons. First of all, the purpose of using machine learning algorithm in this thesis was to find a probabilistic measure that a data point belongs to a certain classifier. Therefore, out of the different methods discussed above, Naïve Bayesian classifier and logistic regression are the best to fit into this kind of problems. The reason for choosing Logistic regression over Naïve Bayesian classifier is the assumption of completely independent features in Naïve Bayesian classifier. Noteworthy that logistic regression is consistent with Naïve Bayesian classifier with this property but it is not rigidly tied to it. More precisely, logistic regression uses the conditional likelihood maximization algorithm that adjusts its parameters to maximize the fit to data; even if they were inconsistent with Naïve Bayes parameter estimates. Another reason is that it has been shown in a published work [30] that in several data sets logistic regression outperforms Naïve Bayes classifier when many training examples are available (and that was the case in this thesis).

2.2.4. Gene Set Enrichment Analysis

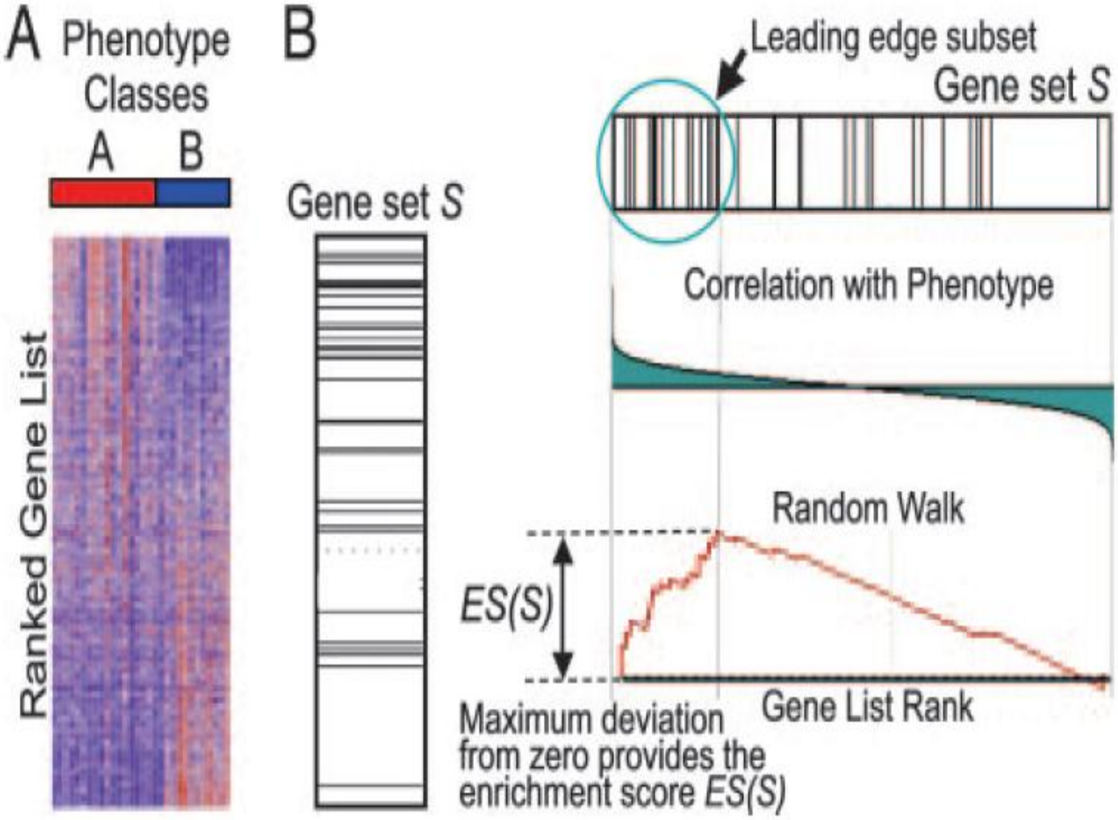
Even though clustering and classification have made a great contribution in the medical and pharmacogenomics fields, these methods are still based only on selecting a set of over-expressed and a set of down-regulated genes to drive the model. Choosing such a set of genes to discern telltale biological clues is not without drawbacks [31]**.** As discussed in [31] some of the limitations that might rise are: (i) Sometimes the relative differences between two biological states are modest relative to microarray technology noise and thus produce no gene above the statistical significance threshold, (ii) Alternatively, sometimes a very long list of genes might satisfy the significance requirements making the biological interpretation a daunting task, (iii) Finally, the distressingly lack of consistency of the ranked list of genes for the same biological system between different labs.

To tackle these limitations a method called Gene Set Enrichment Analysis (GSEA) has been proposed [31]. GSEA evaluates microarray data at the level of gene sets. These gene sets are to be defined based on prior knowledge which can be either from using the coexpression experiments or by obtaining them from databases or mining biomedical texts. As shown in **Figure 2.4,** assume that we have a ranked list of genes ***A or B*** where the genes have been rank-ordered according to their differential expression from the most up-regulated at the top to the most down-regulated at the bottom. In addition, assume that there is a set of genes ***S*** that represents the query signature and contains a set of up-regulated genes and a set of down-regulated genes based on prior knowledge. Then the major contribution of GSEA is to determine whether the up-regulated query genes in the set $\boldsymbol{S}$ tend to occur toward the top of the list $\boldsymbol{A}$ and down-regulated genes in the set $\boldsymbol{S}$ tend to occur toward the bottom of the list $\boldsymbol{A}$ thus indicating high positive connectivity or vice versa indicating high negative connectivity. A zero connectivity score is assigned where the enrichment scores for the up- and down- regulated genes have the same sign indicating that the genes in the signature $\boldsymbol{S}$ are randomly scattered throughout the ranked list $\boldsymbol{A}$. This method has become one of the most popularly used methods in drug mode of action discovery, cancer biomarkers detection and other phenotypic pathway analysis.

2.3. Biomedical Text-Mining

Text mining (TM) is an interdisciplinary field that combines knowledge amongst data mining, computational statistics and computer science [32]. Many techniques are being implemented in TM field including ontology, taxonomy creation, clustering, classification and document summarization [32]. Therefore the general idea for text-mining is to transform a bag of words or data objects into a structured data format based on term frequency and subsequently apply data mining techniques to infer knowledge. The major challenge for TM is the fact that texts, from computer point of view, are unstructured information that needs preprocessing and normalization techniques in order to be computationally meaningful data.

Figure 2.4: Gene Set Enrichment Analysis Overview



This figure shows an overview for the whole GSEA concept. Panel A shows a set of genes ranked from the most positively relevant to the most negatively relevant to a particular phenotype (drug, disease or any others) in addition to a set of genes S with their location regarding the ranked list of genes (A and B). Panel B shows a plot of the running sum for S in the data set including the maximum deviation from zero which represents the actual enrichment score. Leading edge subset represents a set of genes that contributed more in improving the positivity of the enrichment score**. Adopted from [15]**

In general, biomedical TM can be defined as the computational discovery of new, previously unknown information, by automatically extracting information from different written resources [33]. TM integrates a broad spectrum of heterogeneous data sources and thus providing tools for analyzing, extracting and visualizing information, with the aim of helping researchers to transform biomedical data into usable information and knowledge [34]. Generally speaking TM consists of two major steps: information retrieval (IR) and information extraction (IE) [35] where each of which has its own role in achieving accurate and consistent text mining results.

2.3.1. Information Retrieval

The process of TM starts with IR that finds abstracts related to specific biological entities of interest. These entities can be genes, proteins, chemical compounds or diseases. There are two very common searching approaches in IR [36]: (i) rule-based or knowledge-based; and (ii) statistical or machine learning based. The rule based approach uses patterns that rely on basic biological insights or by encapsulating representative relationships between entities in what's called frames. An example of such a frame <Drug A activate Protein C>. The statistical approach uses syntactic parse trees which can be decision based to classify related biomedical literature. More precisely, it builds decision tree using a set of training documents and then classify documents accordingly as being described in previous section.

2.3.2. Information Extraction

IE aims to extract pre-defined types of facts or relationships between biological entities. IE can roughly be divided into two approaches as well [36]**:** (i) co-occurrences based approach, which identifies entities that co-occur within the text; and (ii) Natural language processing (NLP) based approach where the syntax (the orderly manner in which words are put together to form phrases and sentences) and semantics (the meaning that is implied by words and sentences) are combined together for more accurate predictions. As explained in [33] the NLP starts by taking the text to identify sentence and work boundaries, and a part of speech tag (noun or verb) is assigned to each word. A syntax tree is then derived for each sentence to delineate noun phrases and represent their interrelationships. Simple dictionaries are subsequently used to semantically tag the relevant biological entities. Finally, a rule set is used to extract relationships on the basis of the syntax tree and the semantic labels.

Noting that, co-occurrence methods tend to give better recall but worse precision than NLP methods [37] [38]. Therefore, those are assumed to be well suited as parts of exploratory tools because of their ability to identify relationships of almost any type of entities [39]. In addition, co-occurrence methods can also be used to extract a specific type of relationships such as activation, inhibition, phosphorylation and others. This is basically done by combining the entities together with the keywords in a customized text-categorization system to identify the relevant abstracts or sentences [40] [41]. A drawback of these methods is that they have difficulties distinguishing between direct and indirect relationships (for example, whether a gene is directly or indirectly activated by a particular drug) [33].

2.3.3. Beyond NLP and Co-occurrences

Even though co-occurrence and NLP have important contribution in knowledge discovery, more computational analysis still needs to be done for better understanding. Many methods have been proposed to go deeper in TM and infer some indirect relationships [42] [43]. Note that, all of these methods are predicting indirect relationships using co-occurrences in a naïve way. For example if we want to investigate an indirect relationship between $gene\ A$ and $gene\ C$ then the method would first find the set of $genes\ B$ that are related to $gene\ A$, then it will find the set of $genes\ X$ related to $gene\ B$ and finally predict a relationship between $gene\ A$ and $gene\ C$ if: i). $gene\ C \subseteq$ set of $genes\ X$ ii). $gene\ C \nsubseteq$ set of $genes\ B$. Furthermore, NLP-based IE for text-mining have not been popularly used because there is only few NLP systems that are able to accurately extract sufficient large number of direct relationships that allow for indirect relationships inferring. This is happening due to the fact that there is no full-text access to all published papers and some results might have been assumed as trivial and no one has ever published them [33]. Consequently, one can conclude that text-mining performance can be improved by integrating other sources of data that is not necessarily available in the text.

2.4. Biological Data Integration

Data mining approaches that integrate text-mining with other biological data sources do not only have the potential to predict indirect relationships, but they have the potential to make biological discoveries, to understand biological pathways, to interpret many genetic behaviors and to unveil new drug indications that are irretrievable with using text-mining approaches independently.

2.4.1. Gene List Decryption Based

Most attempts to integrate text-mining with other biological data sources are performed either with the goal of enriching the list of up-regulated or down-regulated genes obtained from microarray. This is done because the list of up-regulated or down-regulated genes is very cryptic and requires lots of filtering and interpretation in order to be useful for knowledge discovery [44]**.** Therefore these methods were focusing on eliciting some correlation between these genes in order to understand the biological meaning of these differentially expressed genes in what's called Gene Ontology.

Another attempt to understand the biological meaning of this list by using text-mining was directed toward pathway analysis and/or regulatory network analysis. Pathway analysis mainly focuses on the functional, regulatory and physical interaction between genes instead of just summarizing a list of genes into some Gene Ontology terms. These attempts are basically based on mapping differentially expressed genes derived from microarray experiments onto precompiled pathways derived by manually analyze the literature [45]. KEGG database is one of the most popularly used tools for pathway analysis purposes. Noting that, some tools like STRING [46] defines some algorithms to rank these pathways according to their relevancy to the gene list, because large genes list might produce large number of pathways as well. These pathways can be interconnected to build what's called gene-regulatory network both by using microarray data source and text-mining data source and combine them to result in a combined score represent the interaction between those genes [46].

2.4.2. Gene Network Based

Because of its robustness, networks formalize the major and the initial step for text and data integration. There are several valuable and exploratory tools that are being used to build protein networks based on text-mining and high-throughput experiments [46] [47]. In addition, some authors used networks to integrate different data sources to provide insights into the molecular basis of a disease. For instance, a literature-based protein networks can be integrated with genetic-linkage and gene-expression data to identify some marker genes for a disease, based on their interaction with genes that are already known to have a role in that disease (seed genes) [48]. The resulting molecular networks can be searched for sub-networks that may harbor disease-relevant genes and thus a better understanding for the disease pathophysiology. This can be done by using close nodes from graph theory.

2.5. Drug Repositioning

The repositioning of drugs, already approved for human use, mitigates the costs and risks associated with early stages of drug development and offers shorter routes to approval for therapeutic indications [2]. As the Nobel laureate and the pharmacologist James Black said, "The most fruitful basis for the discovery of a new drug is to start with an old drug."

Pharmaceutical companies save up to 40% of the overall cost of launching a drug to market by skipping many toxicological and pharmacokinetics assessments tests [3]**.** Many examples have shown the successfulness of drug repositioning include, the indication of retinoic acid for acute pro-myelocytic leukemia [49], the indication of thalidomide for severe erythema nodosum leprosum [49], the indication of cimetidine for lung adenocarcinoma [2] and the indication of miltefosine for the treatment of visceral leishmaniasis [50].

To summarize, high-throughput microarray technology, biological databases and text-mining data integration process is a very powerful technique that can provide a closer insight into many biological problems including drug repositioning. Out of the different method proposed to integrate them, two methods are known to be most popularly used in that field [45]: (i) build a list of differentially expressed gene using microarray data and then use the text-mining techniques to prioritize this list of genes in regard to a particular disease, finally build the gene-regulatory network or run pathway analysis (ii) build a set of relationships between different biological entities (genes, diseases or drugs) using text-mining techniques and validate these associations by resorting to microarray data. According to our knowledge, there is not any method that has been proposed to integrate the ranked list of genes obtained from high-throughput technology with the ranked list of genes obtained from text-mining for the purpose of knowledge discovery.

Chapter 3: Computational Techniques for Drug-Repositioning

From the introduction we realized that with the huge amount of researches that are being done on disease and drug discovery, a huge amount of biomedical information are now stored in the web either as microarray data, scientific papers or databases. Since then, the prevailing approach to drug repositioning becomes based on utilizing these valuable data sources in screening libraries of lead compounds or diseases against biological targets of interest. This approach fits to the concept of molecular connectivity map which starts to gain a huge popularity in computational sciences. In computational biology, connectivity map can be defined as using a statistical measure in order to find positively or negatively connected biological entities (drugs, diseases or genes) based on specific features [10]. Noteworthy, that this important bio-computational concept have been used to build associations based on different data sources and different frameworks.

For example **Lamb et al.** [10] built in a connectivity map to associate small molecules, genes and diseases using gene expression signature as their data source and enrichment set analysis as their statistical measure. From the other hand, **Li J et al.** [51] built in a connectivity map to associate disease related genes with drugs using molecular interaction networks and PubMed abstracts as their data source and regularized log-odds function as their statistical measure.

3.1. Computational Methods for Drug Repositioning

Generally speaking, many attempts have been proposed in the field of computational biology for the purpose of drug/disease or drug/target prediction and prioritization based on a particular connectivity score. In this section, I discuss some of the most popularly used ones: similarity based, microarray based, text-mining based and data integration based approaches.

3.1.1. Similarity Based Drug-Disease Prediction (Chemogenomic Approach)

Most approaches that fall in this category were developed to integrate many similarity measures to predict new drug-target or drug-disease associations. Drug-drug similarity, protein-protein similarity and disease-disease similarity are all integrated with a drug-protein interaction network or drug-disease interaction network in order to predict new drug-protein or drug-disease associations [52] [53] [54].

Gottlieb et al. [55] developed a new approach "PREDICT" that can directly predict drug-disease associations including both FDA approved drugs and other molecules in the experimental phase. The algorithm works in three phases: (i) building five drug-drug similarity measures and 2 disease-disease similarity measures; (ii) building a classification features and subsequent learning classification rule that can distinguish between true and false drug-disease associations by using these similarity measures; and (iii) applying a logistic regression classifier to predict any new possible drug-disease associations. Thus for a given drug-disease association from the gold standard, the authors computed an association score by considering all the other known drug-disease association. Let K denote the set of known drug-disease associations. Given an-unknown association between drug dr_i and disease ds_j (dr_i, ds_j), the algorithm would compute a prediction score using the following formula

$$Score(dr_i, ds_j) = \max_{i',j':(dri',dsj')\in K} \sqrt{S(dr_i, dr_{i'}) * S(ds_j, ds_{j'})}$$ **Equation 3.1**

Noteworthy that the similarity scores were all normalized to be in the range between zero and one. The similarity measures that have been used in this work are:

For drugs

(i) Chemical Similarity: This is based on the Jaccard score between the drug's fingerprints, that is, the ratio between the intersection and the union when considering each fingerprint as a set of elements.

(ii) Side effect based: This is based on the Jaccard score between text-mining curated known or predicted side effects.

(iii) Target sequence based: Based on Smith-Waterman sequence alignment score between different drug targets.

(iv) Target Closeness in Protein-Protein interaction network (PPIN): This is based on the shortest path distance between each pair of drug targets in the protein-protein network. Distances were transformed to similarity values based on the following formula: $S(p_i, p_j) = A \quad e^{-bD(pi,pj)}$ Where $S(p_i, p_j)$ represents the similarity measure; D (p_i, p_j) is the shortest path between these proteins in the PPI network; and (A and b) are constants that need an expert knowledge.

According to the author's cross validation experiment A can have a default value of e * 0.9 and b can have a value of 1.

(v) GO based: This is based on the gene ontology similarity of drug targets.

For Diseases

(i) Phenotype similarity: text-mining based phenotype detection for diseases. These phenotypes have been extracted using OMIM database.

(ii) Semantic phenotypic similarity: This one is based on the hierarchical structure of the HPO [56] database that map ontology nodes with OMIM diseases to construct a semantic similarity score.

(iii) Genetic Based: This one is based on the Jaccard score between different disease's signatures.

Even though this technique attained high sensitivity and specificity in cross-validation experiments but it is not without limitations. Firstly, the proposed method used 5 different drug similarity measures and 3 different disease similarity measures. This will make it hard to include a drug without having its chemical structure, its side effects, its target's sequence, its target (PPIN) and its target's gene ontology. The same thing is applicable to diseases. Furthermore this method does not consider the similarity between drug's molecular actions. It only considers the known targets for drugs to define similarities. Sometimes there might be hidden or unknown drug targets that are not considered in this study. In addition it might be the case that some drug-target information has been published in papers but without having them entered into drug-target databases that the authors used. Finally this method does not map between the up/down regulated genes for drugs and diseases. This might result in a bias since drugs trigger its action on target genes and have a consequence effect on other off-target genes.

3.1.2. Microarray Based Approaches

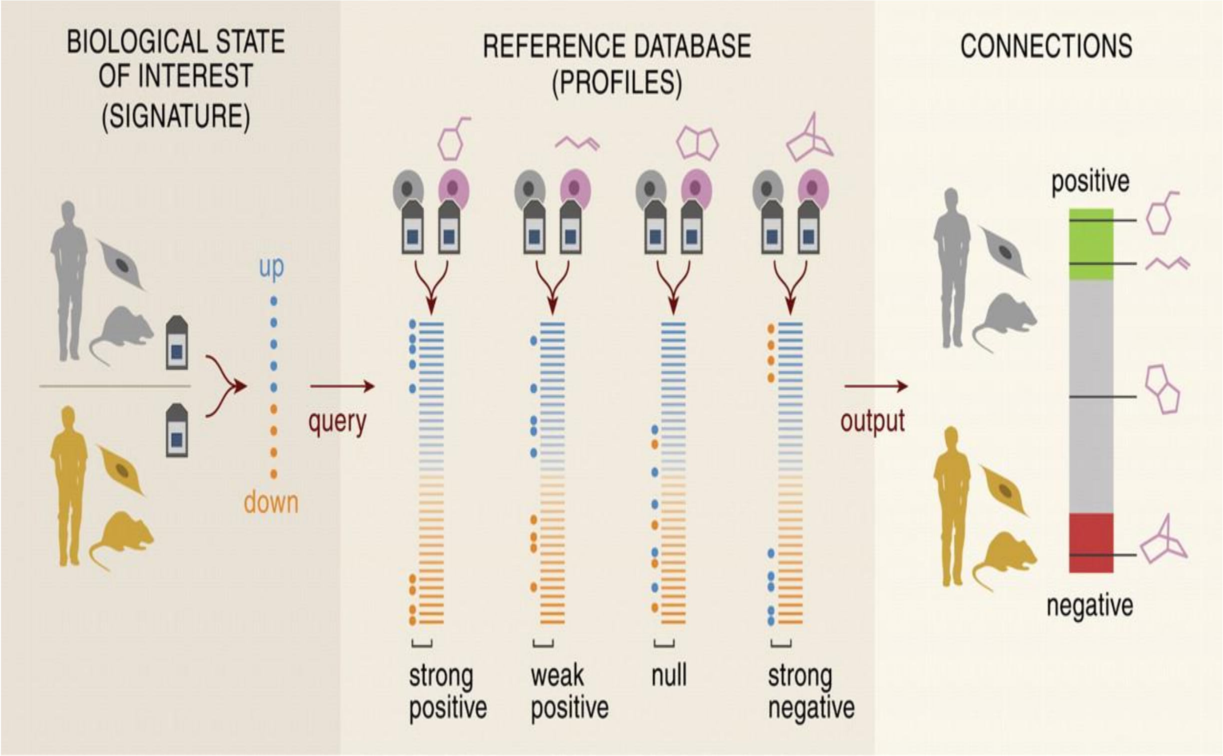
From the moment microarray gene expression profiling was unleashed on the world it was obvious that biomedical research was going to change [57]. Gene expression microarrays have been regularly and broadly applied in clinical studies of human diseases. Comparative gene expression analysis of benign and malignant tumors, peripheral and secondary organs, mutated and un-mutated, chromosome trans-located and original are all used to study the molecular pathophysiology of a disease or diagnostic marker [7] [8] [9]. Microarrays have been also used to discover the molecular effect of drug compounds [10] [58] [2].

Lamb et al [10] studied 164 distinct small molecule and perturbagens that have been selected to represent a broad range of toxicological effect. These expression profiles include U.S. Food and Drug Administration (FDA)- approved drugs, experimental drugs and other bioactive molecules. The authors created a reference gene-expression profiles by using a nonparametric fashion. Each profile was compared to its corresponding control by using z-score then all genes were ranked according to their differential expression relative to the control from the most up-regulated genes on the top to the most down-regulated genes on the bottom. Later on, the authors used a nonparametric, rank-based pattern matching strategy based on the Kolmogorov-Smirnov static which has been formalized in GSEA [31]. Once a researcher has a query signature, the system will work by finding the similarity to each of the reference expression profiles in the data set. This similarity is basically trying to find whether each up-regulated query genes tend to appear near the top of the list and down-regulated query genes near the bottom (positive connectivity) or vice versa (negative connectivity), yielding a connectivity score ranging from +1 to -1.

Noteworthy that the resulted high positive associated instances are having very similar mode of action to the instance from which the query signature is obtained. On the other hand, the resulted high negative associated instances are counteracting the effect of the instance from which the query signature is obtained. Therefore, if the query signature is referring to a disease, then all the negatively connected instances will be considered for further treatment investigation. The authors biologically validated that such approaches can be used to better understand the molecular mechanism of compounds i.e. identification of gedunin (triterpenoid natural product purified from medicinal plants) as an HSP90 inhibitor. **Figure 3.1** shows the whole process starting from having a query signature for a particular disease or drug ending by suggesting a list of drug that can be used to treat that particular disease.

One drawback of this approach was their finding that signatures were conserved across cell types and settings, the thing that did not allow the exact understanding of the molecular effect of a specific drug or disease.

Figure 3.1 General Framework for the Connectivity Map Project



This figure shows the whole framework for the connectivity map project. This framework needs to have a drug specific ranked list of genes called reference database and a query signature contains an up-tag and down-tag set of genes. Once the signature gets dropped into the database, the system will compute the enrichment score between this signature and the ranked lists in the reference database. Finally the drugs will be ranked from the most positively enriched (indicating similarity) to the most negatively enriched (indicating dissimilarity). **Adopted from** [10]**.**

To tackle this problem **F. Iorio et al**. [15] reported another interesting work by using the same data in [10] for the purpose of building drug-drug network and identifying drug communities**.** The authors computed for each drug a “consensus” synthetic transcriptional response summarizing the transcriptional effect of the drug across multiple treatments on different cell lines and/or at different dosages. To do so, the authors sought to run a rank-merging procedure that first compares the ranked lists obtained from the same drug treatment using the Spearman’s Footrule similarity measure [59]. The algorithm then merges the two lists that are most similar to each other (the ranked lists with lowest distance measure), following the Borda Merging Method [60]. This will result in a single ranked list that replaces the original two lists. This procedure keep on running in a hierarchical way until only one and representative ranked list remains to represent each drug. Mathematical description and an example of this procedure are provided in **chapter 4**.

This approach helped in capturing the consensus transcriptional response of a compound across different experimental settings and thus reducing non relevant effects due to toxicity, dosage and cell line. After obtaining a single and representative ranked list for each drug the authors defined a distance measure between every drug and all the other drugs based on GSEA. To find the distance between drug ***A*** and drug ***B*** the authors first considered the most 250 up-regulated genes (up250) and the most 250 down-regulated genes (down250) once from drug *A* ranked list and once from drug *B* ranked list. Then to find the enrichment score of these up/down tag from one drug (let's say *A*) regarding the ranked list of the other (let's say *B*) the authors used the following formula

$$TES_{A,B} = 1 - \frac{ES_{B,up250} - ES_{B,down250}}{2} \qquad \textbf{Equation 3.2}$$

Where $TES_{A,B}$ is the total enrichment score of the up-tag and down-tag signature from drug *A* regarding the ranked list of drug *B*. ES_B is the enrichment score of either the up-tag or the down-tag of drug *A* regarding the ranked list of drug *B*. Noting that ES_B ranges from -1 to +1. The closer the up/down tag is to the top, the closer is the value to +1. The closer the up/down tag is to the bottom, the closer is the value to -1.

$TES_{A,B}$ quantifies how much the genes in up250 are at the top of the ranked list of *B* and how much the genes in down250 are at the bottom of the ranked list of *B*. The closer these two statements are to the truth; the closer to zero is the value of *TES*. After that they computed the $TES_{B,A}$ in the same way as it is not necessary that $TES_{A,B}= TES_{B,A}$. Finally the average distance between two drugs ($D_{A,B}$) would be computed according to the following

$$D_{A,B} = \frac{TES_{A,B} + TES_{B,A}}{2} \qquad \textbf{Equation 3.3}$$

This distance measure was the input for a community builder algorithm in order to construct drug communities.

Another two interesting examples that used GSEA to build a direct drug-disease connection were the methods developed in [58] and [2]**. Sirota M et al.** [2] developed a framework where they extracted microarray expression profiles for 100 diseases including several human cancer cell lines. The authors then defined a signature (a set of significantly up-regulated genes and significantly down-regulated genes) for each of the 100 diseases using SAM method [20]**.** Then they statistically compared each of the disease signatures to each of the reference ranked list of genes developed in the Connectivity map project [10]. The authors computed an enrichment score for every pairing of drug and disease where +1 indicates similar correlation of signatures and -1 indicates opposite signature. In other words, -1 enrichment score represents a contradicting behavior between the drug effect and the disease pathophysiology and thus there would be a paramount potential for this drug to treat that particular disease. To evaluate significance, the authors generated 100 random drug ranked lists and measured their enrichment with each disease signature.

These randomly generated scores have been compared with actual enrichment scores for evaluation purposes. Biological validation has been applied on some of their results for example; the prediction of cimetidine as a candidate therapeutic in the treatment of lung adenocarcinoma. A drawback of this approach was diminishing or diluting the effect for a compound that has inconsistent effects on different cell lines.

Using microarray data only would not clarify whether drug performance on a specific cell line (breast cancer that has been extensively used to measure transcriptional response of drugs in Connectivity Map) is relevant to all types of diseases [6]. In addition, relying on gene microarray data alone may fail to match disease and drug effects that are not manifested at the gene expression level [31]. Therefore, it was necessary to integrate microarrays with other data sources that can build associations between different biological entities to improve results. Many methods have been proposed in the field of text-mining that were able to build associations between biological entities in a score based ranking scheme that represent the likelihood of associations between these biomolecules. Most of these methods are discussed in the next section.

3.2. Text-Mining Based Approaches

TM approach has shown many successful stories in this field. As we discussed in the introduction, TM consists of two major steps, information retrieval (IR) and information extraction (IE) [35]. IR will try to find literature or abstracts related to a particular drug/disease/or gene specified by the user. IE is then used to tabulate the relevant entities or knowledge from the retrieved documents either based on the co-occurrence or by using natural language processing. Text mining approach has been widely used to connect diseases, drugs and genes to build a connectivity map or a network between those entities [61] [43] [51].

3.2.1. Regularized Log Odds Function Based TM

Li J. et al. [51] proposed a computational framework to develop a disease specific drug-protein connectivity map by integrating molecular interaction network mining and text mining techniques. The basic idea was to generate a list of disease-related proteins and a list of disease related drugs as two-attribute dimension for drug-target map. The proposed paradigm starts by incorporating disease-specific seed genes/proteins derived from prior knowledge which can be either from OMIM database, expert knowledge or a set of differentially expressed genes from microarray expression profiles. This seed of genes is improved by expanding and re-ranking them in the functional context through reprioritizing them in disease-related molecular interaction networks. For this reason, a protein-protein interaction database has been used to include all direct neighbors of seed genes. Then they reconstructed the molecular interaction networks between those proteins by just considering the extended list of the seed list proteins. The relevancy of every single protein has been determined by using the resulted protein-protein network according the following formula:

$$R(p_i) = k * (\ln(\sum_{p_j \in NET} conf(p_i, p_j)) - \ln(\sum_{p_j \in NET} N(p_i, p_j))) \qquad \textbf{Equation 3.4}$$

Where p_i and p_j are proteins belonging to the disease related interaction network *NET*. *k* is an empirical constant (*k*=2 in their study). *Conf(p_i, p_j)* is the confidence of the interaction between protein p_i and p_j. *N(p_i, p_j)*= 1 if p_i interacts with p_j and equals to zero otherwise.

Noting that, this parameter will assign a score for every gene indicating the likelihood of their association to a particular disease. After obtaining this ranked list the authors extracted all the abstracts that have mentioned any of these proteins. Using these abstracts they have computed an enrichment score between drugs and proteins according to the following formula

$$Enrichment_{DP} = \ln(Abst_{DP} * N + \lambda) - \ln(Abst_{P} * Abst_{D} + \lambda) \qquad \textbf{Equation 3.5}$$

Where $Abst_{DP}$ is the total number of abstracts where *drug D* and *protein P* have been co-mentioned. $Abst_{P}$ and $Abst_{D}$ are the total number of abstracts in which protein *P* and drug *D* have been mentioned respectively. *N* is the total number of abstracts collected regarding a particular disease. λ is a constant that have been added to avoid out-of-bound errors when either one of $Abst_{DP}$, $Abst_{P}$ or $Abst_{D}$ is equal to zero.

A major limitation of this approach is that the scoring function is based on the connectedness of genes, where non-seed genes were not considered. This has shed the light to the conclusion that their results were biased toward seed genes. In fact, out of the top scored 20 genes, 19 were related to seed genes and only one was a novel prediction [61]. Trying to tackle this problem, **Ozgür A et al.** [61] developed different centrality measures based paradigm to prioritize genes without being biased toward the seed genes. This approach is discussed in next section.

3.2.2. Patterns Recognition based TM

Another interesting approach that uses the concept of ranking biological entities regarding an association score based on text mining is PolySearch. **Cheng D et al.** [43] developed this web-based text mining system for extracting relationships between human diseases, gene, mutations, drugs and metabolites. PolySearch displays links and ranks text, as well as sequence data in multiple forms and formats. A distinguished feature of PolySearch over other biomedical text mining tools is the fact that it exploits the presence of many other biological databases to improve the results.

PolySearch employs a text ranking scheme to score the most relevant sentences and abstracts that associate query terms, association words and database terms. This ranking scheme is based on a pattern recognition system that was defined by the authors. A central premise to their ranking strategy was the assumption that the greater the frequency of co-mentioning of two biological entities, the more significant is the association. In details, Polysearch identifies four different sentences R1, R2, R2 and R4 to compute the association score between the query word, the association word and the database word. R4 sentence is a sentence that contains one of the database terms (that is not necessarily related to the query word). R3 sentence is a sentence that has one of the database terms and the query word. R2 and R1sentences contain one of the database terms, one of the query terms and at least one association word. The only difference between R1 and R2 is that R1 has to pass one of the three patterns identified by the authors namely; compact patterns, general patterns and relaxed patterns. A brief description for every one of them is listed here:

- **Compact patterns:** An association word (activate, inhibit, activation, inhibition …. etc.) and the required word (drug, disease, gene …. etc.) must be within 10 words of the query word from the user. Noting that a stop word such as “that”, “which”, “whereas” or “no” cannot be within this pattern.

- **General patterns:** All the mentioned words must be within 15 words of the query word where a stop word cannot be within this pattern.
- **Relaxed patterns:** All the mentioned words must be within 40 words of the query word where stop words can be within this pattern.

Finally R1 sentences will be given a value of 50, R2 sentences will be given a value of 25, R3 sentences will be given a value of 5 and finally R4 sentences will be given a value of 1. All of these will be used to compute the Polysearch Relevancy Index (PRI) by simply taking the sum of these scores.

Just like relying on microarrays alone, relying on text mining tools is not without some limitations. For example PolySearch uses a relatively simple dictionary approach to identify biological or biomedical associations. This means PolySearch cannot identify novel or newly named diseases, genes, or drugs. In addition, PolySearch only considers parsing published papers leaving no room for utilizing other data sources like microarrays or biological networks. Similarly, the authors in [51] suggested further improvements to their text mining approach by integrating experimental gene expression or protein expression data as their method had a paramount bias toward the initial seed genes.

3.3. Integrating Biomedical Text-Mining with Network Biology

Different algorithms have been proposed in order to filter out un-necessary false positive genes and prioritize more important genes related to a particular disease or drug. That was a crucial step in order to decrease the amount of noise and lack of accuracy of the co-occurrences based approaches in text-mining. For the same particular reason it was necessary to consider the fact that biological networks have been found to be comparable with communication social networks [62]. For example biological and communication networks share the scale-freeness property suggesting the necessity of exploiting social network analysis algorithms in studying and analyzing biological networks. Therefore, most attempts to integrate biological data were based on the concept of integrating a protein-protein interaction network together with a list of genes that can be extracted either from a text or is based on an expert point of view [51] [61] [63].

Ozgür A et al. [61] developed a framework to alleviate the problem of being biased toward the initial seed genes. They integrated a text-mining curated protein-protein network that is related to a particular disease with social network analysis centrality measures to predict unknown disease-gene associations. To build the text-mining based protein-protein network the authors started by extracting a list of seed genes that is related to a particular disease from OMIM database. Then they used sentences parsing in order to build syntactic parse tree represents the syntactic constituent structure of a sentence. This tree has been used later to build a dependency tree that captures the semantic relationships between words belonging to a particular sentence. Dependency parses of the sentences that contain at least two seed or neighbor genes were extracted and the shortest path distances between genes were measured. Finally support vector machine classifier has been used to predict possible gene interactions. After building the disease specific protein-protein network the authors considered all the seed genes in addition to their neighbors for further analysis. Finally to prioritize genes related to a particular disease they have used degree, eigenvector, betweenness and closeness network centrality metrics. A brief description of every one of these metrics is provided here:

- **Degree Centrality:** It measures the number of nodes that are connected to a particular node. Nodes with higher degree centrality are more important.
- **Eigenvector Centrality:** It measures how important is the node by measuring the importance of its neighbor nodes and not only counting the number of its neighbors. A node is more central if it is connected to many central nodes.
- **Closeness Centrality:** It measures how close a particular node is to other nodes in the network. The smaller the distance to other nodes the higher its centrality is.
- **Betweenness Centrality:** It measures how important is a particular node in keeping the network connected. In other words, it can be described as the number of shortest paths between pairs of nodes that run through the node of interest.

The possible limitation for this approach is again the high dependency on published work. The authors only considered the protein-protein interactions that have been parsed from literature. Indeed there are many biologically validated protein-protein interaction networks that are available and might have also been involved in the study. Therefore, even with this integrative paradigm, information from microarrays and biological networks are not fully utilized to improve performance. Accordingly, our systematic approach is based on integrating text-mining, microarray expression profiles and biological network in a single paradigm with equal contribution from each data source to the final prediction score.

3.5. Contribution in More Details

From all the mentioned above I sought that there is a great demand to integrate biological data from different resources to elicit better knowledge. More precisely, using microarray data only would not clarify whether drug performance on a specific cell line (say breast cancer) is relevant to all types of diseases. In addition, relying on gene microarray data may fail to match diseases and drugs that are not manifested at the gene expression level. In the suggested approach, these problems were tackled by integrating information from text-mining that extract drug effect without any bias towards a specific cell line. On the other hand, using the text-mining based approaches is limited to find knowledge from published papers and leaving no room for utilizing other data sources like microarrays or biological networks that might result in some implicit predictions. In the suggested approach, these problems were tackled by integrating information from microarray data and biological networks. Finally the states of art integrative approaches are highly dependent on published work. For instance, biologically validated protein-protein interaction network and microarray data are not fully utilized to improve performance. Accordingly, the suggested approach is based on integrating text-mining, microarray expression profiles and biological network with a significant contribution of each data source to the final predictions.

A ranked list of genes for each drug and each disease was first generated by using microarray expression data. Then another ranked list of genes for each drug and each disease were computed by using text-mining together with biomolecular network. The ranked lists for each entity have been integrated into one and representative ranked list that have been used to build the drug-disease connectivity map based on enrichment statistical measure. As shown in **Figure 3.2**, the framework is divided into three phases where each phase is subdivided into many steps. **Phase 1** majorly functions in defining the initial set of genes, **phase 2** majorly functions in data collection, preprocessing and ranked lists building **phase 3** majorly functions in enrichment computing and analysis.

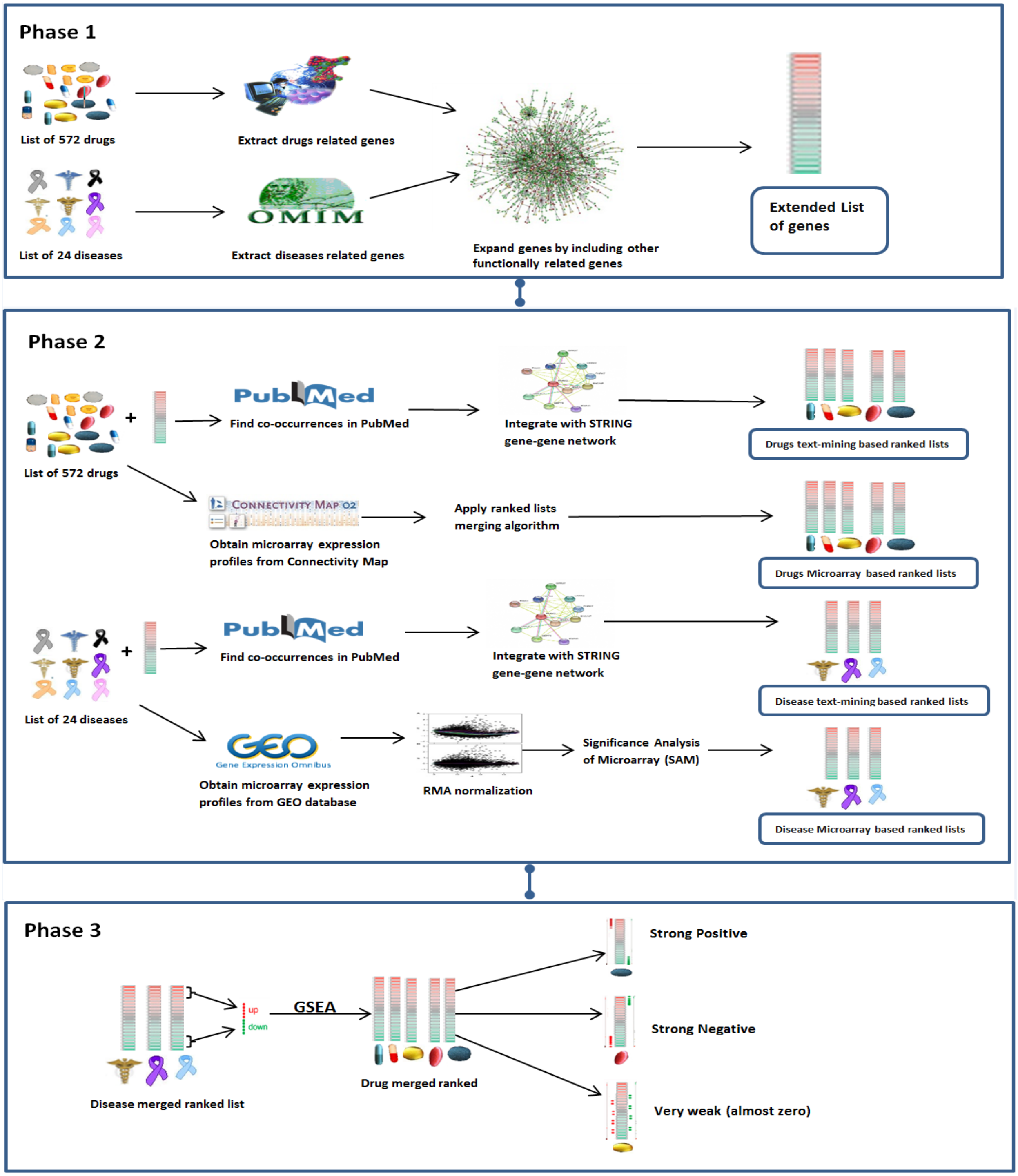
Phase 1 is subdivided into two major steps. In **step1,** genes that are biologically validated to be targets to the set of drugs and/or involved in the pathophysiology of the set of diseases were extracted. These genes formulated the seed genes for the suggested framework. In **step 2**, other genes that are functionally related to these genes were included by using a functional protein-protein network.

Phase 2 is subdivided into three major steps. In **step 1** microarray expression profiles, which are related to the set of diseases and drugs, have been extracted. In addition, the ranked list of genes for disease's and drug's samples were computed by simply comparing the control samples with the diseased or drug treated samples. These genes have been ranked from the most up-regulated (at the top) to the most down-regulated (at the bottom). These ranked lists represent the microarray based ranked lists for drugs and diseases. In **step 2** natural language processing was used to query the PubMed abstracts and found the co-occurrences between each gene and each disease or drug. More precisely, a set of keywords that represent an activation relationship and a set of keywords that represent an inhibition relationship were used together with biological entities to build co-occurrences matrices. A relevancy score was computed to check whether there is an activation or inhibition relationship between each gene and each drug or disease. According to their relevancy scores, genes were ranked from the most positively relevant (at the top) to the most negatively relevant (at the bottom). In **step 3** pertinency scores for genes that had zero relevancy score with any drug or disease were computed. In other words, for each drug or disease, genes with zero relevancy scores were extracted and checked to determine whether they are correlated more with the up-regulated or the down-regulated genes related to that drug or disease. By the end of this step, two ranked lists of genes for each drug and each disease were available for further analysis.

Phase 3 is subdivided into two major steps. In **step 1** each biological entity, drugs or disease, were represented by a single and representative ranked list of genes by combining its two computed ranked lists (microarray based and text-mining/network based) using Borda Merging method. In **step 2** enrichment score between the up/down tag of each disease versus the ranked list of every single drug we found the enrichment score between the up/down tag of each disease versus the ranked list of each drug were computed. These scores were filtered by excluding all the un-significant connections and finally the performance was computed by comparing with a gold standard.

The data sources used in this study and the rationale behind using them are discussed in **chapter 4**.

Figure 3.2 Three Phase's Flow Diagram of the Proposed Paradigm


Phase 1
List of 572 drugs
Extract drugs related genes
List of 24 diseases
OMIM
Extract diseases related genes
Expand genes by including other functionally related genes
Extended List of genes
Phase 2
+
PubMed
Find co-occurrences in PubMed
Integrate with STRING gene-gene network
Drugs text-mining based ranked lists
List of 572 drugs
Connectivity Map 02
Obtain microarray expression profiles from Connectivity Map
Apply ranked lists merging algorithm
Drugs Microarray based ranked lists
+
PubMed
Find co-occurrences in PubMed
Integrate with STRING gene-gene network
Disease text-mining based ranked lists
List of 24 diseases
GEO Gene Expression Omnibus
Obtain microarray expression profiles from GEO database
RMA normalization
Significance Analysis of Microarray (SAM)
Disease Microarray based ranked lists
Phase 3
up
down
GSEA
Disease merged ranked list
Drug merged ranked
Strong Positive
Strong Negative
Very weak (almost zero)

In here, phase 1 starts by extracting all genes related to our set of drugs and diseases. These genes will be further extended by including other functionally related genes using a protein-protein interaction network to result in an extended list of genes. Phase 2 starts by extracting the microarray data for drugs and diseases in addition to extracting the co-occurrences of these entities with genes and specified keywords using PubMed. This data will be further analyzed to result in a microarray and a text mining based ranked list of genes for each drug and each disease. In phase 3 the text-mining based and the microarray based ranked lists will be merged and an enrichment score will be computed between the up/down tags for each disease versus the ranked lists of drugs.

Chapter 4: Data Collection, Normalization and Preprocessing

Since the proposed approach was totally dependent on data with its different forms, this chapter was designed to discuss in details the process of collecting, preprocessing and normalizing this huge amount of data. This chapter describes all different databases that have been used to generate data including gene list, microarray gene expression profiles, text mining co-occurrences, and gene-gene network. In addition, this chapter describes all normalization algorithms and preprocessing steps that have been used for the purpose of generating the closest results to the truth. Finally this chapter digs deeper to indicate the scientific reason beyond choosing every single algorithm or any processing step.

4.1. Generating Gene Lists

For the purpose of generating the gene list to consider in this analysis we have reviewed some concepts regarding drugs and their targets in published literature. We started from a position where we wanted to understand the definition for each of them and therefore make our decision.

A drug can be defined as a chemical substance that, when absorbed into the body of a living organism, alters normal bodily function through directly or indirectly affecting biomolecules called targets. Drug targets are biomolecules that are inhibited, activated or modulated upon drug inhalation, drug exposure, drug injection or oral administration of a drug. The efficiency or the ability of a drug or any other therapeutic substance in accessing a target or changing its normal behavior is called the *druggability* of that particular target. Many experiments have been conducted for the purpose of studying the druggability of human genome.

In an experiment to quantitatively assess the druggability of human genome a conceivable results indicated that there is only 10% of druggable genes in human genome, 10% are involved in disease's pathophysiology and only 5% are both druggable and relevant to disease [64]. Therefore, genes that are related to drug's mode of action or disease's pathophysiology were included in this study. This saves processing time and memory by excluding irrelevant genes from further analysis. Excluding irrelevant genes results in a clearer picture about drugs or diseases molecular action by getting rid of noise that might result from these genes. Finally to guarantee that there is no much information lost, other functionally related genes were included thereby reflecting better understanding for drug mode of action or disease pathophysiology. Later in **chapter 6**, the results show that including this list of gene would not really affect the performance of the predictive paradigm when compared to including the whole list of genes.

To generate drug related genes, drugbank.xml file was downloaded from the DrugBank database [65]**.** A simple Java Xml Parser was used in order to extract all the targets that are related to the drug set. Similarly all disease related genes were extracted from OMIM database [66]. Finally homo_sapiens.interactions.txt file was downloaded from Reactome database [67] to extract all the other genes that are functionally related to the seed list of genes. Therefore, I ended up having 2741 genes as pre-final list of genes. Out of these 2741, only 2379 genes, that have a corresponding probe-set in microarray platform from which microarray data were extracted, were considered in further analysis. Later in the text, this set of genes (the 2379 genes) will be annotated as "Final list of genes".

4.2. Microarray Data

Microarray expression profiles in addition to statistical measures for its meta-analysis are increasing dramatically. For the purpose of satisfying the best knowledge retrieval strategy many facts must be taken in consideration. First of all, a researcher might consider that microarray expression profiles are available in multiple platforms using disparate technologies. Second of all, it would be important to consider that microarray expression profiles are accumulating in public repositories from different laboratories. This would raise an important point before using data directly; different laboratories means different platforms to generate data and most importantly different normalization algorithms to process data. Many experimental results have shown that result's bias and inconsistency might result in incase any of these consideration has been underestimated in the analysis.

4.2.1. Disease's Microarray Expression Profiles

4.2.1.1. Data Collection

Before start generating expression profiles, it has been considered that the consistency of data sets generated using different platforms has been addressed previously and the results were confounding and conflicting [68]. For further assessment, **Severgnini M et al.** [68] designed a standardized strategy to compare transcriptional responses of cell lines using two different platforms. The chosen platforms have been selected from similar technologies where both of them have the same protocol and use the same chemical substances to end up with microarray expression profiles. The results have clearly stated that using microarrays from different platforms will be platform-specific due to many experimental factors including; target preparation and processing, probe design, signal generation and others. This would need a lot of sequence analysis and standardization methods for the purpose of improving comparability between different platforms.

To obtain expression profiles for set of diseases, it was crucial to consider a comprehensive database that store expression profiles covering diverse set of diseases and thereby Gene Expression Omnibus (GEO) was considered. GEO is a public repository that archives and freely distributes microarray, next generation sequencing and other forms of high-throughput functional genomic data submitted by the scientific community [69]**.**

At the beginning, a list of diseases that include almost all types of tumors in addition to other diseases like Diabetes Mellitus, Skeletal Muscle Disorder and Congenital Disorders was considered in the study. To avoid any possible bias that might result from using different platforms; only microarray experiments that have been scanned using Affymetrix Human Genome U133A Array platform (GPL96) were considered. The rationale behind this is because the largest source of drug's microarray expression profiles uses GPL96 to generate data. 24 different diseases including 13 types of cancer have been considered for further analysis. Disease's names, brief description, their microarray GEO accession numbers are all provided in **Table 4.1**. Noteworthy that, GEO offers the opportunity to either extract the raw data files (un-normalized) or the previously normalized data files. Accordingly, raw data files (CLE files) for each experiment were extracted for the purpose of considering one normalization algorithm to process these data sets.

Table 4.1 shows disease's names, GEO number and title of the disease expression profiles we extracted from GEO

Disease Name	GEO Accession	Experiment Title
Lung cancer	GSE10072	Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival
Anemia	GSE16334	Expression data from normal and Fanconi anemia low density bone marrow cells
Breast cancer	GSE15852	Expression data from human breast tumors and their paired normal tissues
Leukemia	GSE22529	Gene expression profiles in CLL
Nevus	GSE3189	Novel genes associated with malignant melanoma but not benign melanocytic lesions
Melanoma	GSE3189	Novel genes associated with malignant melanoma but not benign melanocytic lesions
Rheumatoid arthritis	GSE12021	Identification of inter-individual and gene-specific variances in mRNA expression profiles in the RA SM
Osteoarthritis	GSE12021	Identification of inter-individual and gene-specific variances in mRNA expression profiles in the RA SM
Osteoporosis	GSE7429	Gene Expression of Circulating B Lymphocytes for Osteoporosis
Ovarian cancer	GSE6008	Human ovarian tumors and normal ovaries
Prostate cancer	GSE8218	Gene expression data from prostate cancer samples
Sarcoma	GSE21122	Whole-transcript expression data for soft-tissue sarcoma tumors and control normal fat specimens
Follicular thyroid carcinoma	GSE27155	Human thyroid adenomas, carcinomas, and normals
Papillary thyroid carcinomas	GSE27155	Human thyroid adenomas, carcinomas, and normals
Diabetes mellitus	GSE25724	Expression data from type 2 diabetic and non-diabetic isolated human islets
Liver cancer	GSE2109	Expression Project for Oncology (expO)
Colon cancer	GSE2109	Expression Project for Oncology (expO)
Congenital disorder	GSE8440	Expression data from Congenital disorders of Glycosylation type-1 patients (CDG-I)
Glioblastoma	GSE2485	Gene expression of pseudopalisading cells in human glioblastoma
Huntington's disease	GSE1751	Human blood expression for Huntington's disease versus control
Hutchinson–Gilford Progeria Syndrome	GSE3860	Comparison of Hutchinson–Gilford Progeria Syndrome fibroblast cell lines to control fibroblast cell lines
Polycystic ovary syndrome	GSE5090	PCOS patients vs control subjects
Duchenne muscular dystrophy	GSE3307	Comparative profiling in 13 muscle disease groups
Muscular Dystrophies	GSE3307	Comparative profiling in 13 muscle disease groups

4.2.1.2. Data Normalization

In microarray technology, an mRNA molecule or a gene is represented on an array by a probe set composed of 10-20 probe pairs. Each probe pair is composed of a perfect match (*PM*) probe, a section of the mRNA molecule of interest, and a mismatch (*MM*) probe that is created by changing the middle base of the *PM* with the intention of measuring non-specific binding [70].

After scanning the arrays hybridized to labeled RNA samples of interest (drug, disease, and control), intensity values PM_{ij} and MM_{ij} are recorded for arrays $i=1,....m$ and probe pairs $j=1,....,n$ for any given probe set. To define a measure of expression representing the amount mRNA species it is necessary to summarize probe intensities for each probe set. Many model-based approaches have been proposed to normalize and better represent the microarray expression data. Noting that, the outputs from these approaches are highly disparate as shown in **Figure 4.1**. Therefore, one can consider using a single normalization algorithm in case of running a systematic analysis that includes different data files from different laboratories. Many studies suggest that subtracting *MM* as a way of correcting for non-specific binding is not the best way to normalize microarray expression data. Empirical results demonstrate that mathematical subtraction does not translate to biological subtraction. In fact, *MMs* are found to be a mixture of probes for which (i) the intensities are largely due to non-specific binding and background noise and (ii) the intensities include transcript signal just like the *PMs*. Log scale robust multi-array analysis (RMA) was one of the most robust models [70]**.**

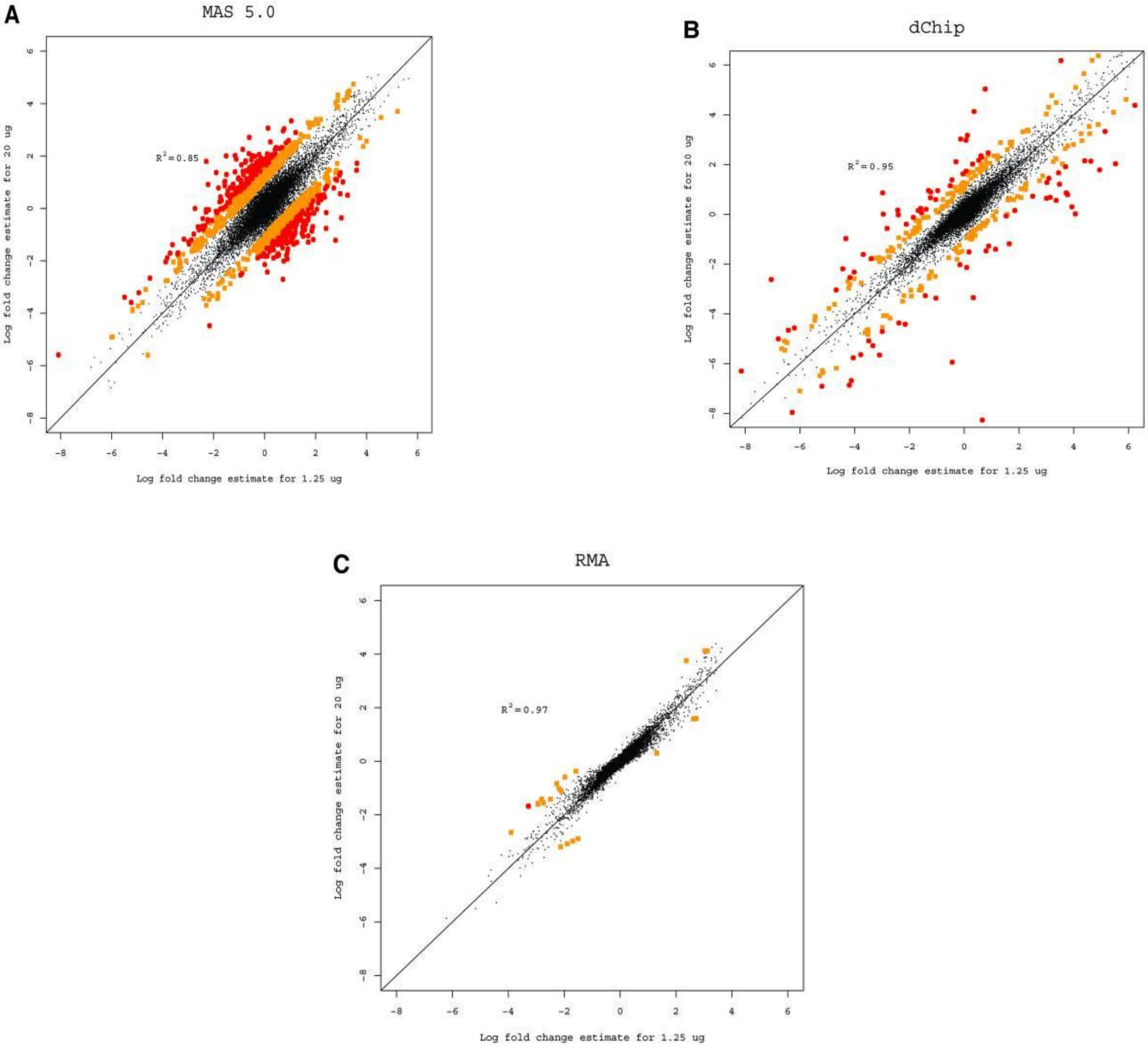
RMA [70] assumes that each array has a common mean background level and it adjusts *PM* intensities to remove background effect using a transformation function. Then it normalizes the arrays using quintile normalization algorithm. Finally the expression measure for each probe set will be background-adjusted, normalized and log transformed. *PM* intensities follow a linear additive model

$$T(PMij) = ei + aj + \varepsilon_{ij} \qquad \textbf{Equation 4.1}$$

For *i=1…..m* (number of arrays), *j=1….n* (number of probe-pairs) and *z=1…..p* (number of probe-sets). Where *aj* is log scale affinity effect for probes *j=1…….J*, *ei* representing the log scales expression level on arrays *i=1…….I*, and *εij* representing an independent identically distributed error term with mean zero.

In this thesis, RMA normalization algorithm has been used in order to normalize the microarray data. This is because RMA has been proven to have the lowest False Discovery Rate, among all other normalization algorithms, when it comes to predicting differentially expressed genes [70]. The CLE files from each experiment have been normalized independently using RMAExpress software (http://rmaexpress.bmbolstad.com/). After normalization, the expression intensities for genes included in the final list of genes were extracted. The average intensities for genes that are represented with more than one corresponding probe set were computed to represent the expression value for that gene.

Figure 4.1: Differentially Expressed Genes Using Different Normalization Algorithms



This figure shows the fold change estimates of gene expression for a data normalized with three different normalization algorithms. Circles and squares represent genes demonstrating 2 to 3 fold change. (Adopted from [70])

4.2.2. Drug's Microarray Expression Profiles

The Connectivity Map (CMap) [10] project was one of the richest projects regarding the availability and diversity of drug treated microarray expression profiles. In addition, most of the experiments have been done using the GPL96 platform indicating better comparability with the disease's microarray expression profiles. The CMap website (http://www.broad.mit.edu/cmap/) contains a collection of genome-wide transcriptional expression data for different cells that have been treated with different drugs at different concentrations. As of **January 10th/2012** CMap website was having a collection of 6100 ranked list of genes that are related to 1309 unique compounds. To obtain the drug's microarray based ranked list "rankMatrix.txt" data file was downloaded with its associated annotation file "cmap_instance_02.xls" from CMap website. Probe-sets ranks related to each chemical instance were extracted and again the rank for each gene was computed by averaging the ranks for all corresponding probe-sets. Finally for each drug, the gene rank score was normalized to be from 1 to 2379.

4.3. Generating Text Mining Data

For the purpose of running text mining experiment, it was crucial to consider a comprehensive database that almost covers and stores all papers in the biomedical field. The reason beyond that is the need to find all possible co-occurrences between drugs, diseases and genes to formularize an idea about the connectivity between them. All the required properties that can best fit the model were found in PubMed database.

PubMed is a service of the US National Library of Medicine that provides access to abstracts for medical, nursing, dental, veterinary, health care and preclinical sciences journals. Furthermore, PubMed provides many services that make it easier to run systematic analysis, for example: (i) it provides a web service that enables users to automatically run queries on their server, and (ii) it provides a service to automatically include all synonyms for the entered biological entities. This property is with no doubt a very distinct property because biological entities tend to have different synonyms and publications do not consider a common name for a particular entity.

4.3.1. Obtaining the Drug-Gene and Disease-Gene Co-occurrences

Using PubMed I searched for the co-occurrences between two entities and a keyword at a time. More specifically, each query contains a drug or a disease, a gene and a keyword. These keywords were divided into two major sets; a set of keywords representing activation relationship between drug/disease and a gene (activate; agonist; cofactor; synthesis; trigger; induce) and another set representing inhibition relationship between a drug/disease and a gene (antagonist; block; deactivate; inactivate; inhibit; suppress). In a previous study these keywords have been curated manually and validated to have a role in reflecting association relationships between different biological entities [43]. This step has resulted in two lists of occurrences between the mentioned entities once with activation relationships and once with inhibition relationships.

Noteworthy, that all synonyms that are suggested by PubMed were considered while querying the database. For example when considering the abstracts that contain topoisomerase I gene, prostate cancer and an activation keyword, the entered query would be “Top1 AND prostate cancer AND activation AND Homo Sapiens” and in PubMed it would be “Top1[All Fields] AND ("prostatic neoplasms"[MeSH Terms] OR ("prostatic"[All Fields] AND "neoplasms"[All Fields]) OR "prostatic neoplasms"[All Fields] OR ("prostate"[All Fields] AND "cancer"[All Fields]) OR "prostate cancer"[All Fields]) AND activation[All Fields] AND ("humans"[MeSH Terms] OR "humans"[All Fields] OR ("homo"[All Fields] AND "sapiens"[All Fields]) OR "homo sapiens"[All Fields])”.

4.3.2. Obtaining the Gene-Gene Network

Since there were few gaps in the ranked list of genes obtained from text mining, a gene-gene network was used to fill out these gaps. For this purpose, it was important to consider a comprehensive database that has high sensitivity in detecting these gene-gene interactions; thereby STRING database [46] was used.

STRING is a database and a web-tool dedicated to protein-protein interactions, including both physical and functional interactions. STRING is a meta-resource database that weights and integrates information from numerous sources. It uses many experimental repositories, computational prediction methods and biomedical text collection and augments those into a single confidence score. As a text mining source, STRING parses a large body of scientific texts and all abstracts from PubMed. The authors searched for statistically significant co-occurrences of gene names, and used Natural Language Processing to included possible semantically related genes.

In this thesis, the final list of genes were dropped in STRING database and text mining based interactions with confidence score >0.5 were considered. This confidence score has been defined by the authors [46] as being the minimum score to be considered as high confidence based on intensive experimentations and cross validation analysis.

Chapter 5: Building the Connectivity Map

For the purpose of building the connectivity map using GSEA, it was crucial to build ranked list of genes based on their relevancy to a particular disease or drug. This chapter describes the process of obtaining the microarray based and text mining based ranked list of genes for drugs and diseases. In addition, it describes the rank merging method that have been used in order to integrated the text-mining based and the microarray based ranked list of genes for every single entity. Furthermore, this chapter discusses in details and with mathematical explanation the GSEA. It also describes all possible combinations and experiments done in order to generate different connectivity maps for the purpose of comparing performances of these combinations. Finally this chapter describes the process of performance evaluation with a detailed description for process of generating the gold standard.

5.1. Obtaining Microarray Based Ranked Lists

As mentioned in **chapter 2**, microarray needs to have a set of replicas representing query samples and a set of replicas representing control sample. Many statistical measurements have been proposed in order to compare the expression value for every single gene in two biological conditions and finally results in a score to represent the level of differential expression for that gene. Because microarray data for drugs and diseases have been generated from two different sources in two different forms, it was unfortunate to use two different methods to generate the ranked lists for those drugs and diseases.

5.1.1. Obtaining the Microarray based Ranked List of Genes for each Drug

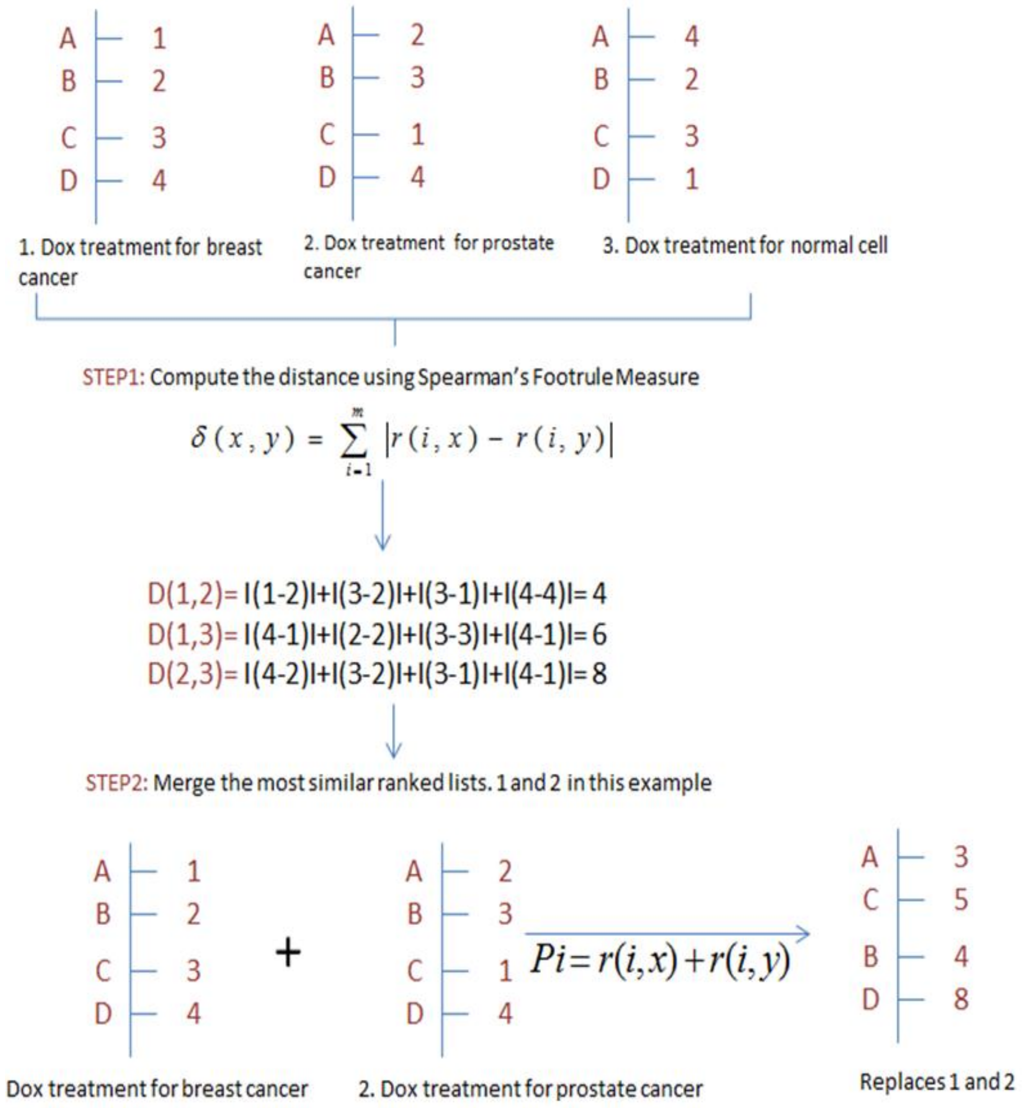
Drug treated microarray data was available in the form of ranked lists of genes instead of query/control samples traditional data form. The authors in CMap project [10] have already compared each profile to its corresponding control by using z-score. Therefore, all genes were ranked according to their differential expression relative to the control from the most up-regulated on the top to the most down-regulated on the bottom and posted the ranks in their website as described in previous chapter.

Having 6100 expression profiles for 1309 compounds indicates the availability, in average, of 5 expression profiles to represent each compound. Therefore, it was necessary to use an effective and an already validated method to merge these profiles and get a single and representative ranked list for each compound. For this purpose, a previously described merging technique in [15] was used. This hierarchical merging procedure that is explained in **figure 5.1** started by measuring the Spearman's Footrule similarity measure [59] between all the instances belonging to one drug according to the following formula:

$$\delta(x,y)=\sum_{i=1}^{m}\left|r(i,x)-r(i,y)\right| \qquad \textbf{Equation 5.1}$$

Where *m* represents the number of probe-sets, *x* and *y* represent two different instances that belong to a single drug, *r(i,x)* represents the rank of gene i in instance *x* and *r(i,y)* represents the rank of gene i in instance *y*.

Figure 5.1 Example on Gene Ranked Lists Merging


A 1
B 2
C 3
D 4
1. Dox treatment for breast cancer
A 2
B 3
C 1
D 4
2. Dox treatment for prostate cancer
A 4
B 2
C 3
D 1
3. Dox treatment for normal cell
STEP1: Compute the distance using Spearman's Footrule Measure
$\delta(x, y) = \sum_{i=1}^{m} |r(i, x) - r(i, y)|$
D(1,2)= I(1-2)I+I(3-2)I+I(3-1)I+I(4-4)I= 4
D(1,3)= I(4-1)I+I(2-2)I+I(3-3)I+I(4-1)I= 6
D(2,3)= I(4-2)I+I(3-2)I+I(3-1)I+I(4-1)I= 8
STEP2: Merge the most similar ranked lists. 1 and 2 in this example
A 1
B 2
C 3
D 4
+
A 2
B 3
C 1
D 4
$Pi = r(i,x) + r(i,y)$
A 3
C 5
B 4
D 8
Dox treatment for breast cancer
2. Dox treatment for prostate cancer
Replaces 1 and 2

This figure shows an example of merging different ranked lists belonging to a particular instance using Borda merging method. In step one the algorithm will compute the Spearman's Footrule Measure. Then the algorithm would start by merging the two closest ranked lists using Borda Merging method. The algorithm will keep on running in a hierarchical way until one and representative ranked list remain to replace that drug.

Then the method merges the two lists that are the most similar to each other in a hierarchical way following the Borda Merging Method. For each *gene i*, a score *pi* was computed from merging instance *x* and instance *y* according to the following

$$pi = ri(x) + ri(y) \qquad \textbf{Equation 5.2}$$

where *r* represents a previously defined function that has been considered as a summation function. Finally all genes have been ranked according to their *pi* value in an ascending order.

5.1.2. Obtaining the Microarray based Ranked List of Genes for each Disease

To obtain the ranked list of genes (according to their differential expression) for each disease, SAM statistical measure was used [20]. SAM identifies genes with significant changes between healthy and diseased samples by assimilating a set of gene-specific t-tests. Each gene is assigned a score that depends on the ratio between the change in that gene expression and the standard deviation across the repeated measurement for that gene. Later on, genes with a score greater than specific threshold are selected as being potentially significant. The relative difference score *d(i)* is computed for every single gene according to the following formula

$$d(i) = \frac{\bar{x}h(i) - \bar{x}d(i)}{s(i) + so} \qquad \textbf{Equation 5.3}$$

Where $\bar{x}h(i)$ and $\bar{x}d(i)$ are the average levels of expression for $gene\, i$ in the healthy and the diseased samples, respectively, $s(i)$ is the standard deviation of repeated expression measurements for gene i and finally so is a positive constant that has been added to ensure that the variance of *d(i)* is independent on the value of gene expression. Using this score, all the genes have been ranked from the most positive (at the top) to the most negative (at the bottom).

5.2. Obtaining Text-Mining Based Ranked Lists of Genes

In chapter 4, the process of generating three different co-occurrences matrices was described. These three matrices were used to build text-mining based ranked list of genes. The first two matrices have been generated by querying PubMed for co-occurrences between a drug/disease, a gene and activation keywords. The second two matrices have been generated by querying PubMed for co-occurrences between a drug/disease, a gene and inhibition keywords. The same have been applied for the third two matrices but with finding co-occurrences without using any keyword.

Further analysis and normalization techniques have been applied on these matrices to get the desired ranked lists of genes. All of that have been done based on the assumption that if a gene co-occurs more frequently with a drug/disease and activation keywords then it is more probable that this drug/disease triggers activation action on that gene. On the other hand, if a gene co-occurs more frequently with a drug/disease and inhibition keywords then it is more probable that this drug/disease triggers inhibition action on that gene. The reason for integrating these ranked lists with microarray ranked lists is based on the assumption that building a ranked list of genes from other source (text mining in this case) might have the potential of filtering and reprioritizing the microarray based ranked lists. Meanwhile microarray based ranked lists might have the potential of filtering and reprioritizing the text mining based ranked lists. This will indeed result in integrated ranked lists that contain information from both sources for the purpose of getting better accuracy measures.

5.2.1. Generating In-Complete Ranked Lists of Genes

Because searching for co-occurrences is vulnerable to false positives, co-occurrences between drugs/diseases and genes without using any keyword were also generated. This is to make a more precise idea about the general co-appearance of those entities together. Later on, relevancy score was computed to make a final judgment about the relationship between disease/drug and a gene.

The rest of this section explains the role of finding the relationship between drugs and genes; meanwhile it is important to consider that the same technique has been applied for finding the relationship between diseases and genes. Assume that *AC (i, j)* and *IN (i, j)* represent two matrices with *i* genes and *j* drugs obtained by using activation and inhibition keywords respectively. Assume that *GE (i, j)* represents a matrix with *i* genes and *j* drugs obtained by finding the co-occurrences without using any keyword. Then the relevancy score (*RE*) between gene *x* and drug *y* is obtained according to the following

$$RE(x,y) = \frac{AC(x,y)}{GE(x,y)} - \frac{IN(x,y)}{GE(x,y)} \qquad \textbf{Equation 5.4}$$

Noting that, this score was used to investigate whether the gene is more enriched to that drug with the activation keywords or the inhibition keywords. After making a judgment by using this formula, another parameter was added to give the priority for genes with more co-occurrences with a particular drug to move to extremes (either to get more positive *RE* score or more negative *RE* score). Thus the normalized relevancy score (*NRE*) would be

$$NRE(x,y) = RE(x,y) + \lambda \qquad \textbf{Equation 5.5}$$

Where $\lambda = \frac{GE(x,y)}{\max\limits_{z=1:i}(GE(z,y))}$ if *RE(x, y)* is positive and $\lambda = -\left(\frac{GE(x,y)}{\max\limits_{z=1:i}(GE(z,y))}\right)$ if *RE(x, y)* is negative and equals to zero otherwise. This normalized score has been used to compute pertinency score for genes with zero relevancy scores regarding any drug or disease as shown in the next step.

5.2.2. Integrating Network Biology for Complete Ranked List of Genes Generation

The co-occurrences based text mining approach, used here, has resulted in zero co-occurrences for some genes in relevant to particular drugs or diseases. That is to say, some genes have never been published in literature with some drugs or diseases. Accordingly, these genes were having zero relevancy scores with those drugs or diseases. To assign a score for these genes, it was crucial to consider if those are related to over-expressed genes or downregulated genes, in relevant to a particular drug/disease, by exploiting the concept of network biology. This network was extracted from STRING database where gene lists represents the nodes and an edge links two genes if they have an interaction confidence with more than 0.5.

To check whether genes with zero *NRE* are more correlated with overexpressed or down-regulated genes, a pertinency score (*PS*) was computed. This score is computed based on the correlation of that gene with the positively relevant genes, negatively relevant genes and their non-zero relevancy scores. Therefore, the pertinency score *PS(x, y)* for gene *x* in terms of drug/disease *y* is computed according to the following formula:

$$PS(x,y) = \frac{\sum_{i=1}^{n} Corr(x,i) * NRE(i,y) + \sum_{j=1}^{m} Corr(x,j) * NRE(j,y)}{\lambda}$$ **Equation 5.6**

Where gene *x* is any gene that has zero relevancy score with drug/disease *y*, *n* represents the number of positively relevant genes in terms of drug/disease *y*, *m* represents the number of negatively relevant genes in terms of drug/disease *y*, *Corr (x,i)* is the weight of the edge in the gene-gene network obtained from STRING database between gene *x* and gene *i*, *NRE(i,y)* is the normalized relevancy score of gene *i* in terms of drug/disease *y* and λ is a positive constant that is equal to $\lambda = \sum_{i=1}^{n+m} \left| NRE(i,y) \right|$

Noteworthy that, because *PS* for genes with non-zero relevancy scores was set to *NRE*, parameter λ was added to guarantee that genes with zero relevancy scores would not take more positive or negative *PS* than genes with non-zero relevancy scores. In other words, the priority to move to extremes was given for genes with non-zero *NRE*. This process has been repeated in an iterative way for 5 times until all genes obtained a non-zero *PS*. Finally all genes have been ranked according to their *PS*.

Pertinency score ranges in its value from a high positive score but with less positivity than the minimum positive *NRE* to a high negative score but with less negativity than the maximum negative *NRE*. High positive and high negative scores indicate a better probability for a particular gene to be positively or negatively, respectively, relevant to a particular drug or disease. **Figure 5.2** shows the different parameters that can play a role in finding a pertinency score for genes with zero *NRE*. In addition, it shows three possible scenarios where pertinency score for a particular *gene z* might be high negative, high positive or close to zero in relevant to *drug A*, *drug B*, and *drug C* respectively. The two major parameters that affect the pertinency score for *gene z* are its weighted correlation with positively relevant or negatively relevant nodes in addition to their non-zero relevancy scores. Up to this point two ranked lists, microarray based and text-mining based lists, were available for each drug and each disease.

5.3. Obtaining Integrated Based Ranked Lists

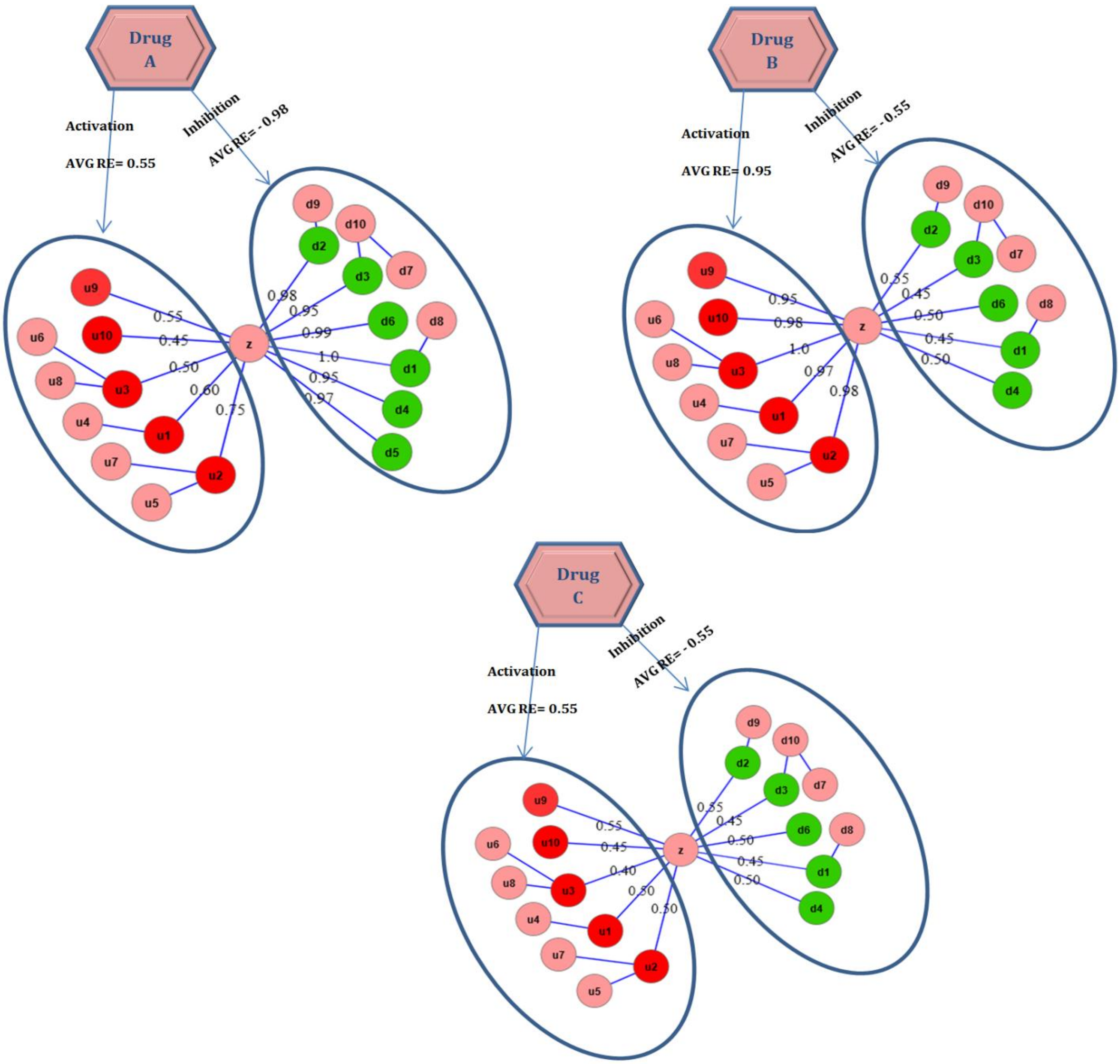
After finding the microarray based and the text mining based ranked lists, they were merged in a representative ranked list of genes for drugs and diseases. Borda merging procedure was implemented using two functions; the geometric mean and the arithmetic mean.

5.3.1. Geometric Mean

The geometric mean is a type of mean or average, which indicates the central tendency or typical value of a set of numbers computed by taking the *nth* root of the cross product of *n* numbers. For each drug and disease we computed a score for each gene that is equal to geometric mean of both scores; the text mining based and the microarray based, according to the following formula

$$Geo\_Mean = \sqrt{Mic\_Score * Text\_Score} \qquad \textbf{Equation 5.7}$$

Figure 5.2. Three Possible Scenarios for Pertinency Score Value Together with Parameters Affecting this Value.



This figure shows different parameters and scenarios in defining a pertinency score for genes with zero relevancies related to a particular drug/disease. In the three cases, *d* nodes represent genes that have been found to be negatively relevant; *u* nodes represent genes that have been found to be positively relevant and *z* node represents a gene that has been found to have a zero relevancy. Red and green nodes represent the genes that have been found to be positively or negatively relevant, respectively, and they are directly interacting with *gene z* based on STRING database. Noting that, the two major parameters that affect the pertinency score for *gene z* are; its weighted correlation with *u* nodes or *d* nodes in addition to their non-zero relevancy scores with a particular drug. The more correlated is *gene z* with *u* nodes, and the more *u*'s are positively relevant to a particular drug the more chance that *gene z* will have a positive pertinency score with that drug. The more correlated is *gene z* with *d* nodes and the more *d* nodes are negatively relevant to a particular drug the more chance *gene z* will have a negative pertinency score with that drug. Then from the example above we assumed that *gene z* would most probably have high negative score, high positive score or a close to zero score regarding *drug A*, *drug B* or *drug C* respectively.

5.3.2. Arithmetic Mean

The arithmetic mean is a type of mean or average, which indicates the central tendency of a collection of numbers taken as the sum of the numbers divided by the size of the collection. For each drug and disease we computed a score for each gene that is equal to arithmetic mean of both scores; the text mining based and the microarray based, according to the following formula

$$Arith\_Mean = \frac{Mic\_Score + Text\_Score}{2} \qquad \textbf{Equation 5.8}$$

Finally all genes have been ranked according to their *Arith_Mean* score.

5.4. Computing Enrichment Scores

To connect drugs and diseases based on the final ranked list of genes in both of them; a modified version of GSEA, previously described in [2], was used**.** More specifically, the enrichment score between a specific disease signature composed of up-tag and down-tag and a ranked list of genes for a drug were computed. For each tag (up or down), enrichment score (ENS) was computed by computing two variables es_{up} and es_{down} independently.

For simplicity I will be talking about computing es_{up} and the exact same thing has been applied to es_{down}. In addition, I will explaining the rule of finding the connection between one disease signature and one drug ranked list of genes and the same is applicable for global connectivity measurements. Let *n* be the total number of genes in the reference drug ranked list (in our case *n*=2379 or *n*=22283), let *s* be the number of genes in the selected up-tag disease signature and let *S* represents the set of up-tag genes (in our case *s*=25 or *s*=250). First a vector *V* of the position (*1..... n*) of each gene in the disease signature in regards the ranked list of a drug is constructed. Those values were sorted in an ascending order such that *V(j)* is the position of gene *j* where gene *j* ∈ *S* for *j =1, 2, ... s*. Then two parameters a_{up} and b_{up} are computed as the following

$$a_{up} = \max_{j=1, j\in S}^{s} \left[\frac{j}{s} - \frac{V(j)}{n} \right] \qquad \textbf{Equation 5.9}$$

$$b_{up} = \max_{j=1, j\in S}^{s} \left[\frac{V(j)}{n} - \frac{(j-1)}{s} \right] \qquad \textbf{Equation 5.10}$$

Later on, the following was considered to compute es_{up}:

If $a_{up} > b_{up}$, then es_{up} is set to a_{up}. Else es_{up} is set to $-b_{up}$.

As mentioned earlier, $\mathrm{es}_{\mathrm{down}}$ is computed in the same way to end up having es_{up} and es_{down} for each disease signature against a ranked list of genes belonging to a particular drug. Finally *ENS* is set to zero if es_{up} and es_{down} have the same algebraic sign, otherwise $ENS= es_{up} - es_{down}$. See supplementary material in [10] for more details.

5.5. Performance Analysis

After building the initial connectivity map, it was crucial to investigate two major metrics to analyze the performance of the proposed paradigm. Firstly, it was important to compute the significance of the obtained associations and get rid of insignificant ones. Secondly, it was important to compare prediction results with real life gold standard to check whether results were or were not consistent with some already approved and well known associations.

For the first task, one sample t-test was used to check whether the predicted association scores are significantly higher or significantly lower than randomly generated scores. For the second task, a gold standard of previously known drug-disease associations was collected and compared with the set of the predicted associations to compute some performance measures that are described later in this chapter.

5.5.1. Random Samples Generation

To run the one sample t-test and thus determine the significance of the observed associations it was necessary to compute some randomly generated associations and compare them with the observed ones. For this reason, 200 random disease signatures were generated each of 50 genes length. This has been done by randomly selecting 50 genes from the whole set of genes and the process was repeated for 200 times. In each run, 25 genes were randomly selected to represent the set of up-regulated genes and the other 25 were representing the set of down-regulated genes. Enrichment scores have been computed for each randomly generated up/down tag versus the ranked lists of drugs. Therefore, each drug was having a set of 200 randomly generated enrichment scores (RGES) with the randomly generated signatures. The RGES relevant to each particular drug were tested for normality and all of these were satisfying normal distributions.

5.5.2. Significance Analysis

One sample t-test has been used to test for the significance of association and to check whether the observed association score was significantly higher or lower than the randomly generated scores. The reason for choosing t-test but not z-test is that because the variance of the population is not known. Several steps have been done in order to obtain the results and conclusions for t-test for every single association score.

5.5.2.1. Null (H_0) and the Alternative hypothesis (H_1)

H_0: $\mu_{observed} = \mu_{random}$

H_1: $\mu_{observed} \neq \mu_{random}$

Because the purpose of the test was to check if the randomly generated scores are significantly different than observed scores.

5.5.2.2. State the decision rule

In this step, α was set to 0.05 and degree of freedom was set to *n-1* where *n* represents the number of randomly generated samples (200 in this case). Critical value was computed using the value of α together with the degrees of freedom. Noteworthy that the confidence interval is computed according to the following formula

$$CI = \mu_{random} \pm t_{\frac{\alpha}{2}, n-1} \left(\frac{s}{\sqrt{n}} \right)$$

Equation 5.11

Where μ_{random} represents the average of the randomly generated score, s represents the standard deviation of the randomly generated scores and n represents the number of randomly generated samples. Thus if the t-score for a particular prediction score is greater than the positive critical value or is less than the negative critical value, then the null hypothesis is rejected and the prediction score is assumed to be significant. Mathematically, the null hypothesis is rejected where $t > t_{\frac{\alpha}{2},n-1}$ or $t < -t_{\frac{\alpha}{2},n-1}$ such that

$$t = \frac{\mu_{random} - \mu_{observed}}{s/\sqrt{n}}$$ **Equation 5.12**

Where $\mu_{observed}$ represents the actual association score, μ_{random} represents the average of the randomly generated score, s represents the standard deviation of the randomly generated scores and n represents the number of randomly generated samples.

5.5.3. Extracting Gold Standard

In this thesis, the performance of the proposed paradigm was tested at two different levels where each of which was having its own gold standard.

In the first level, I wanted to show that using integrated ranked list of genes, as input to GSEA, would output better results than considering ranked lists of genes from microarray or text-mining independently. Furthermore, I wanted to investigate the effect of considering subset of genes in this analysis rather than considering the whole set of genes.

For this purpose two different data sources were used to build the gold standard. The first was PolySearch **(http://wishart.biology.ualberta.ca/polysearch/index.htm)**

[43] because it computes a relevancy score that does not only rely on co-occurrences but it defines a pattern recognition system instead. From PolySearch, connections with non-zero relevancy scores were obtained by simply inputting disease's names one by one in the server and retrieve all related drugs from that tool. In here, the automated disease synonyms function was on for all diseases. Finally, all drugs with corresponding synonym in the drug list, in this study, were considered. To avoid being biased toward text-mining approach, another set of associations that have been previously used by **Gottlieb A et al** [55]**,** were included**.** The authors used the drug-disease associations that are stored in the registry of federally and privately supported clinical trials (RFPSCT) conducted around the world (http://clinicaltrials.gov/).

In the second level, I wanted to compare the proposed paradigm with one approach from each of the previously mentioned categories for drug repositioning. More precisely, I wanted to compare the proposed method with a text-mining based approach, a microarray based approach and a similarity based approach.

For this purpose, the proposed method was compared with PolySearch [43], ConnectivityMap [10] and PREDICT [55] since those are one of the most robust methods in text-mining, microarray and similarity based categories, respectively. The gold standard was built according the following scenario. Let *PS*, *CM* and *PR* denote three different drug-disease association sets that were generated using PolySearch, ConnectivityMap and PREDICT methods, respectively. Let $G= \{PS\} \cup \{CM\} \cup \{PR\}$ and let G_X denote the set *G* excluding set *X*. Therefore the proposed paradigm was compared with PolySearch, ConnectivityMap and PREDICT, independently, using three different gold standards; G_{PS}, G_{CM} and G_{PR}, respectively.

Noteworthy that in both levels, the gold standard should not be mistaken for true confirmed drugs with therapeutic or toxicological values. Instead, it provides an unbiased disease-drug associations list for performance evaluation purposes only.

5.5.4. Performance Measures

Finally after considering significant connections, the predicted set of associations was compared with the gold standard using different performance test measures. Before introducing these measurements, it is worthy introducing several terms that are commonly used along with these measurements namely; true positive (**TP**), true negative (**TN**), false negative (**FN**), and false positive (**FP**). Let us assume that MT represents the set of associations predicted using the suggested model that is going to be compared with gold standard G. Then, if an association (Dr_i, Ds_J) ∈ MT ∧ (Dr_i, Ds_J) ∈ G then it is considered to be TP. Similarly, if an association (Dr_i, Ds_J) ⋹ MT ∧ (Dr_i, Ds_J) ⋹ G then the test result is TN. Both TP and TN suggest a consistent result between the gold standard and the proposed paradigm. However, if (Dr_i, Ds_J) ∈ MT ∧ (Dr_i, Ds_J) ⋹ G, then the association is a FP. Conversely, if (Dr_i, Ds_J) ⋹ MT ∧ (Dr_i, Ds_J) ∈ G then the association is a FN. Both FP and FN indicates that the test results are opposite to the actual set of associations.

Using these terms, performance measurements were done according to the following: (1) **Sensitivity:** is the percent of correctly identified drug-disease connections and equal to (TP/ (TP + FN)). Sensitivity of 1 means that each association in the gold standard was correctly identified as so in the prediction algorithm. (2) **Specificity:** is the percent of correctly identified non drug-disease connections and equals to (TN/ (TN+ FP)). A specificity of 1 means that every association that is not in the gold standard, was correctly predicted to be so in the prediction algorithm. (3) **Accuracy**: is the proportion of correctly predicted drug-disease associations and equal to (TP + TN)/ (TP + TN + FP + FN). An accuracy of 1 means that all predicted associations are actually in the gold standard and all the non-predicted associations are actually not in the gold standard. (4) **Precision**: is the probability of correct positive drug-disease predictions and equals to (TP/ (TP + FP)). A precision of 1 means that every association predicted by the algorithm, does actually belongs to the gold standard.

Results for these performance measures and further biological discussion of the obtained results are discussed in **chapter 6**.

Chapter 6: Results and Discussion

The enrichment technique discussed in **chapter 5** takes a set of overexpressed genes (up-tag), a set of down-regulated genes (down-tag) and finds the enrichment score between these signatures and the ranked list of genes for the set of drugs as shown in **figure 3.2**. A high negative score between a specific drug and a specific disease indicates that this drug might be used for the treatment of that disease. From the other hand, a high positive score indicates that both the drug and the disease share the same molecular action and thus this drug might have the same pathological effect of that disease.

6.1. Evaluating the Connectivity Map

Four different connectivity maps were generated using the enrichment analysis measure. The first two have been generated by using the microarray based drug ranked gene lists with the microarray based disease signatures once by using the 2379 genes and once by using the whole set of 22283 probe-sets namely; Mic_2379 and Mic_22283 respectively. The third one has been generated by using the text mining based drug ranked gene lists with the text mining based disease signatures namely Text-Mining. The forth one has been generated using arithmetic mean since it outperformed the geo-metric mean one. Results are shown in **Figure 6.1**.

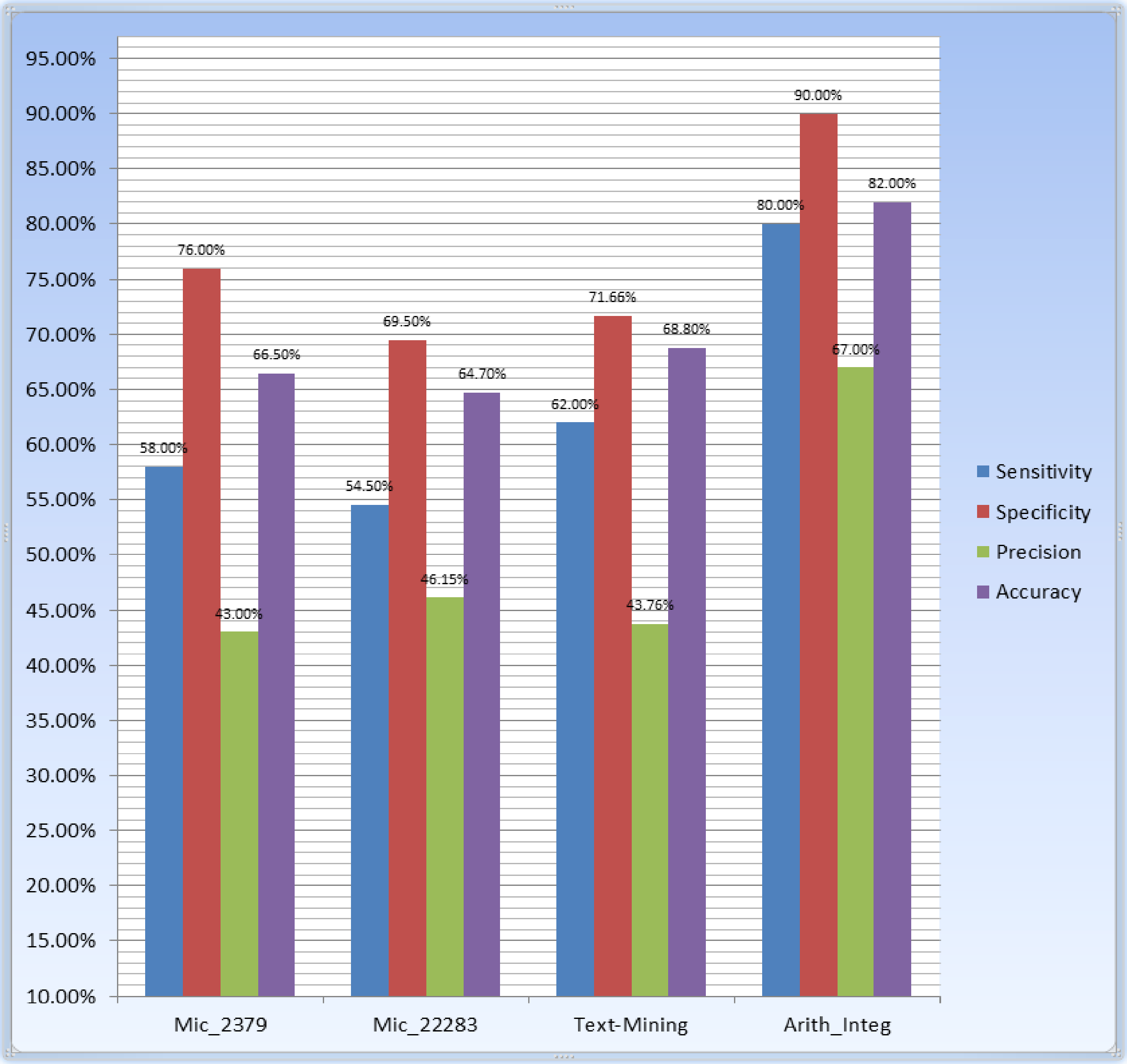
As indicated previously, 2%-4% of the gene list was enough to represent the up-tag and the down-tag [15] [10] and thus we decided to use most 25 up-regulated genes as up-tag and the most 25 down-regulated genes as down-tag. To further assess the usability of this percentage, this experiment was repeated on different lengths from (15-50) in an interval of 5. The results show that considering more than this percentage would result in high false positive rate. On the other hand, considering less than this percentage would result in very small number of predictive associations.

Many things can be inferred by looking at **Figure 6.1.** First of all it has been found that all methods that have been used to build the connectivity map have reported a low precision or positive predictive score value. This might be due to two reasons. The first is the unsupervised nature of the GSEA that tends to build unrealistic associations between biological entities. The second might be related to the nature of the gold standard obtained in this experiment. Firstly, the gold standard was not big enough to report the actual associations between the 406 drugs and the 24 diseases in the dataset. Secondly, the gold standard came from two different sources that have totally different nature than the GSEA. Accordingly the false positives were way higher than the true positives even with one of the most popularly used methods; the microarray expression profiles based method.

6.2. Comparing Different Connectivity Maps

Figure 6.1 shows that using just a sub-set of genes that are related to the set of drugs and diseases does not really harm microarray performance results. For instance, the connectivity map that has been generated using just 2379 genes shows comparable performances with the one that has been generated using the whole expression profiles with 22283 probe-sets. Indeed Mic_2379 shows better sensitivity and specificity when compared to Mic_22283.

Figure 6.1 Performance results for four different connectivity maps



This figure shows the Sensitivity, Specificity, Precision and Accuracy for four different methods that have been used to generate drug-disease connectivity map. In here Mic_2379 and Mic_22283 represents connectivity maps that have been generated using microarray data only once by considering 2379 genes and once by considering the 22283 probe-sets. Text-Mining represents the connectivity map that has been generated using text-mining only. Arith_Integ represents the connectivity maps that have been generated using a Borda merged ranked lists that integrate both the text-mining and the microarray based ranked lists respectively.

On the other hand, **Figure 6.1** shows that the connectivity map with the best sensitivity measure is the one that has been generated by using an integrated rank list of genes. The arithmetic mean integrated reported a sensitivity score of 80% whereas Mic_2379, Mic_22283 and Text-Mining reported sensitivity scores of 58%, 54.5% and 62%, respectively. Furthermore the integrated ranked list based method has reported better precision, specificity and accuracy scores when being compared with other methods. This indicates the superiority of the proposed paradigm in predicting drug-disease associations.

6.3. Comparing Different Algorithms for Drug-Repositioning

In the second part of this analysis, the proposed approach was compared with a text-mining based, microarray-based and similarity based approaches. More precisely, the proposed paradigm was compared with PolySearch [43], ConnectivityMap [10] and PREDICT [55]. To make the comparison fair, a unique gold standard was generated for each comparison as being described in **section 5.5.3**. Receiver Operator Characteristic (ROC) curve was used to assess the performance for each technique. That is because it shows the trade-off between True Positive Rate and False Positive Rate and therefore reflects a better measure to evaluate different methods. For each disease, the predicted associations and the prediction scores together with the corresponding gold standard for that disease were used as an input to ROC. **Table 6.1** shows the ROC score for heuristically selected cancer diseases namely; breast cancer, colon cancer, liver cancer, lung cancer and prostate cancer. In addition, it shows the average ROC score for all diseases using different methodologies.

Table 6.1: ROC scores using different methodologies

ROC Score	ROC for comparison 1		ROC for comparison 2		ROC for comparison 3	
Disease	Integrative Approach	Connectivity Map	Integrative Approach	PolySearch	Integrative Approach	PREDICT
Breast Cancer	0.882	0.55	0.72	0.66	0.75	0.52
Colon Cancer	0.91	0.58	0.82	0.8	0.84	0.52
Liver Cancer	0.89	0.61	0.77	0.74	0.77	NA
Lung Cancer	0.87	0.54	0.81	0.5	0.78	NA
Ovarian Cancer	0.82	0.5	0.76	0.58	0.81	NA
Prostate Cancer	0.83	0.6	0.74	0.6	0.78	0.6
All Diseases	0.87	0.61	0.76	0.6	0.75	0.6

Table 6.1 shows that the proposed methodology scored better ROC than all other methodology. This indicates the robustness of the proposed methodology in detecting different associations using any of the existed approaches. For instance, although all these methodologies scored very well when compared with other gold standards, it is obvious that these methodologies failed to detect other associations predicted by other data sources. For example, PREDICT [55] reported an AUC score of 0.9 when the authors evaluated their predictions with a gold standard that has been extracted from RFPSCT but failed to predict associations that resulted from PolySearch and ConnectivityMap. The integrative paradigm on the other hand, was able to score well in detecting different associations from different data sources. This suggests that using an integrative paradigm

that integrates different data sources might have the ability to unveil hidden associations that might not be discovered using these data sources independently.

6.4. Microarray un-predicted Associations

To further investigate the robustness of our method in improving sensitivity we have checked the true associations that have been detected using our proposed approach but not with using microarray based approach. **Table 6.2** lists some of the most negatively enriched associations of those together with their enrichment scores and p-values. Finally these associations were tested for biological validation using DrugBank database [65].

Table 6.2 shows the fact that some of the highly negatively enriched associations were available in the gold standard. In addition it shows that some of those negatively enriched associations have been validated in DrugBank database. Noteworthy that these associations have been predicted using the integrative approach but not with using microarray data independently. This suggests that using microarray data independently might miss some information and thus miss many important associations between biological entities. Furthermore, it suggests that these associations can be considered as an attempting target for drug repositioning. This is because most of these associations are not yet validated in DrugBank database but they have some experimental validation curated from text (some of these associations are discussed in the biological analysis and validation section).

Table 6.2: shows some of the negatively enriched associations that have been detected using our integration based approach but not with using microarray data only based approach.

Drug	Disease	Integrated Based Enrichment	Integrated Based P-value	Polysearch Relevancy Score	Validation in Drug Bank
Dexamethasone	Anemia	-0.7310	$2.67*10^{-110}$	80	No
Enalapril	Muscular Dystrophies	-0.7036	$2.77*10^{-106}$	18	No
Lovastatin	Nevus	-0.6900	$1.42*10^{-106}$	31	No
Mifepristone	Polycystic ovary syndrome	-0.6770	$2.96*10^{-104}$	308	No
Enalapril	Duchenne muscular dystrophy	-0.64657	$7.31*10^{-100}$	85	No
Ganciclovir	Congenital disorder	-0.5378	$7.22*10^{-81}$	10	No
Thalidomide	Diabetes Mellitus	-0.4370	$7.03*10^{-71}$	12	No
Sirolimus	Leukemia	-0.6116	$1.44*10^{-99}$	236	No
Etoposide	Leukemia	-0.5573	$1.42*10^{-84}$	588	Yes
Doxorubicin	Leukemia	-0.4308	$1.30*10^{-74}$	1325	No
Methotrexate	Leukemia	-0.4977	$2.4*10^{-82}$	658	No
Cyclosporine	Anemia	-0.535	$4.39*10^{-93}$	225	No
Dacarbazine	Melanoma	-0.4273	$2.1*10^{-75}$	3163	Yes
Flutamide	Breast Cancer	-0.5129	$5.66*10^{-87}$	0	No
Flutamide	Prostate Cancer	-0.4179	3.0510^{-73}	1392	Yes
Desipramine	Colon Cancer	-0.3732	$8.71*10^{-58}$	87	No

6.5. Microarray Predicted Associations

In addition to investigating associations that are only predicted using the integrative approach, we have investigated the presence of some associations that have been reported both by using our approach and microarray based approach. The associations together with their enrichment scores, p-values and their validation status are reported in **table 6.3**. **Table 6.3** provides some of the very interesting results that indicate a complementary relationship between the integrative approach and the microarray based approach for the purpose of increasing confidence about a particular association.

On the other hand the associations reported in **Table 6.3** provide an idea about the robustness of the proposed method in discovering knowledge that can be contradicting with reality when using microarray independently. For instance some of the reported associations in **Table 6.3** show positive associations in case of microarray based approaches whereas they show negative associations when using the integrative approach. This contradiction can be misleading and confusing especially that we have manually checked and validated that all associations in **Table 6.3** are more of having a drug-disease treatment relationship but not drug-disease side effect relationships (some of these associations are discussed in the Discussion section). This in other words indicates that these associations are better to have negative enrichment scores but not positive enrichment scores.

Finally it is important to note that although some of these associations were negatively enriched, when using microarray data independently, the integrative approach was able to provide more negative enrichment score, smaller p-value and thus more confidence about ties of these biological entities.

Table 6.3 shows some of negatively enriched associations that have been detected using integrative approach and microarray based approach.

Disease	Drug	Integ_Based_En	Mic_Based_En	Integ_Based P-value	Mic_Based P-value	Polysearch RE	Validation in Drug Bank
Diabetes Mellitus	Ethambutol	-0.58916	0.2748	$2.72*10^{-95}$	$7.91*10^{-48}$	10	No
Rheumatoid Arthritis	Amitriptyline	-0.5518	0.2446	$1.03*10^{-92}$	$1.46*10^{-34}$	80	No
Ovarian Cancer	Amitriptyline	-0.5070	-0.2094	$1.48*10^{-86}$	$9.23*10^{-35}$	118	No
Prostate Cancer	Mitoxantrone	-0.5435	-0.2591	$2.14*10^{-95}$	$5.84*10^{-48}$	1004	No
Rheumatoid Arthritis	Tacrolimus	-0.5226	0.2369	$3.69*10^{-89}$	$4.66*10^{-36}$	1576	No
Huntington’s Disease	Daunorubicin	-0.5153	-0.5681	$2.59*10^{-89}$	$1.12*10^{-96}$	117	No
Lung Cancer	Daunorubicin	-0.4486	-0.2240	$3.52*10^{-79}$	$4.59*10^{-37}$	0	No
Anemia	Albendazole	-0.4307	0.2792	$1.47*10^{-70}$	$1.46*10^{-40}$	152	No
Rheumatoid Arthritis	Triamcinolone	-0.4366	-0.1975	$1.95*10^{-70}$	$3.52*10^{-27}$	176	No
Diabetes Mellitus	Imatinib	-0.3795	-0.2787	$2.30*10^{-55}$	$1.06*10^{-36}$	86	No

6.6. Discussion

Getting inspired from the fact that the repositioning of drugs that has already been approved to have a safe human use mitigates the costs and risks associated with the early stages of drug development [3], we have proposed a novel approach for drug repositioning. Drug repositioning became one of the most important areas of research regarding its importance in adopting new therapeutic indications by exploiting the rigorous and already existing safety tests required by different agencies. In this thesis, a novel integrative framework was suggested for the purpose of drug repositioning. The framework designed in a way that exploit the biomedical knowledge stored in microarray expression profiles, biomedical literature and network biology for the purpose of predicting new indications for already marketed drugs. This completely unsupervised paradigm was able to record the best sensitivity measure when compared with using text-mining or microarray data independently. In addition this paradigm shows a great superiority when compared with different state of art methods that use other data sources to make predictions. This suggests that the proposed paradigm was able to provide a more comprehensive drug-disease association prediction when compared with other methods.

As previously discussed we have reported some important and negatively enriched associations that have not been predicted when using microarray data independently (see **Table 6.2**). Furthermore, it was interesting that some of the negatively enriched associations predicted by our approach are found to have some biological sense when manually curate the literature. These associations have been found to be positively enriched when using microarray data independently (see **Table 6.2**). This indicates the powerfulness of the proposed integrative method in inferring and predicting novel knowledge. To further assess the comprehensiveness of the resulting associations, we have checked the association scores between entities in **Table 6.3** in other connectivity maps described in chapter 5.

6.6.1. Biological Analysis for Associations in Table 6.2

Cortisones and hydrocortisones are naturally occurring glucocorticoids that are used as a replacement therapy in adrenocortical deficiency states. Dexamethasone which is assumed to be the synthetic analog for these compounds is an-anti-inflammatory agent that works as a glucocorticoid agonist by simply binding to specific cytoplasmic glucocorticoid receptors [65]. Dexamethasone has been widely used to relieve inflammation, treat certain forms of arthritis, skin, kidney, eye, severe allergies and other disorders. According to the proposed approach, Dexamethasone was predicted to have potential effect in Anemia disease management as it is highly negatively enriched with this disease.

Indeed several studies have reported this indication previously for this compound. For instance **Bernini JC et al.** [71] have done a study on two groups of children where the children in one group have been treated with Dexamethasone and others were treated with placebo. The authors found that Dexamethasone treated group showed a significant shorter stay in hospital, prevention of clinical deterioration and a reduced need for blood transfusions. These results have led to the conclusion that intravenous Dexamethasone has a beneficial effect on children with sickle cell disease with mild to moderate severe chest syndrome. On the other hand **Gupta N et al.** [72] ran an experiment on 8 different patients having Anemia and/or Leukemia for two years combined treatment with Rituximab, Cyclophosphamide and Dexamethasone. The results suggest that this combination should be considered in the management of this particular disease. Furthermore **Hatano K et al.** [73] have done a very recent study on 72 patients with prostate cancer and they found that a low dose combination of Docetaxel, Estramustine and Dexamethasone is active and tolerable with beneficial effects on Anemia and bone pain in patients with prostate cancer.

Enalapril is an angiotensin-converting enzyme (ACE) inhibitor drug. It is mainly used for the treatment of essential or renovascular hypertension and symptomatic congestive heart failure [65]. According to the proposed paradigm, Enalapril was predicted to have high negative enrichment with Muscular Dystrophies and Duchene muscular dystrophy diseases. Navigating in literature, these associations were found to have biological sense**.** For instance **Cozzoli A et al.** [74] did an experiment on mdx mouse model to study the effect of Enalapril on Dystrophies recovery. Enalapril was found to cause a dose-dependent increase in fore limb strength where the highest dose has led to a complete recovery in comparison with the control model. This suggests the ability of Enalapril in blunting some muscular functional impairment that might result from angiotensin activation of pro-inflammatory pathways. **Ramaciotti C et al.** [75] did another experiment on 50 patients to study the effect of Enalapril on Duchenne muscular dystrophy. The results reported that 43% of patients were able to respond to Enalapril and get recovery.

Flutamide is a non-steroidal anti-androgen compound that has a potent rule in the management of locally confined stage B2-C and stage D2 metastatic carcinoma of the prostate [65]. Based on our paradigm Flutamide was found to have a high negative enrichment with prostate cancer indicating the robustness of our approach in detecting such associations. In addition Flutamide has been found to have a high negative enrichment with breast cancer. Many papers in literature have validated the anti-neoplastic effect of Flutamide on estrogen receptor negative subtype of breast cancer. For instance **Naderi A and Liu J** [76] investigated the therapeutic effects of persistent ERK phosphorylation in combination with AR inhibition using Flutamide. The results demonstrate a significant reduction in breast cancer cell viability and growth indicating its promising effect in management apocrine breast cancer. A similar finding has been reported by **Naderi A and Hughes-Davies L** [77] which demonstrates that a combined treatment with Flutamide together with ErbB2 pathway inhibition can be effective in reducing breast cancer cell viability.

Sirolimus or Rapamycin is a macrolide compound obtained from Streptomyces hygroscopicus that works as a potent immunosuppressant for organ transplant rejection and possesses both antifungal and antineoplastic properties. According to our findings Sirolimus is highly negatively enriched with Leukemia. A very recent study has demonstrated the effect of Sirolimus derivative (Everolimus) in a combined treatment with all-trans retinoic acid (ATRA) on acute myeloid leukemia (AML) [78]. The results showed growth inhibition and apoptosis in AML cell lines. Another recent study has showed that a combined treatment of Rapamycin and Dexamethasone in cell lines and xengraft model of leukemia cell lines showed a significantly greater apoptosis and cell cycle arrest in some cell lines [79]. Those were not the only studies mentioning the role of Rapamycin in inducing leukemia cell apoptosis. Indeed many other studies have mentioned and validated this therapeutic indication clinically [80].

6.6.2. Biological Analysis for Associations in 6.3

One of the negatively enriched associations reported in **table 6.3** is the Amitriptyline drug with rheumatoid arthritis (RA) disease. Noting that, this association was positively enriched in case of using microarray data independently. Therefore, we reviewed literature to check whether amitriptyline can really counteract the RA effect or it works in parallel with its pathophysiology. Amitriptyline is a tertiary amine tricyclic antidepressant that is used for treatment of depression, chronic pain, irritable bowel syndrome and many other diseases [65]. **Bird H and Broggini M** [81] did a study on 191 patients with RA to check the effect of Amitriptyline and another drug. The results showed that Amitriptyline was effective in treatment of depression with improvements in RA associated pain and disability. On the other hand **Frank RG et al** [82] did another experiment on 47 patients with RA and they realized that Amitriptyline was the most effective drug in reducing pain. The authors suggested using a moderate dose of Amitriptyline as an adjunct drug for the treatment of pain in both depressed and non-depressed patients with RA.

Tacrolimus also shared a similar story with Amitriptyline where it has a high negative and a high positive enrichment score with RA using the proposed method and microarray, respectively. Tacrolimus is an immunosuppressive antibiotic whose main first use was to reduce patient's immune system and thus reducing the risk of liver transplantation rejection [65]. Later on, its usage has been extended to include many other organ transplantations. According to our paradigm Tacrolimus can play a major rule in RA management. In literature, Tacrolimus were found to be significantly effective for suppressing the activity of AR which makes it a promising candidate for this disease's treatment [83] [84]. In another experiment of 123 patients, a combination of Tacrolimus with other antirheumatic drugs has been found to have a significant effect in RA management when compared with using antirheumatic drugs only [85]. These very interesting findings suggested that using of Tacrolimus in patients with inadequate response to antirheumatic drugs is useful and could became one of the most promising options for those patients.

Mitoxantrone is a DNA-reactive agent that interferes with DNA and RNA and assumed to be a potent inhibitor of topoisomerase II. This compound has an anti-neoplastic properties and it is used for the treatment of secondary progressive, progressive relapsing remitting multiple sclerosis [65]. In one of the experiment, Mitoxantrone was found to be able to induce Fas receptor expression on primary prostate cancer cell lines which translated into enhancement of apoptosis of all cancer cell lines treated [86]. Furthermore Mitoxantrone has shown potential effect as an anti-neoplastic compound when combined with other drugs. For instance **Pinto AC et al.** [87] found that a combined treatment of Mitoxantrone with Imatinib can significantly induce tumor growth inhibition. In addition, it has been found that Mitoxantrone in combination with prednisone can work as second-line chemotherapy in Docetaxel-refractory patients [88]. The survival rate was prolonged and the side effects were completely low for such a combination.

6.6.3. Analysis of Different Experiments

From all previously mentioned examples we realize that even though microarray data has an enormous contribution to drug discovery, integrating biological data from other sources might improve its efficiency. Indeed we found that depending on one source of information might not only lose some information but it can also result in associations that contradict with reality.

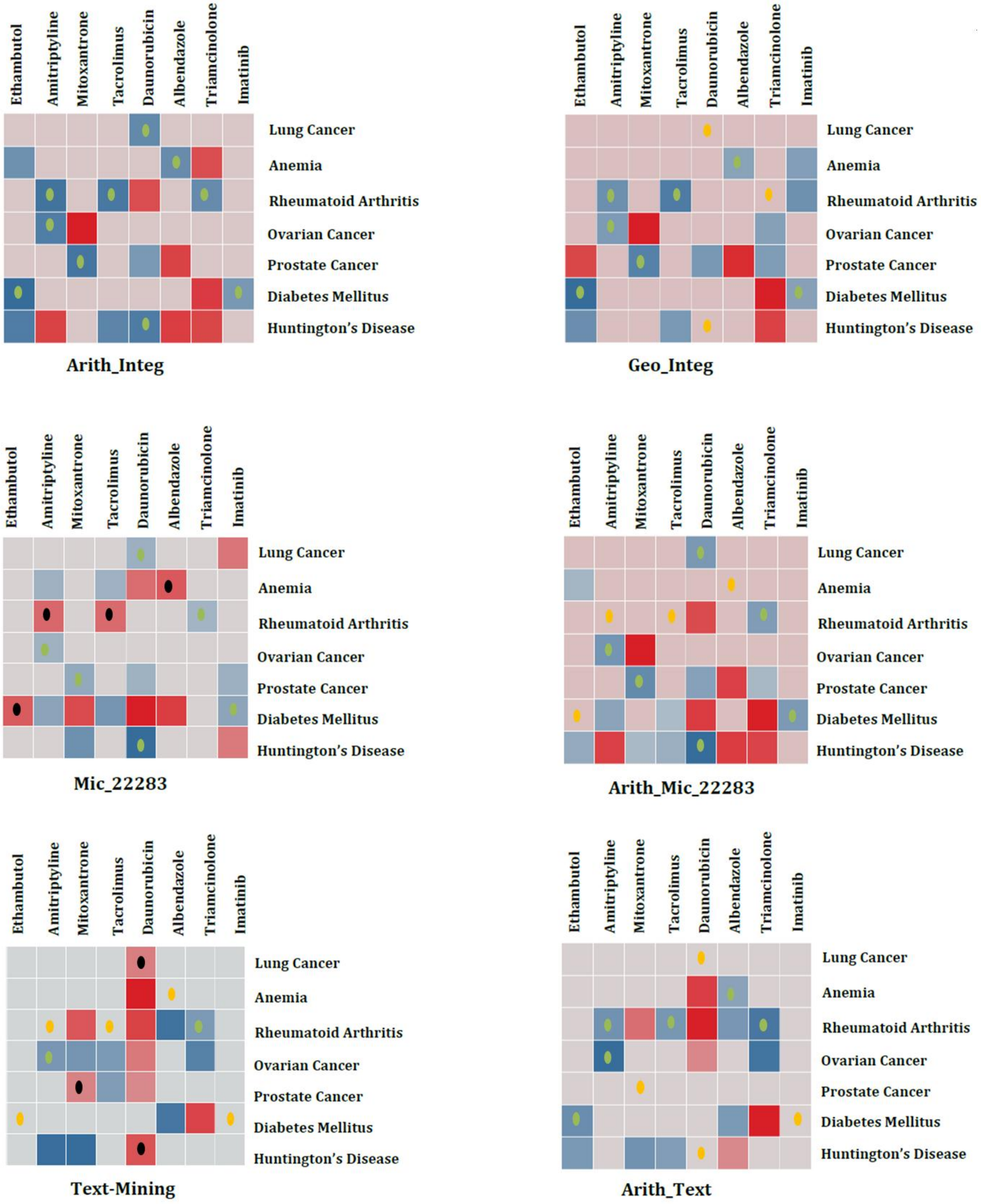
To further analyze this property we sought to study the prediction scores for associations in **table 6.3** using six different connectivity maps (see **figure 6.2**). Arith_Integ, Geo_Integ, Mic_22238 and Text-Mining were discussed previously in chapter 4. Arith_Text and Arith_Mic22238 were generated by taking the average enrichment score between Arith_Integ from a side and Mic_22238 and Text-Mining from the other side, respectively. To facilitate the comparison between these connectivity maps different HeatMaps were generated where each blue square represents a negative enrichment score and each red square represents a positive enrichment score. In addition, green, black and yellow dots were added to tell if a specific association is true positive and consistent with reality, true positive and contradicting with reality or false negative, respectively. These colored dots were only added to the associations in **table 6.3**. Noteworthy that true positive or false negative judgment were given based on the gold standard whereas contradicting or being consistent with reality judgment were given through manually navigating PubMed abstracts.

Starting from the integration based connectivity map, Geo_Integ connectivity map was able to capture 70% (green dots) of the associations with 0% black dots. This indicates the robustness of Geo_Integ in detecting these associations but with a less sensitivity than Arith_Integ. Mic_22283 was able to detect 100% of the associations but 40% of them were black dots and thus contradicting with reality. On the other hand Arith_Mic_22283 was able to predict 60% of the associations with 0% black dots. We argue that Arith_Mic_22283 performs better than Mic_22283 as the former could not detect 40% of the associations but the later resulted in 40% misleading associations. This is another proof to indicate the great demand for data integration with all of its form. In other words, although a simple averaging scheme to generate Arith_Mic_22283, Arith_Mic_22283 was able to result in more accurate associations in addition to its power in improving sensitivity when compared with Mic_22283 (see **figure 6.2**).

On the other hand, Text-Mining was able to detect 50% of associations where 60% of them were black dots. This suggests that neither the predicted associations nor gold standard was biased toward Text-Mining. In addition it might indicate that the robustness of the integration based connectivity maps (Arith_Integ and Geo_Integ) came from the complementary relationship between the microarrays based ranked gene lists and the Text-mining based ranked gene lists. This strongly supports the assumption that the ranked gene list from either source can correct or at least modify the ranked gene list from the other source.

Arith_Text was able to detect 60% of the associations with 0% black dots. This is the exact same result that has been obtained with Arith_Mic_22283. Both results are found to be obviously superior than the results obtained with using text mining or microarray data independently (Text-Mining or Mic_22283 respectively). Note that the whole connectivity map for Arith_Integ is provided in **figure 6.3.**

Figure 6.2 HeatMaps for the enrichment scores between biological entities mentioned in table 6.3.

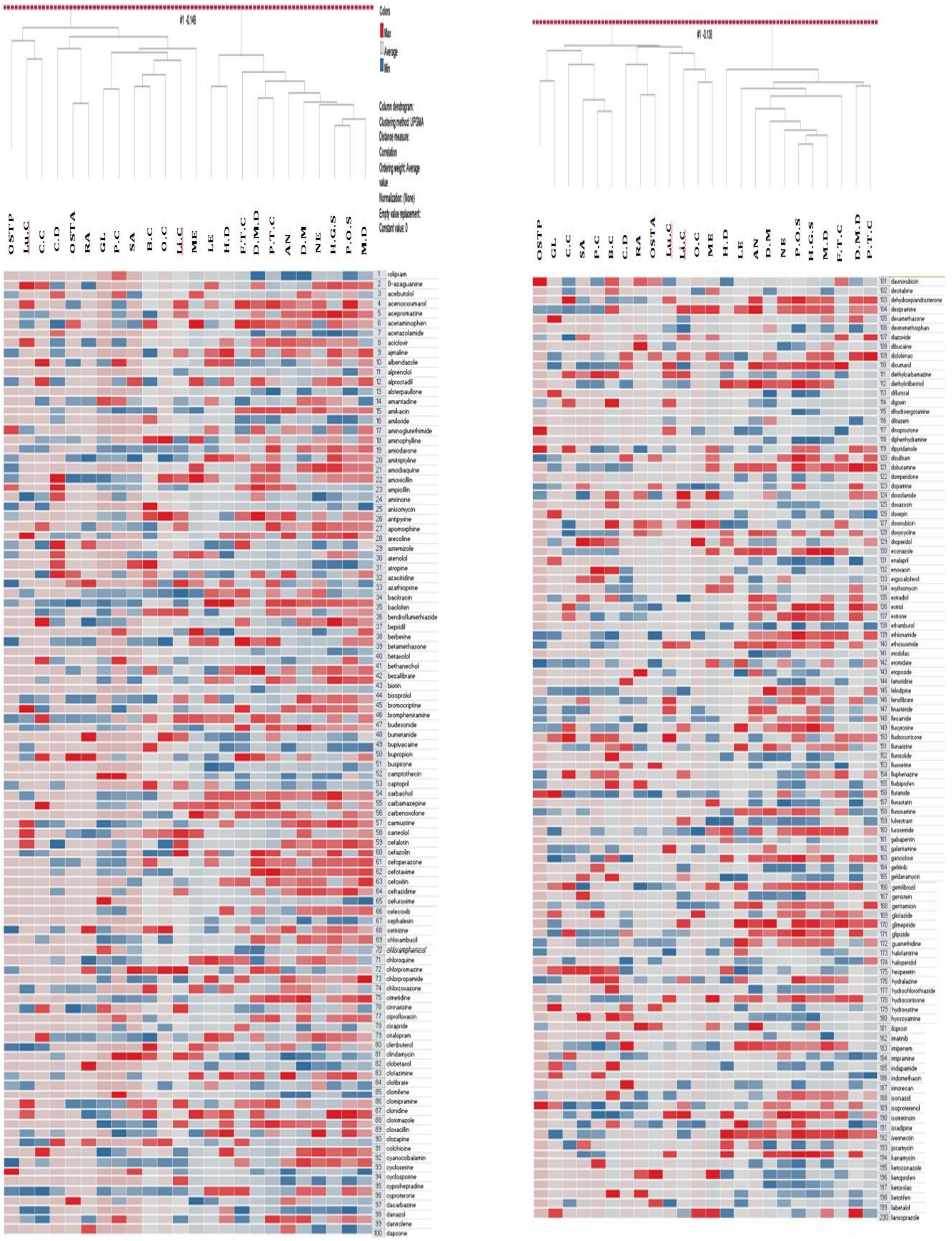


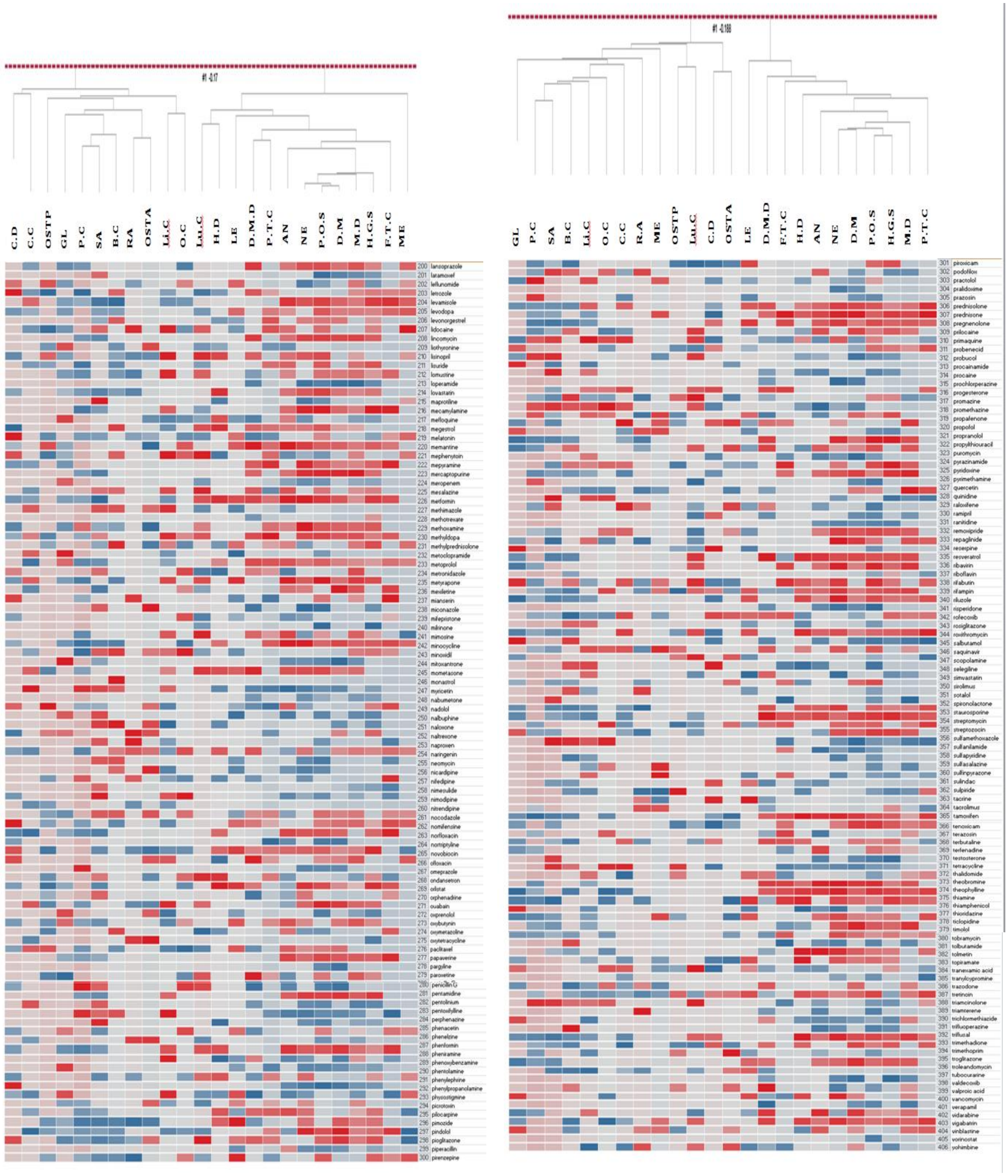
This figure shows the enrichment score between different biological entities mentioned in **table 6.3**. These association scores are based on the previously described different paradigms for getting different connectivity maps. The scores have been curated and converted to Heat Maps to facilitate the comparison process. A deep blue color indicates high negative enrichment whereas a deep red color indicates high positive enrichment. Green, black and yellow dots have been added to indicate whether a particular association is true positive and consistent with reality, true positive and contradict with reality or a false negative respectively.

6.6.4. Exploring the Whole Connectivity Map

To improve the usability of the results, a visualized version of the predicted associations was generated as a HeatMap. A researcher can have a look at any disease of interest and find the corresponding enrichment score with the studied drugs. Spotfire TIBCO software (http://spotfire.tibco.com/) was used for the purpose of visualizing the whole connectivity map. Noting that, including the whole connectivity map in one HeatMap (HM) would not result in good visualization nor would it give clear labeling. Therefore, the connectivity map was split into four different HMs. Two-dimensional Hierarchical clustering with average linked measure was used to cluster related diseases and drugs for each set.

Figure 6.3 Whole HeatMap for the Computed Connectivity Map





This figure shows the whole connectivity map for all drugs and diseases in this study. Blue rectangle represents a strong negative association whereas a red rectangle represents strong positive association. In here, drugs are arranged in rows whereas diseases are in columns. Symbols have been used to represent diseases names according to the following: Lu.C= lung cancer, AN=anemia, B.C= breast cancer, LE=leukemia, NE=nevus, ME=melanoma, RA=rheumatoid arthritis, OSTA= osteoarthritis, OSTP= osteoporosis, O.C= ovarian cancer, P.C= prostate cancer, SA= sarcoma, F.T.C=follicular thyroid carcinoma, P.T.C=papillary thyroid carcinoma, D.M=diabetes mellitus, Li.C=liver cancer, C.C= colon cancer, C.D=congenital disorder, GL= Glioblastoma, H.D=huntington's diseases, H.G.S=Hutchinson-Gilford syndrome, P.O.S=polycystic ovary syndrome, D.M.D= Duchenne muscular dystrophy, M.D= muscular dystrophies.

6.6.5. Work Limitations

Finally it is reminding worthy that although of the interesting results obtained using this integration scheme, the methodology is not without caveats. First of all, other biological information from different sources and databases can be integrated and further improve the results. For example gene to phenotypes analysis can be used to further filter and re-prioritize genes related to a particular disease or drug. In addition, integrating other biological entities that can play a rule in drug or a disease molecular action, like metabolic network analysis or miRNA-gene interaction, might also improve the results. Second of all, one can consider a more complex framework to prioritize genes from the most over-expressed to the most down-regulated using text mining based approach. For instance, further improvement can be achieved by going beyond prioritization method that is based on co-occurrences and simple natural language processing. For example, identifying a framework that is able to extract abstracts and analyze them both syntactically and semantically might decrease the false positives when predicting a positive or a negative relevancy score between a gene and a drug or a disease. Noteworthy that such a framework might need to compromise between complexity that might result in more accurate results and simplicity that might result in computationally less intensive operations.

Chapter 7: Applying Integration Concept for the Construction of Functional miRNA-Disease Interaction Using Regression Model

To further investigate the comprehensiveness of the biological data integration approach in predicting novel associations, the model was applied for constructing functional miRNA-disease interactions. The main reason to apply this approach on such data is the growing body of evidence associating microRNAs (miRNAs) with human diseases, in addition to the large amount of high-throughput data on diseases and miRNAs. miRNAs are new key players in the disease paradigm demonstrating roles in several human diseases. The functional association between miRNAs and diseases remains largely unclear and far from complete. With the advent of high-throughput functional genomics techniques, it is now possible to infer functional association between diseases and biological molecules by integrating disparate biological information.

7.1. Background

MicroRNAs (miRNAs) are small RNA molecules that regulate genes by triggering target degradation or translational repression [89]. miRNAs play a key role in diverse biological processes including differentiation, cell cycle and apoptosis [90]. About 3% of the human genes encode for miRNAs, each miRNA is estimated to regulate hundreds of genes, and over 50% of the human protein-coding genes are regulated by miRNAs. Computational predictions estimated that there are around 1,700 miRNAs in human genome [91]. This makes miRNAs one of the most abundant classes of regulatory genes in humans.

MicroRNAs expression is altered in several diseases including cancer and thus it is very likely that alteration in miRNA expression could lead to human diseases [92]. Several studies have investigated the role of miRNAs in cancer using mRNA and miRNA expression profiling [89] and suggest that most diseases are attributed to more than one miRNA that affect hundreds of genes.

There are several lines of evidence suggesting functional association between cancers and miRNAs. First, miRNAs are shown to control cell proliferation and apoptosis [90]. Thus their dysregulation may contribute to proliferative disease. In addition, several miRNAs showed to act as tumor suppressor or oncogenes [93]. Second, genome-wide association studies demonstrated that most human miRNAs are located at fragile sites in the genome or regions that are commonly altered or amplified in human cancer [94]. Mutation of miRNAs, dysfunction of miRNA biogenesis and dysregulation of miRNAs and their targets may result in various diseases.

The question still remains how miRNA alteration might cause a disease. All these evidences support the strong necessities in understanding the functional association between miRNAs and diseases. Many studies have produced large number of miRNA-disease associations and showed that the mechanisms of miRNAs involved in diseases are very complex. Uncovering disease-miRNA associations will help pharmaceutical community to understand the underlying mechanisms in diseases and thus narrow down the search pace for new therapeutic targets. This would result in better insights into the functional role of newly discovered miRNAs in certain diseases.

Disease gene signatures bear a signature of regulatory activity of miRNAs as it is anticipated that the collective effect of miRNAs may lead to dramatic changes in the expression of their targets that may lead to diseases. Although integrating bioinformatics approaches with miRNA expression data can predict miRNAs deregulated in certain diseases, only very few miRNAs have been functionally validated in disease context, and the underlying mechanisms of why and how miRNAs become deregulated are unknown. Better understanding of the regulatory role of miRNAs in cancer development and progression requires exploring their cooperative influence on target genes context.

Characterizing the effect of miRNA on target-context protein partners gained considerable body of attention in the past few years. Protein degree in PPI networks showed to be correlated with the number of targeting miRNAs [95]. Topological features of proteins in PPI showed to be useful to eliminate false discoveries in miRNA-target prediction algorithms [95]. These observations shed light on the influence of miRNAs on the PPI subnetwork involving the targets and highlight the importance of considering target protein partners when searching for functional miRNA-disease interactions.

To summarize the contribution of this chapter, a logistic regression model was used to identify miRNAs whose target's protein contexts are enriched in disease gene signatures. The model was applied to identify diseases-miRNA associations by integrating disease gene signatures extracted from microarray experiments and PubMed abstracts, with miRNA-gene interactions resulting from integrating predicted miRNA-gene interactions and their influence on target protein context. This integrative approach has enabled the prediction of functional association between miRNAs and diseases. The results of the model were validated against a miRNA-disease interactions gold standard using ROC analysis. Finally, more analysis has been done on newly identified miRNAs in prostate cancer and characterizes their functional role.

7.2. Materials and Methods

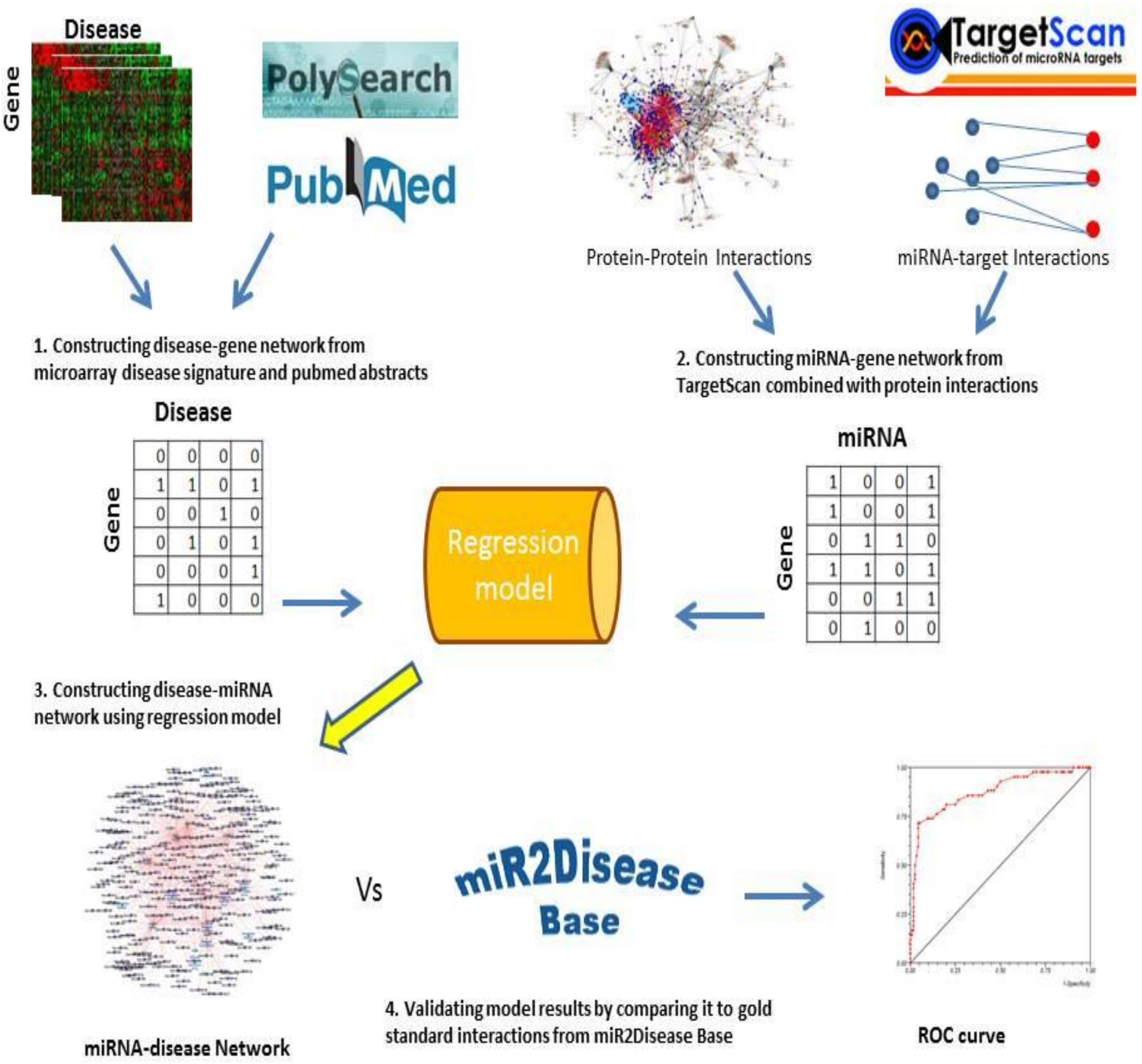
In this section, I describe how the miRNA-target and disease-gene networks were constructed and preprocessed as input to the proposed regression model. First, the steps to define gene-disease and miRNA-target interaction networks to define signatures for each disease and miRNA respectively are described. The regression model used to associate miRNAs with disease is then explained. Finally, validation steps to validate the predicted results from the proposed model are discussed. The whole framework used in this study is shown in **figure 7.1**.

7.2.1. Identification of Disease-Gene Signature

Gene-disease interactions were retrieved from two independent sources. The first source was microarray expression profiles related to 24 diseases including 13 cancers from Gene Expression Omnibus. That was the same data used to build the drug-disease connectivity map. All preprocessing and normalization steps were discussed in **chapter 4**.

The second source of data was PubMed publications. But instead of using the proposed text-mining approach, discussed in **chapter 4**, PolySearch [43] web server was used to generate data. The reason choosing PolySearch, but not our proposed text-mining technique, is that PolySearch employs text ranking scheme to score relevant sentences and thus it could result more accurate results since it has access to sentences as being described in **chapter 3**.

Figure 7.1: General Framework for disease-miRNA interaction predictions


Disease
Gene
PolySearch
PubMed
Protein-Protein Interactions
TargetScan
Prediction of microRNA targets
miRNA-target Interactions
1. Constructing disease-gene network from microarray disease signature and pubmed abstracts
2. Constructing miRNA-gene network from TargetScan combined with protein interactions
Disease
Gene
Regression model
miRNA
Gene
3. Constructing disease-miRNA network using regression model
Vs
miR2Disease Base
4. Validating model results by comparing it to gold standard interactions from miR2Disease Base
miRNA-disease Network
ROC curve

0	0	0	0
1	1	0	1
0	0	1	0
0	1	0	1
0	0	0	1
1	0	0	0

1	0	0	1
1	0	0	1
0	1	1	0
1	1	0	1
0	0	1	1
0	1	0	0

Four major steps to construct functional disease-miRNA associations. First, disease-gene interactions that were constructed by integrating disease signatures from microarray gene expression data and from PubMed abstracts. Second, miRNA-gene associations were constructed by integrating computationally predicted miRNA-target interactions and protein networks. The aim of integrating protein networks is to reduce noisiness in the predicted data. Proteins that are not targeted by a miRNA but their partners are, are considered as indirect miRNA-target association. Third step is to process the two inputs (gene-disease and miRNA-gene) as input to the regression model. The final step is to evaluate the predicted results against gold standard miRNA-target interactions data

Default keywords that were manually curated in [43] were used to relate diseases with genes. The number of abstracts was set to 10,000 and thus obtaining results for the most 10,000 relevant abstracts. For this experiment, we heuristically considered all genes with non-zero relevancy score. 720 genes of relevance to the set of diseases were extracted. Finally the union of the two gene sets (from microarray and text-mining) was considered to build a bipartite graph *DiseaseSig* between genes and diseases.

7.2.2. Constructing miRNA-target Interactions

Human miRNA-target computational predictions were taken from TargetScan 5.1 [96] which showed to outperform all other miRNA-target prediction methods [97]. These interactions are direct interactions between miRNAs and their targets. Another set of non-direction miRNA-target interactions was considered, by considering direct neighbors of target proteins in PPI network. Undirectional functional protein interactions, or PPI, were extracted from Reactome [98] , which includes proteins physically interacting, proteins sharing biological function and regulatory interactions. Proteins that are not targeted by miRNAs but at least five of their neighbors are targeted by miRNAs, were considered indirectly influenced by miRNAs. In this study, both direct and indirect miRNA-target interactions (*NetmiR*) were combined and used it as input to regression model.

7.2.3. Logistic Regression Model

Regularized Logistic Regression was used to predict miRNA-Disease functional interactions. Disease-gene (*DiseaseSig*) and miRNA-gene (*NetmiR*) interactions have been used as response and predicted variables, respectively, as input to the regression model. Let *DiseaseSig* represent the gene signature of a particular disease, *NetmiR*$^{(i)}$ be the miRNA-target influence profile a specific *miRNA(i)* for *i=1.....m*. let *miR-Dis* is the gold standard for miRNA-Disease functional interactions. Then the logistic regression's cost function in relevant to a particular disease *dis(z)* can be written as

$$\underset{j=1\ldots n}{J(\beta j)} = \frac{1}{m}\sum_{i=1}^{m} Cost(H(NetmiR^{(i)}), y^{(i,z)}) \qquad \textbf{Equation 7.1}$$

Where $y^{(i,z)}$ is 1 if *miRNA(i)-dis(z)* pair ∈ *miR-Dis* and zero otherwise, and

$$Cost(H(NetmiR), y) = \begin{cases} -\log(H(NetmiR)) & if\ y=1 \\ -\log(1-H(NetmiR)) & if\ y=0 \end{cases} \qquad \textbf{Equation 7.2}$$

Therefore the whole cost function can be re-written as

$$J(\beta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i,Z)} \log H(NetmiR^{(i)}) + (1-y^{(i)})\log(1-H(NetmiR^{(i)})\right] \qquad \textbf{Equation 7.3}$$

Further details about logistic regression and solving its cost function are provided in **chapter 2**.

7.2.4. Evaluating the Performance of Regression Model

The predicted disease-miRNA interactions of the regression model were validated against a gold standard of disease-miRNA associations manually extracted from miR2disease and HMDD databases [99]. The gold standard network contains 743 interactions between the 24 disease and 305 miRNAs. Area under curve (AUC) is used to assess the performance of the proposed model and compare it with other results. The purpose of this step was to show that integrating multiple data sources, microarrays and PubMed abstracts, to define disease gene signatures and integrating the influence of miRNAs on the target protein context is valuable to uncover disease-miRNA interactions. MiRNAs associated with prostate cancer were further analyzed and prediction results were validated using two independent prostate miRNA profiling studies. The aim was to assess the diagnostic and prognostic value of the new predictions of the method.

7.3. Results and Discussion

7.3.1. Constructing miRNA-target and disease-gene networks

MiRNA-target network and gene-disease network were first constructed to be used as predicted and response variables, respectively, as input to the regression model. MiRNA-target network was constructed by integrating results from TargetScan and protein interactions. This study only focused on genes that are targeted by a miRNA and interact with proteins at the protein level. 3,235 genes were obtained and found to be targeted by 305 miRNAs. To build disease gene interactions, disease gene signature from microarray data (1942) and PubMed abstracts (720) were combined and considered in this analysis. This combination resulted in a set of 2,061 genes across 24 diseases. Finally the intersection between the 3235 and 2061 gene lists were considered leading to a new list of 658 genes.

7.3.2. Regression is able to identify miRNAs from downregulated gene sets

The performance of the proposed paradigm was assessed by using several gene lists reported by recently published studies that used microarray analysis to reveal genes whose expression is affected by pre-miRNA treatment. For example, in [100] LNCaP cell lines were treated with pre-miRNA (pre-miR-1, pre-miR206, and pre-miR27b) and downregulated genes were identified using differential gene expression analysis. The downregulated gene lists that were used as *DiseaseSig* and *NetMiR* were used to evaluate the performance of the regression model to identify the influential miRNAs after treatment. MiRNA prediction scores from the regression model were used to assess the enrichment of miRNAs' targets in the gene set.

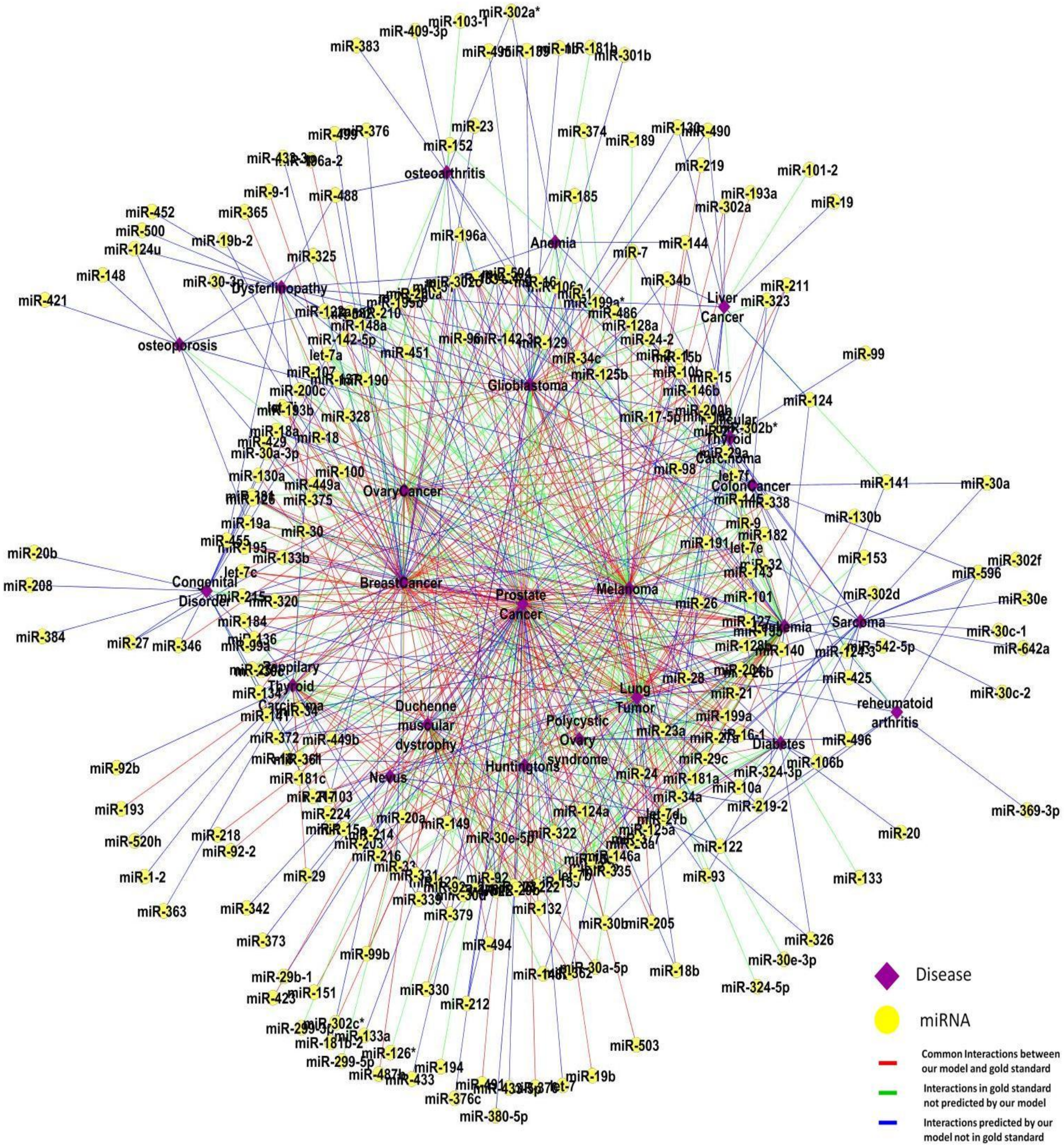
In the pre-miR-1 downregulated genes, the regression model ranked miRNA-1 first with the highest coefficient value. In the pre-miR-206 downregulated genes, the regression model showed that miR-1 and miRNA-206 have the highest prediction scores of the downregulated genes. In the downregulated genes after miR-27b treatment, the model showed that miRNA-9 has the highest prediction score and miRNA-27b ranked second. Enrichment results of the proposed model were compared with Fisher test and hypergeometric test and a miRNA enrichment tool Geneset2miRNA [101]. The results of the proposed method demonstrated that it is able to infer correct miRNAs from gene lists downregulated after pre-miRNA treatment and it can better infer the influential miRNAs.

These findings show that integrating the influence of miRNA on the protein context of the target improves miRNA enrichment analysis and demonstrated effectiveness for using regression to predict miRNA-disease functional associations.

7.3.3. Reconstructing miRNA-disease functional association

After demonstrating that regression model has successfully identified miRNAs from downregulated gene lists post to miRNA treatment, regression model was applied to identify miRNAs associated with diseases using miRNA-target and disease-gene networks. In this section, I discuss the network generated using combined microarray and abstracts disease gene signature with PPI based miRNA target network. The proposed model generated 741 interactions between the 24 diseases and the 365 miRNAs. 364 interactions were common with the gold standard, 157 were in the gold standard and missed by the proposed model, and 220 were identified by the model and not in the gold standard (**Figure 7.2**). 37 new interactions were predicted between miRNAs and prostate cancer. Further diagnostic and prognostic characterizations of the 37 prostate miRNAs were conducted.

Figure 7.2: Predicted disease-miRNA functional association



Predicted miRNA-disease interactions using regression model. Combined microarray and abstract disease gene signature were used as response variable with PPI-based miRNA-target signatures as predicted variable. We mapped all the common interactions between the predicted interactions and the gold standard data. We also showed the novel interaction predicted by our model and the interactions missed by our model. Results showed that results are biased to cancer diseases (prostate, breast, ovary, glioblastoma, melanoma as they have more complete gene signatures.

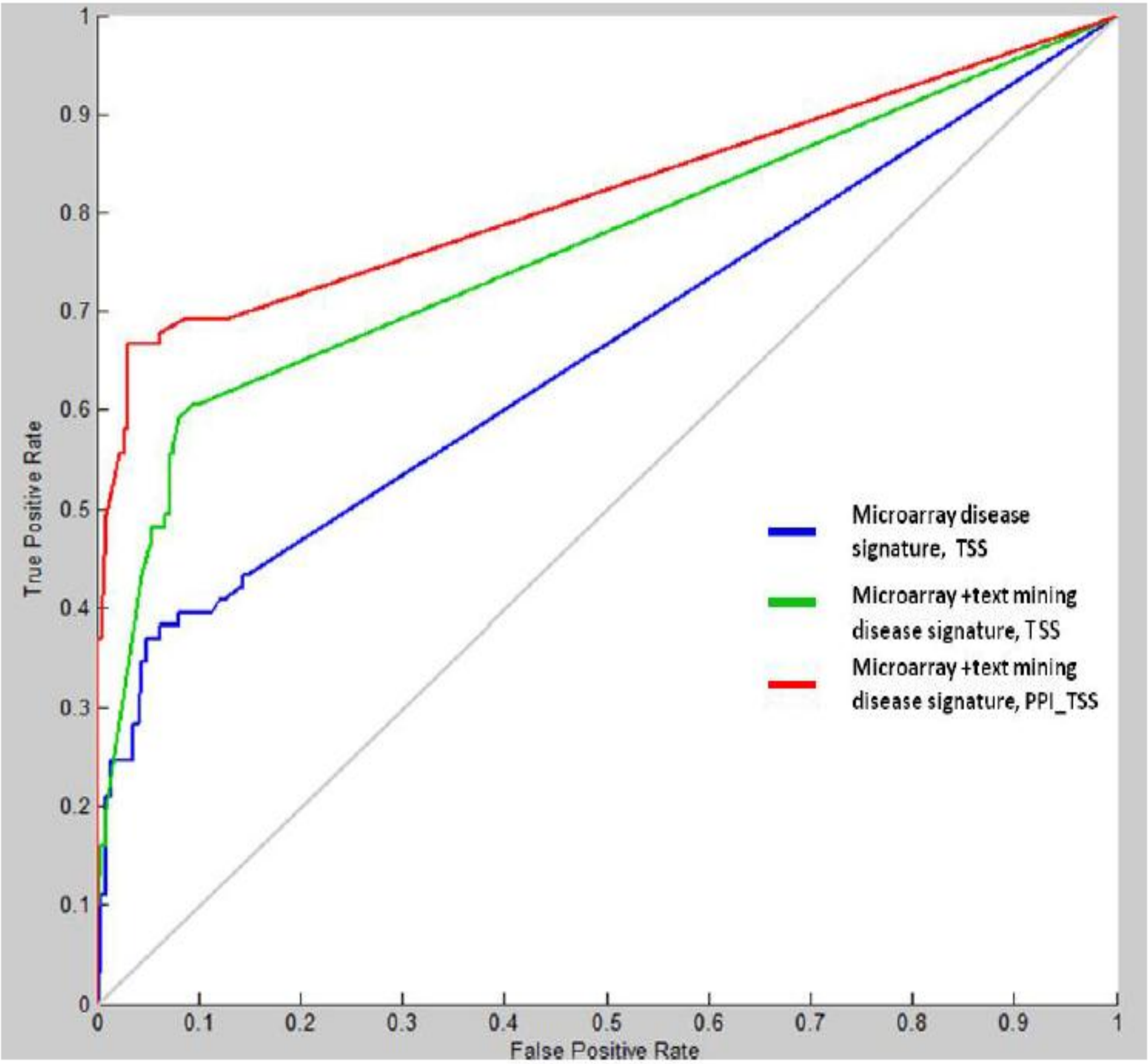
7.3.4. Assessing the performance of the proposed method

The performance of the regression model was evaluated on a gold standard miRNA-disease interactions obtained from miR2Disease database that contains experimentally verified miRNA-disease associations [99].

740 interactions between the 24 diseases and the miRNAs were extracted. The performance of the model was evaluated using several combinations. We first used the microarray gene signature-disease network vs miRNA-target network obtained from TargetScan to predict miRNA-disease associations. We then combined disease gene signature from PubMed with the microarray gene signature vs the targetscan miRNA target network. In the third test, we used the combined microarray and text signature vs TargetScan and PPI based miRNA-target network. The goal of this step was to assess if including more disease signatures and miRNA targets would increase the performance of the model. The last combination is to use PITA miRNA-target algorithm instead of TargetScan to assess the performance of the model when changing the input data sets. The model returns miRNA-disease association values that ranged from 0 to 1.

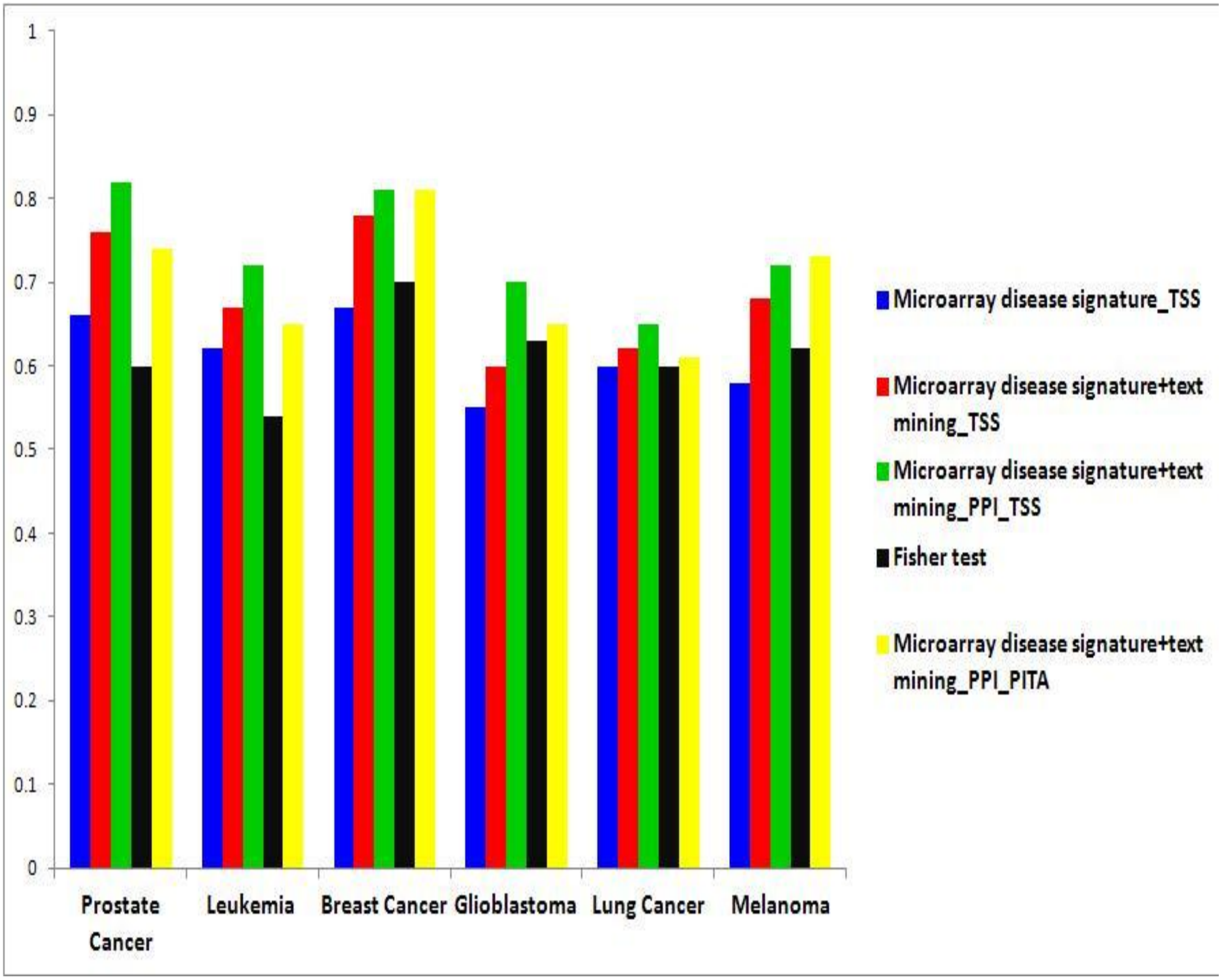
ROC curve analysis was conducted to assess the performance of the model against different network construction strategies. ROC curves for prostate cancer (**Figure 7.3**) showed that integrating disease signature from abstracts increased the performance of the model, and integrating indirect miRNA-target association increased the performance of the model even more. ROC curve analysis was also conducted on six other cancer diseases and found consistent results in all the diseases (**Figure 7.4**).

Figure 7.3: Comparative analysis using different integrative biology approaches to predict miRNA-prostate cancer interactions



ROC curve analysis of prostate cancer using three different inputs to the regression model. ROC curve of integrating microarray and abstract based disease gene signature with PPI-based miRNA target showed AUC of 0.81.

Figure 7.4: Comparative analysis using different integrative biology approaches



Area under curve values using different inputs in different cancer types. We compared the ROC results from different combinations of inputs. Integrating multiple data to define disease gene signatures and including protein networks to define miRNA signature improves the accuracy of the model. Different miRNA-target interaction data leads to different results. This is due to the completeness of miRNA-target interactions

The reason to focus on these cancer diseases, is the fact that they have the highest number of miRNAs associated with them. AUC results from the proposed regression model were compared with results obtained using Fisher test and found that regression performs better than Fisher test-based (miRNA-target, disease-gene). This suggests that the performance of the regression model is robust and can be adapted to different networks.

7.3.5. Discussion

Over recent years, miRNAs have emerged as major players in the complex networks of gene regulation and have been implicated in various aspects of human diseases. Deciphering functional associations between miRNAs and diseases is a major step toward understanding the underlying patterns governing miRNA disease associations. In addition, it gives better insights into the functional role of miRNAs in disease development. The accumulated data on miRNA expression levels in tumors demonstrate that miRNAs are promising diagnostic candidates to distinguish different tumors and different subtypes of tumors as well as to predict their clinical behavior. The observations supported the role of miRNAs as either prognostic and/or diagnostic markers. miRNAs have therapeutic applications by which disease-causing miRNAs could be antagonized or functional miRNAs could be restored.

Regression modeling demonstrated promise to construct miRNA-target networks [102]. Motivated by this work, logistic regression model was used in order to predict functional associations between miRNAs and diseases based on gene signatures of each. Since there is an explosion of disease microarray data, it was used to define a signature for each disease.

To assess the noisiness in the disease signature, disease-gene signatures from PubMed abstracts were used to generate signature that cover wider spectrum of genes. For the miRNA-gene network, only genes that are interacting with other proteins or genes and are directly or indirectly influenced by the miRNAs were considered. This is because these genes are anticipated to have higher influence on disease progression compared to genes that are targeted by miRNAs and not propagating their influence on the protein network.

We first evaluated the performance of regression as a miRNA enrichment analysis method as a proof of concept. Regression successfully identified miRNAs from downregulated genes after miRNA treatment. The performance was further evaluated on disease -miRNA interaction networks. Disease-miRNA association network was extracted from miR2Disease and HMDD that contain manually curated database for microRNA deregulation in human diseases. ROC curve analysis showed that integrating microarray and text abstracts to define disease signature gives better performance compared to using the signatures separately. Similarly, integrating miRNAs' indirect influence on genes to define miRNA target signature demonstrated better performance compared to using the direct influence alone. This suggests that refining signatures is a key step for accurate regression modeling.

Two key issues might be having big effect on the accuracy of the model. The first one is the completeness and noisiness in the disease and miRNA signature. The more complete and refined the signature is, the more accurate the model is. Since microarray disease gene signature might harbor many off target genes that are irrelevant to the disease, more robust disease gene signature that is based on integrating more evidences is essential for the success of the modeling process. Similarly, incomplete miRNA-target interactions showed to affect the performance of the model. Using miRNA-target interactions from PITA showed less accuracy compared with TargetScan results. This suggests that miRNa-target data plays critical role in regression modeling to predict functional associations between miRNAs and diseases.

The second issue is the gold standard data. Gold standard data was biased toward certain diseases like prostate cancer, breast cancer, and glioblastoma that have around hundred associated miRNAs. However, other diseases like sarcoma and colon cancer are associated with very few miRNAs like let-miR-7a and miR-21, respectively. This has big impact on false discovery rates and thus AUC performance measure. A more curated miRNA-disease interactions network is required to have more accurate performance evaluation. Unfortunately, a complete manually curated miRNA-disease database is not available. Therefore miR2Disease and HMDD were combined to trade-off the incompleteness in the used miRNA-target interactions.

Chapter 8: Conclusion and Future Work

8.1. Conclusion

Predicting functional associations between diseases and drugs using omic-data integration is valuable and promising to reveal biological mechanisms underlying diseases and drugs mode of action. Utilizing freely available data sources, two models with different statistical measures were built. The first uses enrichment analysis statistical measure to build associations between drugs and diseases, whereas the second uses regression model to predict miRNA-diseases functional associations. Both models were based on biological data integration and have shown better performance when integrating multiple data sources of different nature.

For the first model, I showed that merging the concepts of microarray technology, text-mining and network biology using computational biology techniques for the purpose of drug repositioning, have the potential to speed up drug discovery and testing processes. In this thesis, I tackle the great demand in integrating biological data from different sources to elicit better knowledge regarding drug discovery. The power of text-mining methods in discovering hidden or indirect relationships, the power of microarrays in providing a global view of drugs/diseases molecular effects and the power of gene network in understanding the functional and behavioral correlation between genes were all utilized to build a novel integrative paradigm for drug repositioning. The proposed paradigm was able to predict many associations that could not be detected when using microarray or text-mining data independently.

In addition, motivated by the fact that uncovering miRNA-disease functional association is a key step to understand disease development, another integrative approach was built to predict such associations. The integrative approach showed that integrating disease signature from microarray data and PubMed abstract with miRNA target interactions, to build miRNA-disease functional associations, showed promise to uncover significant associations between diseases and miRNAs. Regression model demonstrated effectiveness for miRNA enrichment analysis. Integrating multiple data sources and biological networks to define more accurate disease and miRNA signature uncovered novel biological associations between miRNAs and disease. Newly predicted miRNAs associated with prostate cancer showed diagnostic and prognostic potential. This concludes that the proposed model gives more insight into disease and functional role of miRNAs in disease development.

8.2. Future Work

For future work, I will consider integrating other information sources for the hope that more information with further improve the results. For example gene to phenotypes analysis can be used to further filter and repriotize genes related to a particular disease or drug. Furthermore, since we are dealing with a biological system (cellular system) it is most probably that all biological entities inside are interacting. Thus I will try to build a more comprehensive integrative approach that uses metabolic network analysis and miRNA-gene interaction to predict a more extensive drug-disease connectivity map.

As I mentioned in the limitations section, further improvements can be done on the text-mining technique. I will be working on developing a more comprehensive framework to prioritize genes that goes beyond co-occurrences and simple NLP based approaches. More accurately, I will be working on a framework that is able to extract abstracts and analyze them both syntactically and semantically to extract more accurate drug-gene and disease-gene associations. Noteworthy that such a framework might need to balance between complexity that might result in more accurate results and simplicity that might result in computationally less intensive operations.

Finally I will improve the system by utilizing social networks measures and techniques. Since biological networks are very similar to social networks in their properties, different centrality measures (degree, closeness, betweenness and eigenvector) might uncover some really valuable information in predicting drug-disease associations.

References

[1] H. A. Mucke, "Drug Repositioning: Extracting Added Value from Prior R&D Investments," [Online]. Available: http://www.insightpharmareports.com/uploadedFiles/Reports/Reports/Drug_Repositioning/DrugRepositioning_Report_SamplePages.pdf. [Accessed 5 April 2012].

[2] M. Sirota, J. Dudley, J. Kim, A. Chiang, A. Morgan, A. Sweet-Cordero, J. Sage and A. Butte, "Discovery and preclinical validation of drug indications using compendia of public gene expression data," *ci Transl Med,* pp. 77-96, 2011.

[3] J. DiMasi, R. Hansen and H. Grabowski, "The price of innovation: new estimates of drug development costs," *J Health Econ,* vol. 22(2), pp. 151-185, 2003.

[4] B. Lewin, Genes VIII, Pearson Education, Inc, 2004.

[5] I. Lobo, "Environmental influences on gene expression," *Nature Education,* vol. 1, 2008.

[6] V. Kulasingam and E. Diamandis, "Strategies for discovering novel cancer biomarkers through utilization of emerging technologies," *Nature Clinical Practice Oncology,* vol. 10, pp. 588-99, 2008.

[7] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield and E. and Lander, "Molecular Classification of Cancer: Class Discovery and Class Predictiion by Gene Expression Monitoring," *Science,* vol. 286, pp. 531-537, 1999.

[8] P. Charles, S. Therese, E. Michael, d. Matt, J. Stefanie, R. Christian, P. Jonathan, R. Douglas, J. Hilde, A. Lars, F. Oystein, P. Alexander, W. Cheryl and Z. Shirley, "Portraits of Human Breast Tumor," *Nature,* vol. 406, pp. 747-752, 2000.

[9] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack and A. and Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Science,* vol. 96, pp. 6745-6750, 1999.

[10] J. Lamb, E. Crawford, D. Peck, J. Modell, I. Blat, M. Wrobel, J. Lerner, J. Brunet, A. Subramanian, K. Ross, M. Reich, H. Hieronymus, i. G. We, S. Armstrong, S. Haggarty, P. Clemons, i. R. We, S. Carr, E. Lander and T. Golub, "The Connectivity Map: using gene-expression to connect small molecules, genes and disease.," *Science,* vol. 313, pp. 1929-1935, 2006.

[11] L. Hunter, r. taylor, s. leach and r. and simon, "GEST: gene expression search tool based on a novel bayesian similarity metric," *Bioinformatics,* pp. s115-s122, 2001.

[12] M. E. Wall, R. Andreas, M. R. Luis, D. Berrar, W. Dubitzky and M. Granzow, Singular value decomposition and principal component analysis in A Practical Approach to Microarray Data Analysis, Norwell, MA : Kluwer, 2003.

[13] V. R, G. A, L. M and H. D, "Novel unsupervised feature filtering of biological data," *Bioinformatics,* vol. 22(14), pp. e507-e513, 2006.

[14] A. O, B. PO and B. D, "Singular value decomposition for genome-wide expression data processing and modeling," *Proc Natl Acad Sci U S A,* vol. 97(18), pp. 10101-10106, 2000.

[15] F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi and D. di Bernardo, "Discovery of drug mode of action and drug repositioning from transcriptional responses," *Proc Natl Acad Sci U S A,* vol. 107(33), pp. 14621-14626, 2010.

[16] i. J. DeRis, L. Penaland, P. Brown, M. Bittner, P. Meltzer and M. Ray, "Use of cDNA microarray to analyse gene expression patterns in human cancer 14:457–460.," *Nature Genetic,* vol. 14, pp. 457-460, 1996.

[17] Y. Chen, E. Dougherty and M. Bittner, "Ratio-based decisions and the quantitative analysis of cDNA microarray image," *Biomedical Optics,* vol. 2, pp. 364-374, 1997.

[18] A. Long, H. Mangalam, B. Chan, H. ,. W. TolleriL and P. Baldi, "Improved statistical inference from DNA microarray data analysis using analysis of variance and Bayesian statistical framework," *Biological Chemistry,* vol. 276, pp. 19937-19944, 2001.

[19] S. Dudoit, Y. Yang, T. Speed and M. Callow, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Statistica Sinica,* vol. 12, pp. 111-139, 2002.

[20] V. Tusher, R. Tibshirani and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences,* vol. 98, pp. 5116-51121, 2001.

[21] K. Seo, L. Jae and S. Suk, "Comparison of various statistical methods for identifying differential gene expression in replicated microarray data," *Statistical Methods in Medical Research,* vol. 15, pp. 3-20, 2006.

[22] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques, Waltham, MA, USA: Morgan

Kaufmann, 2012.

[23] D. Jiang, C. Tang and Aidong Zhang, "Cluster analysis for gene expression data: a survey," *Knowledge and Data Engineering, IEEE Transactions,* vol. 16, pp. 1370-1386, 2004.

[24] T. C., Z. L., Z. A and R. M, "Interrelated two way clustering: An unsupervised approach for gene expression data analysis," in *BIBE: 2nd IEEE International Symposium on Bioinformatics and Bioengineering* , Maryland, 2001.

[25] Tang, Chun and Zhang, "An iterative strategy for pattern discovery in high-dimensional data sets," in *Proceedings of 11th Internation Conference on Information and Knowledge Management*, McLean, 2002.

[26] H. LJ, K. S and Y. S, "Exploring expression data: identification and analysis of coexpressed genes," *Genome Research,* vol. 9(11), pp. 1106-1115, 1999.

[27] M. Eisen, T. Spellman, P. Brown and D. and Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Science,* vol. 95, pp. 14863-14868, 1998.

[28] S. Tavazoie, D. Hughes, M. Campbell, R. Cho and G. Church, "Systematic determination of genetic network architecture," *Nature Genetic,* pp. 281-285, 1999.

[29] C. Marcilio, I. ,. G. de Souto, D. S. Costa, T. ,. B. de Araujo and a. A. S. Ludermir, "Clustering cancer gene expression data: a comparative study.," *Bioinformatics,* pp. 497-506, 2008.

[30] A. J. M. Ng, "On Discriminative vs. Gnerative Classifiers: A comparison of Logistic Regression and Naive Bayes," *Neural Information Processing Systems,* 2002.

[31] A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander and J. Mesirov, "Gene set enrichment analysis: a knowledge based approach for interpreting genome-wide expression profiels," *Proceedings of the National Academy of Science,* vol. 43, pp. 15545-15550, 2005.

[32] F. Ingo, H. Kurt and M. David, "Text Mining Infrastructure in R," *Journal of Statistical Software,* vol. 25, 2008.

[33] L. Jensen, J. Saric and P. Bork, "Literature mining for the biologist: from information retrieval to biological discovery," *Nat Rev Genet,* vol. 7, pp. 119-129, 2006.

[34] M. Krallinger, R. Erhardt and A. Valencia, "Text-mining approaches in molecular biology and biomedicine," *Drug Discov Today,* vol. 10, pp. 439-445, 2005.

[35] L. Hirschman, J. Park, J. Tsujii, L. Wong and C. Wu, "Accomplishments and challenges in literature data mining for biology," *Bioinformatics,* vol. 18, pp. 1553-15561, 2002.

[36] Y. Yang, S. Adelstein and A. Kassis, "Target discovery from data mining approaches," *Drug Discov Today,* vol. 17, pp. s16-s23, 2012.

[37] J. Wren and H. Garner, "Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network Bioinformatics," *Bioinformatics,* vol. 20, pp. 191-199, 2004.

[38] J. Ding, D. Berleant, D. Nettleton and E. Wurtele, "Mining MEDLINE: abstracts, sentences, or phrases?," *Pac Symp Biocomput,* pp. 236-337, 2002.

[39] P. Bowers, M. Pellegrini, M. Thompson, J. Fierro, T. Yeates, D. Eisenberg and Prolinks, "a database of protein functional linkages derived from coevolution," *Genome Biol,* vol. 5, 2004.

[40] J. Cooper and A. Kershenbaum, "Discovery of protein-protein interactions using a combination of linguistic, statistical and graphical information," *BMC Bioinformatics,* pp. 143-149, 2005.

[41] A. Ramani, R. Bunescu, R. Mooney and E. Marcotte, "Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome," *Genome Biol,* vol. 6, 2005.

[42] D. Hristovski, B. Peterlin, J. Mitchell and S. Humphrey, "Using literature-based discovery to identify disease candidate genes," *Int J Med Inform,* vol. 74, pp. 289-298, 2005.

[43] D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju and D. Wishart, "PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites," *Nucleic Acid Research,* vol. 36, pp. 399-405, 2008.

[44] T. Werner, "Bioinformatics applications for pathway analysis of microarray data," *Curr Opin Biotechnol,* vol. 19, pp. 50-54, 2008.

[45] A. Faro, D. Giordano and C. Spampinato, "Combining literature text mining with microarray data: advances for system biology modeling," *Brief Bioinform,* vol. 13(1), pp. 61-82, 2012.

[46] L. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork and C. von Mering, "STRING 8--a global view on proteins and their functional interactions in 630 organisms," *Nucleic Acids Res,* vol. 37, pp. 412-418, 2009.

[47] R. Hoffmann and A. Valencia, "A gene network for navigating the literature 2004," *Nat Genet,* vol. 36, pp. 664-671, 2004.

[48] M. Krauthammer, C. Kaufmann, T. Gilliam and A. Rzhetsky, "Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease," *Proceedings of the National Academy of Science, USA,* vol. 101(42), pp. 15148-15153, 2004.

[49] J. Aronson, "Old drugs--new uses," *Br J Clin Pharmacol,* vol. 64(5), pp. 563-568, 2007.

[50] S. Sundar, T. Jha, C. Thakur, S. Bhattacharya and M. Rai, "Oral miltefosine for the treatment of Indian visceral leishmaniasis," *Trans R Soc Trop Med Hyg,* vol. 1000, pp. S26-S33, 2006.

[51] J. Li, X. Zhu and J. Chen, "Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts," *PLoS Comput Biol,* vol. 5(7), 2009.

[52] Y. Yamanishi, M. Akari, A. Gutteridge, W. Honda and M. Kanehisa, "Prediction of drug-target interaction networks from the integration of chemical and genomic spaces," *Bioinformatics,* pp. 232-240, 2008.

[53] K. Bleakley and Y. Yamanishi, "Supervised prediction of drug-target interactions using bipartite local models," *Bioinformatics,* vol. 25(18), pp. 2397-2403, 2009.

[54] D. di Bernardo, M. Thompson, T. Gardner, S. Chobot, E. Eastwood, A. Wojtovich, S. Elliott, S. Schaus and J. Collins, "Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks," *Nat Biotechnol,* vol. 23(3), pp. 377-383, 2005.

[55] A. Gottlieb, G. Stein, E. Ruppin and R. Sharan, "PREDICT: a method for inferring novel drug indications with application to personalized medicine," *Mol Syst Biol,* pp. 490-496, 2011.

[56] P. Robinson and S. Mundlos, "The human phenotype ontology," *Clin Genet,* vol. 77(6), pp. 525-534, 2010.

[57] M. Chee, R. Yang, E. Hubbell, A. Berno, X. Huang, D. Stern, J. Winkler, D. Lockhart, M. Morris and S. Fodor, "Accessing genetic information with high-density DNA arrays," *Science,* vol. 274(5287), pp. 610-614, 1996.

[58] G. Hu and P. Agarwal, "Human disease-drug network based on genomic expression profiles," *PloS One,* vol. 4(8), 2009.

[59] P. Diaconis and R. Graham, "Spearman’s footrule as a measure of disarray," *J R Stat Soc,* vol. 39, pp. 262-268, 1977.

[60] S. Lin, "Space oriented rank-based data integration," *Stat Appl Genet Mol Biol,* vol. 9(1), 2010.

[61] A. Ozgür, T. Vu, G. Erkan and D. Radev, "Identifying gene-disease associations using centrality on

a literature mined gene-interaction network," *Bioinformatics,* vol. 24(13), pp. 277-285, 2008.

[62] B. Junker, D. Koschützki and F. Schreiber, "Exploration of biological network centralities with CentiBiN," *BMC Bioinformatics,* pp. 212-219, 2006.

[63] J. Chen, B. Aronow and A. Jegga, "Disease candidate gene identification and prioritization using protein interaction networks," *BMC Bioinformatics,* pp. 63-73, 2009.

[64] A. Hopkins and C. Groom, "The druggable genome," *Nat Rev Drug Discov,* vol. 1(9), pp. 727-730, 2002.

[65] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. Guo and D. Wishart, "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs," *Nucleic Acids Res,* vol. 39, pp. D1035-D1041, 2011.

[66] M. Baltimore, "Online Mendelian Inheritance in Man," Johns Hopkins University, [Online]. Available: http://omim.org/. [Accessed 1 Feburary 2012].

[67] G. Wu, X. Feng and L. Stein, "A human functional protein interaction network and its application to cancer data analysis," *Genome Biol,* vol. 11(5), 2010.

[68] M. Severgnini, S. Bicciato, E. Mangano, F. Scarlatti, A. Mezzelani, M. Mattioli, R. Ghidoni, C. Peano, R. Bonnal, F. Viti, L. Milanesi, G. De Bellis and C. Battaglia, "Strategies for comparing gene expression profiles from different microarray platforms: application to a case control experiment," *Anal Biochem,* vol. 353(1), pp. 43-56, 2006.

[69] T. Barrett, D. Troup, S. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. Kim, A. Soboleva, M. Tomashevsky and R. Edgar, "NCBI GEO: mining tens of millions of expression profiles--database and tools update," *Nucleic Acids Res,* pp. D760-D765, 2007.

[70] R. Irizarry, B. Bolstad, F. Collin, L. Cope, B. Hobbs and T. Speed, "Summaries of Affymetrix GeneChip probe level data," *Nucleic Acids Res,* vol. 31(4), 2003.

[71] J. Bernini, Z. Rogers, E. Sandler, J. Reisch, C. Quinn and G. Buchanan, "Beneficial effect of intravenous dexamethasone in children with mild to moderately severe acute chest syndrome complicating sickle cell disease," *Blood,* vol. 92(9), pp. 3082-3089, 1998.

[72] N. Gupta, S. Kavuru, D. Patel, D. Janson, N. Driscoll and e. al, "Rituximab-based chemotherapy for steroid-refractory autoimmune hemolytic anemia of chronic lymphocytic leukemia," *Leukemia,* vol. 16(10), pp. 2092-2097, 2002.

[73] K. Hatano, K. Nishimura, Y. Nakai, T. Yoshida, M. Sato and e. al, "Weekly low-dose docetaxel combined with estramustine and dexamethasone for Japanese patients with metastatic

castration-resistant prostate cancer," *Int J Clin Oncol,* pp. 10147-10159, 2012.

[74] A. Cozzoli, B. Nico, V. Sblendorio, R. Capogrosso, M. Dinardo and e. al, "Enalapril treatment discloses an early role of angiotensin II in inflammation- and oxidative stress-related muscle damage in dystrophic mdx mice," *Pharmacol Res,* vol. 64(5), pp. 482-492, 2011.

[75] C. Ramaciotti, L. Heistein, M. Coursey, M. Lemler, R. Eapen and e. al, "Left ventricular function and response to enalapril in patients with duchenne muscular dystrophy during the second decade of life," *Am J Cardiol,* vol. 98(6), pp. 825-832, 2006.

[76] A. Naderi and J. Liu, "Inhibition of androgen receptor and Cdc25A phosphatase as a combination targeted therapy in molecular apocrine breast cancer," *Cancer Lett,* vol. 298(1), pp. 74-87, 2010.

[77] A. Naderi and L. Hughes-Davies, "A functionally significant cross-talk between androgen receptor and ErbB2 pathways in estrogen receptor negative breast cancer," *Neoplasia,* vol. 10(6), pp. 542-550, 2008.

[78] H. Yoshida, T. Imamura, A. Fujiki, Y. Hirashima, M. Miyachi and e. al, "Post-transcriptional modulation of C/EBPα prompts monocytic differentiation and apoptosis in acute myelomonocytic leukaemia cells," *Leuk Res,* vol. 36(6), pp. 735-741, 2012.

[79] C. Zhang, Y. Ryu, T. Chen, C. Hall, D. Webster and M. Kang, "Synergistic activity of rapamycin and dexamethasone in vitro and in vivo in acute lymphoblastic leukemia via cell-cycle arrest and apoptosis," *Leuk Res,* vol. 36(3), pp. 342-351, 2012.

[80] D. Pan, Y. Li, Z. Li, Y. Wang, P. Wang and Y. Liang, "Gli inhibitor GANT61 causes apoptosis in myeloid leukemia cells and acts in synergy with rapamycin," *Leuk Res,* vol. 36(6), pp. 742-750, 2012.

[81] H. Bird and M. Broggini, "Paroxetine versus amitriptyline for treatment of depression associated with rheumatoid arthritis: a randomized, double blind, parallel group study," *J Rheumatol,* vol. 27(12), pp. 2791-2798, 2000.

[82] R. Frank, J. Kashani, J. Parker, N. Beck, M. Brownlee-Duffeck and e. al, "Antidepressant analgesia in rheumatoid arthritis," *J Rheumatol,* vol. 15(11), pp. 1632-1640, 1988.

[83] M. Miyata, T. Asano and S. Satoh, "Effect of additional administration of tacrolimus in patients with rheumatoid arthritis treated with biologics," *Fukushima J Med Sci,* vol. 57(2), pp. 54-63, 2011.

[84] M. Kitahama, H. Okamoto, Y. Koseki, E. Inoue, H. Kaneko and e. al, "Efficacy and safety of tacrolimus in 101 consecutive patients with rheumatoid arthritis," *Mod Rheumatol,* vol. 20(5),

pp. 478-485, 2010.

[85] S. Kawai, T. Takeuchi, K. Yamamoto, Y. Tanaka and N. Miyasaka, "Efficacy and safety of additional use of tacrolimus in patients with early rheumatoid arthritis with inadequate response to DMARDs--a multicenter, double-blind, parallel-group trial," *Mod Rheumatol,* vol. 7(9), pp. 458-468, 2011.

[86] J. Symes, M. Kurin, N. Fleshner and J. Medin, "Fas-mediated killing of primary prostate cancer cells is increased by mitoxantrone and Docetaxel," *Mol Cancer Ther,* vol. 7(9), pp. 3018-3028, 2008.

[87] A. Pinto, J. Moreira and S. Simões, "Liposomal imatinib-mitoxantrone combination: formulation development and therapeutic evaluation in an animal model of prostate cancer," *Prostate,* vol. 71(1), pp. 81-90, 2011.

[88] C. Thomas, B. Hadaschik, J. Thüroff and C. Wiesner, "Patients with metastatic hormone-refractory prostate cancer. Second-line chemotherapy with mitoxantrone plus prednisone," *Urologe A,* vol. 48(9), pp. 1070-1074, 2009.

[89] S. Sevli, A. Uzumcu, M. Solak, M. Ittmann and M. Ozen, "The function of microRNAs, small but potent molecules, in human prostate cancer," *Prostate Cancer Prostatic Dis,* vol. 13(3), pp. 208-217, 2010.

[90] V. Ambros, "The functions of animal microRNAs," *Nature,* vol. 431(7006), pp. 350-355, 2004.

[91] A. Gordanpour, R. Nam, L. Sugar and A. Seth, "MicroRNAs in prostate cancer: from biomarkers to molecularly-based therapeutics," *Prostate Cancer Prostatic Dis,* vol. 15(4), pp. 314-323, 2012.

[92] M. Ozen, C. Creighton, M. Ozdemir and M. Ittmann, "Widespread deregulation of microRNA expression in human prostate cancer," *Oncogene,* vol. 27(12), pp. 1788-1793, 2008.

[93] B. Zhang, X. Pan, G. Cobb and T. Anderson, "microRNAs as oncogenes and tumor suppressors," *Dev Biol,* vol. 302(1), pp. 1-12, 2007.

[94] G. Calin, C. Sevignani, C. Dumitru, T. Hyslop, E. Noch, S. Yendamuri, M. Shimizu, S. Rattan, F. Bullrich, M. Negrini and C. Croce, "Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers," *Proc Natl Acad Sci U S A,* vol. 101(9), pp. 2999-3004, 2004.

[95] M. Sualp and T. Can, "Using network context as a filter for miRNA target prediction," *Biosystems,* vol. 105(3), pp. 201-209, 2011.

[96] A. Grimson, K. Farh, W. Johnston, P. Garrett-Engele, L. Lim and D. Bartel, "MicroRNA targeting

specificity in mammals: determinants beyond seed pairing," *Mol Cel,* vol. 27(1), pp. 91-105, 2007.

[97] D. Yue, H. Liu and Y. Huang, "Survey of Computational Algorithms for MicroRNA Target Prediction," *Curr Genomics,* vol. 10(7), pp. 478-492, 2009.

[98] G. Wu, X. Feng and L. Stein, "A human functional protein interaction network and its application to cancer data analysis," *Genome Biol,* vol. 11(5), 2010.

[99] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang and Y. Liu, "miR2Disease: a manually curated database for microRNA deregulation in human disease," *Nucleic Acids Res,* pp. D98-D104, 2009.

[100] R. Hudson, M. Yi, D. Esposito, S. Watkins, A. Hurwitz, H. Yfantis, D. Lee, J. Borin, M. Naslund, R. Alexander, T. Dorsey, R. Stephens, C. Croce and S. Ambs, "MicroRNA-1 is a candidate tumor suppressor and prognostic marker in human prostate cancer," *Nucleic Acids Res,* vol. 40(8), pp. 3689-3703, 2012.

[101] A. Antonov, S. Dietmann, P. Wong, D. Lutter and H. Mewes, "GeneSet2miRNA: finding the signature of cooperative miRNA activities in the gene lists," *Nucleic Acids Res,* pp. W323-W328, 2009.

[102] Y. Lu, Y. Zhou, W. Qu, M. Deng and C. Zhang, "A Lasso regression model for the construction of microRNA-target regulatory networks," *Bioinformatics,* vol. 27(17), pp. 2406-2419, 2011.

[103] C. Chong and D. J. Sullivan, "New uses for old drugs," *Nature,* vol. 448(7154), pp. 645-651, 2007.

[104] D. Maglott, J. Ostell, K. Pruitt and T. Tatusova, "Entrez Gene: gene-centered information at NCBI," *Nucleic Acids Res,* pp. D54-D62, 2005.

[105] N. Andrew, "CoursEra," 5 4 2012. [Online]. Available: https://class.coursera.org/ml-003/class/index. [Accessed 5 4 2013].