Indian Statistical Institute
Kolkata

Insilico Identification of Disease Genes Using Microarray Data and Protein-Protein Interaction Networks

Ekta Shah

Roll Number : MTC1303

Thesis submitted in partial fulfilement of the requirement for the degree of Master of Technology in Computer Science at Indian Statistical Institute, 2015

M.TECH (CS) DISSERTATION THESIS COMPLETION CERTIFICATE

Student : Ekta Shah (MTC1213)
Topic : Insilico Identification of Disease Genes Using Microarray Data and Protein-Protein Interaction Networks
Supervisor : Dr. Pradipta Maji

This is to certify that the thesis titled "Insilico Identification of Disease Genes Using Microarray Data and Protein-Protein Interaction Networks" submitted by Ekta Shah in partial fulfillment for the award of the degree of Master of Technology is a bonafide record of work carried out by her under my supervision. The thesis has fulfilled all the requirements as per the regulations of this Institute and, in my opinion, has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other university for the award of any degree or diploma.

Dr. Pradipta Maji
Associate Professor
Machine Intelligence Unit
Indian Statistical Institute,Kolkata

Abstract

One of the important problems in functional genomics is how to select the disease genes. In this regard, the paper presents a new similarity measure to compute the functional similarity between two genes. It is based on the information of protein-protein interaction networks. A new gene selection algorithm is introduced to identify disease genes, integrating judiciously the information of gene expression profiles and protein-protein interaction networks. The proposed algorithm selects a set of genes from microarray data as disease genes by maximizing the relevance and functional similarity of the selected genes. The performance of the proposed algorithm, along with a comparison with other related methods, is demonstrated on colorectal cancer data set.

Acknowledgements

I wish to express my deep sense of gratitude to Dr. Pradipta Maji, for his guidance, encouragement and facilities extended without which this project would not have taken its present shape.
I am also grateful to Dr. Sushmita Paul, for her continuous guidance for successfully completing this project.
I would like to thank my parents and all my faculties and friends and everyone else who helped me with this project. I am very much grateful to all those who have helped me in understanding the topics. The discussions that I had with them have helped me to attain deeper understanding of the topic. Though I wanted to quote their names in the acknowledgements because of the constraints I am not able to do so. But their help has always been remembered.
I would also like to thank "Biomedical Imaging & Bioinformatics Lab"and "Machine Intelligence Unit", Indian Statistical Institute providing me with ample resources and helped me throughout this project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Genetic diseases such as Alzheimer's disease, breast cancer, leukemia, colorectal cancer, down syndrome, and heart disease are caused by abnormalities in genes or chromosomes. A genetic disease may be heritable disorder or may not be. While some genetic diseases are passed down from the parent's genes, others are frequently caused by new mutations or changes to the DNA. In other instances, the same disease, for example, some forms of cancer, may stem from an inherited genetic condition in some people, from new mutations in some people, and from non-genetic causes in other people. As the term genetic disease suggests, these diseases are caused by the dysfunction of some genes. Therefore, such genes are better known as disease genes [3].

Recent advancement and wide use of high-throughput biotechnologies have been producing an explosion in using gene expression phenotype for understanding the function of disease genes [11, 30]. Analyzing the difference of gene expression levels in particular cell types may provide an idea about the propensity of a disease. Specifically, if a set of genes shows a consistent pattern of different expression levels in sick subjects and a control group, then that gene set is likely a strong candidate of playing a pathogenic role. Differences in expression levels can be detected primarily by microarray studies. In this background, microarray gene expression data has been widely used for identification of disease genes using different feature selection algorithms [20, 43, 52, 65].

In [6, 26], it has been shown that the genes associated with the same disorder tend to share common functional features, reflecting that their protein products have a tendency to interact with each other. Hence, another indicative trait of a disease gene is that its protein product is strongly linked to other disease-gene proteins. In this background, the protein-protein interaction (PPI) data have been used in various studies to identify disease genes [39, 57]. Individually microarray data or the PPI network data can be used to identify potential disease genes, although there is a limited chance of finding novel disease genes from such an analysis. In this regard, data integration methods have been developed to identify pleiotropic genes involved in the physiological cellular processes of many diseases.

The integrated approaches assume that the protein products of disease genes tend to be close to differentially expressed genes in the protein interaction network. Chao et al. [81] developed a method by integrating gene expression data and the PPI network data to prioritize cancer-associated genes. Zhao et al. [87] also proposed an approach by integrating gene expression data and the PPI network data to select disease genes. Jia et al. [34] developed a dense module searching method to identify disease genes for complex diseases by integrating the association signal from genome wide association studies data sets into the human PPI network. Li and Li [46] developed another approach to identify candidate disease genes, where heterogeneous genomic and phenotype data sets are used. In this method, separate gene networks are first developed using different types of data sets. The various genomic networks are then merged into a single graph, and disease genes are identified using random walk. In [43], minimum redundancy-maximum relevance (mRMR) [20] approach has been used to select a set of genes from expression data, while maximum relevance-maximum significance (MRMS) criterion [52] has been used in [65]. The selected gene set is then used for identification of intermediate genes between a pair of selected genes using the PPI network data. However, all the methods reported earlier consider gene expression and PPI data separately while selecting candidate genes.

In this regard, this thesis presents a new gene selection algorithm to identify disease genes. It selects a set of disease genes by maximizing the relevance

and functional similarity of the selected genes. A new similarity measure is introduced to compute the functional similarity between two genes. The proposed algorithm judiciously integrates the information of gene expression profiles and PPI networks. The mutual information is employed to compute the relevance of the genes with respect to class labels based on gene expression profiles, while the PPI data is used to calculate the functional similarity between two genes. The mutual information is used to select differentially expressed genes as disease genes using gene expression profiles, on the other hand, the functional protein association network is used to study the mechanism of diseases. The performance of the proposed algorithm, along with a comparison with other related methods, is demonstrated on colorectal cancer data set. An important finding is that the proposed algorithm is shown to be effective for selecting relevant and functionally similar genes from microarray data, and the identified genes are significantly linked with colorectal cancer. Extensive experimental study on colorectal cancer establishes the fact that the genes identified by the proposed method have more colorectal cancer genes than those identified by existing methods and using the gene expression profiles alone, irrespective of any gene selection algorithm. All the results indicate that the proposed method is quite promising and may become a useful tool for identifying disease genes.

Chapter 2

Basic Concept of Microarray Data and Protein-Protein Interaction Network

2.1 Microarray Data

Microarray technology has become one of the indispensable tools that many biologists use to monitor genome wide expression levels of genes in a given organism. A microarray is typically a glass slide on to which DNA molecules are fixed in an orderly manner at specific locations called spots (or features). A microarray may contain thousands of spots and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene. The DNA in a spot may either be genomic DNA or short stretch of oligo-nucleotide strands that correspond to a gene. The spots are printed on to the glass slide by a robot or are synthesized by the process of photolithography.

Microarrays may be used to measure gene expression in many ways, but one of the most popular applications is to compare expression of a set of genes from a cell maintained in a particular condition (condition A) to the same set of genes from a reference cell maintained under normal conditions (condition B). Figure 2.1 gives a general picture of the experimental steps

involved. First, RNA is extracted from the cells. Next, RNA molecules in the extract are reverse transcribed into cDNA by using an enzyme reverse transcriptase and nucleotides labeled with different fluorescent dyes. For example, cDNA from cells grown in condition A may be labeled with a red dye and from cells grown in condition B with a green dye. Once the samples have been differentially labeled, they are allowed to hybridize onto the same glass slide. At this point, any cDNA sequence in the sample will hybridize to specific spots on the glass slide containing its complementary sequence. The amount of cDNA bound to a spot will be directly proportional to the initial number of RNA molecules present for that gene in both samples.

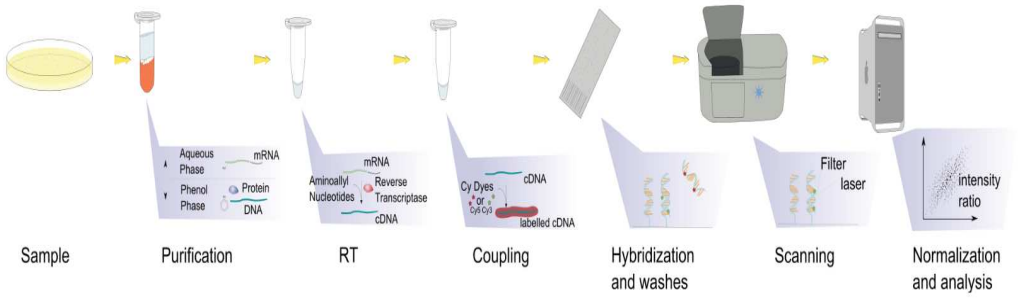

Figure 2.1: Steps required in a microarray experiment

Following the hybridization step, the spots in the hybridized microarray are excited by a laser and scanned at suitable wavelengths to detect the red and green dyes. The amount of fluorescence emitted upon excitation corresponds to the amount of bound nucleic acid. For instance, if cDNA from condition A for a particular gene was in greater abundance than that from condition B, one would find the spot to be red. If it was the other way, the spot would be green. If the gene was expressed to the same extent in both conditions, one would find the spot to be yellow, and if the gene was not expressed in both conditions, the spot would be black. Thus, what is seen at the end of the experimental stage is an image of the microarray, in which each spot that corresponds to a gene has an associated fluorescence value representing the relative expression level of that gene. Further image processing normalization and transformation steps are carried out to generate the gene expression data.

Figure 2.2 gives a general idea of a typical microarray gene expression data. Every row in the expression data represents a gene and every column represents the samples, either diseased or non-diseased. Every entry in the matrix basically represents the expression level of a gene in a particular sample. The last row defines the class to which every sample belongs, i.e. diseased or non-diseased. Thus, the microarray data can be analyzed to monitor the expression level of a set of genes in a particular region over sick and control groups.

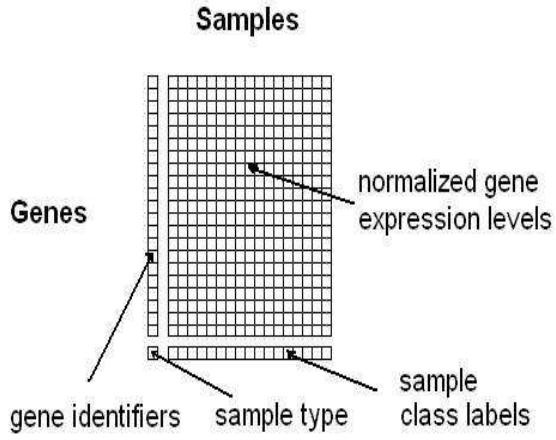

Figure 2.2: Microarray Gene Expression Data Matrix

2.1.1 Image Processing

The first step in the analysis of microarray data is to process the image. Most manufacturers of microarray scanners provide their own software; however, it is important to understand how data is actually being extracted from images, as this represents the primary data collection step and forms the basis of any further analysis. Image processing involves identification of the spots and distinguishing them from spurious signals, determination of the spot area to be surveyed, determination of the local region to estimate background hybridization, and reporting summary statistics and assigning spot intensity after subtracting for background intensity.

2.1.2 Expression Ratios - Primary Comparison

The relative expression level for a gene can be measured as the amount of red or green light emitted after excitation. The most common metric used to relate this information is called expression ratio. It is denoted here as T_k and defined as follows:

$$T_k = \frac{R_k}{G_k} \tag{2.1}$$

where, for each gene k on the array, R_k and G_k represent the spot intensity metric for the test sample and the reference sample, respectively. The spot intensity metric for each gene can be represented as a total intensity value or a background subtracted median value. If we choose the median pixel value, then the median expression ratio for a given spot is:

$$T_{\text{median}} = \frac{R^{\text{spot}}_{\text{median}} - R^{\text{background}}_{\text{median}}}{G^{\text{spot}}_{\text{median}} - G^{\text{background}}_{\text{median}}} \tag{2.2}$$

where $R^{\text{spot}}_{\text{median}}$ and $R^{\text{background}}_{\text{median}}$ are the median intensity values for the spot and background, respectively, for the test sample.

2.1.3 Transformations of the Expression Ratio

The expression ratio is a relevant way of representing expression differences in a very intuitive manner. For example, genes that do not differ in their expression level will have an expression ratio of 1. However, this representation may be unhelpful when one has to represent up-regulation and down-regulation. For example, a gene that is up-regulated by a factor of 4 has an expression ratio of 4 (R/G = 4G/G = 4). However, for the case where a gene is down regulated by a factor of 4, the expression ratio becomes 0.25 (R/G = R/4R = 1/4). Thus, up-regulation is blown up and mapped between 1 and infinity, whereas down-regulation is compressed and mapped between 0 and 1.

$$\text{Up} - \text{regulation} \xrightarrow{\text{mapped}} [1, \infty]$$
$$\text{Down} - \text{regulation} \xrightarrow{\text{mapped}} [0, 1]$$

2.1.4 Data Normalization

The expression ratios and their transformations are reasonable measures to detect differentially expressed genes. However, when one compares the expression levels of genes that should not change in the two conditions (say, housekeeping genes), what one quite often finds is that an average expression ratio of such genes deviates from 1. This may be due to various reasons, for example, variation caused by differential labeling efficiency of the two fluorescent dyes or different amounts of starting mRNA material in the two samples. Thus, in the case of microarray experiments, as for any large-scale experiments, there are many sources of systematic variation that affect measurements of gene expression levels. Normalization is a term that is used to describe the process of eliminating such variations to allow appropriate comparison of data obtained from the two samples.

There are many methods of normalization. The first step in a normalization procedure is to choose a gene set, which consists of genes for which expression levels should not change under the conditions studied, that is, the expression ratio for all genes in the gene set is expected to be 1. From that set, a normalization factor, which is a number that accounts for the variability seen in the gene set, is calculated. It is then applied to the other genes in the microarray experiment. One should note that the normalization procedure changes the data, and is carried out only on the background corrected values for each spot. Total intensity normalization and mean log centring are some normalization methods.

2.2 Protein-Protein Interaction Networks

Protein-protein interactions (PPIs) refer to intentional physical contacts established between two or more proteins as a result of biochemical events and/or electrostatic forces.In fact, proteins are vital macromolecules, at both cellular and systemic levels, but they rarely act alone. Diverse essential molecular processes within a cell are carried out by molecular machines that are built from a large number of protein components organized by their PPIs. Indeed, these interactions are at the core of the entire interactomics system of any living cell and so, unsurprisingly, aberrant PPIs are on the basis of multiple diseases, such as Creutzfeld-Jacob, Alzheimer's disease, and cancer.

To achieve a first level of understanding of such cellular complexity we need to unravel the interactions that occur between all the proteins that integrate a living cell. However, the definition of protein-protein interaction intutively is restricted to the physical contact between the two protein surfaces. Methods currently being used a bias towards detection of higher levels of relations or associations between proteins. SUch protein relations may include several factors like inclusion in multiprotein complexes, common cellular compartments, signalling pathways, etc.

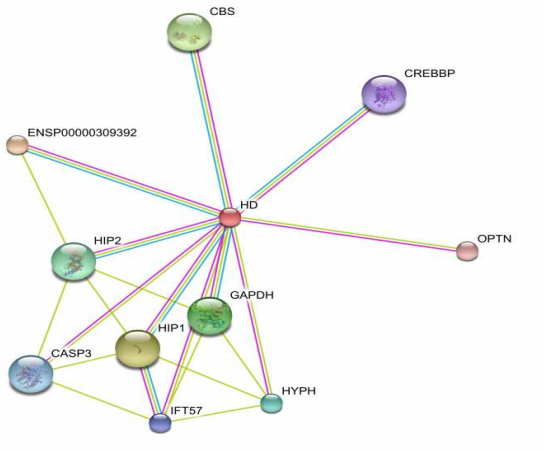

Figure 2.3: Protein-Protein Interaction Network

Large scale identification of PPIs generated hundreds of thousands interactions, which were collected together in specialized biological databases that

are continuously updated in order to provide complete interactomes. The first of these databases was the Database of Interacting Proteins (DIP).[29] Since that time, the number of public databases has been increasing. Examples of such databases include Biomolecular Interaction Network Database (BIND), Biological General Repository for Interaction Datasets (BioGRID), Human Protein Reference Database (HPRD), Known and Predicted Protein-Protein Interactions (STRING), etc. Figure 2.3 gives an brief idea about a PPI Network.

Chapter 3

Existing Approaches for Disease Gene Identification

3.1 Disease Gene Selection from Microarray Data

3.1.1 Survey on Gene Selection Methods

An important application of gene expression data in functional genomics is to classify samples according to their gene expression profiles, such as to classify cancer versus normal samples or to classify different types or subtypes of cancer [27]. The small number of training samples and a large number of genes/mRNAs make gene/mRNA selection a more relevant and challenging problem in expression based classification [74]. Furthermore, additional experimental complications like noise and variability render the analysis of microarray data an exciting domain. This is an important problem in machine learning and referred to as feature selection. In order to deal with these particular characteristics of microarray data, the obvious need for dimension reduction techniques was realized [1, 7, 27, 69] and soon their application became a de facto standard in the field. Lot of gene selection algorithms have been developed to select differentially expressed genes [72].

Because of the high dimensionality of most microarray analyses, fast and

efficient feature selection techniques such as univariate filter methods [7, 22, 33, 41, 45, 75] have attracted most attention. Univariate methods can be parametric [5, 24, 58, 77] or non-parametric [23, 48, 62, 64, 66, 78]. Univariate techniques are fast, scalable, and independent of the classifier. The simplicity of the univariate techniques has made it dominant in the field of gene selection using microarray data. Univariate selection methods have certain restrictions and may lead to less accurate classifiers by, for example, not taking into account gene-gene interactions. Thus, researchers have proposed techniques that try to capture these correlations between genes.

The application of multivariate filter methods ranges from simple bivariate interactions [9] towards more advanced solutions exploring higher order interactions, such as correlation-based feature selection [80, 85] and several variants of the Markov blanket filter method [25, 54, 82]. The minimum redundancy-maximum relevance [19, 67] and uncorrelated shrunken centroid [86] algorithms are two other solid multivariate filter procedures, highlighting the advantage of using multivariate methods over univariate procedures in the gene expression domain. A f-information measure based method has been proposed in [50] for selection of discriminative genes from microarray data. Another gene selection algorithm based on rough-fuzzy hybridization is given in [51].

Feature selection using wrapper or embedded methods offers an alternative way to perform a multivariate gene subset selection, incorporating the classifiers; bias into the search and thus offering an opportunity to construct more accurate classifiers. In the context of microarray analysis, most wrapper methods use population-based, randomized search heuristics [8, 36, 44, 60], although also a few examples use sequential search techniques [32, 83]. An interesting hybrid filter-wrapper approach is introduced in [70], crossing a univariately pre-ordered gene ranking with an incrementally augmenting wrapper method.

The embedded capacity of several classifiers to discard input features and thus propose a subset of discriminative genes, has been exploited by several authors. Examples include the use of random forests (a classifier that combines many single decision trees) in an embedded way to calculate

the importance of each gene [35, 79]. Another line of embedded feature selection techniques uses the weights of each feature in linear classifiers, such as support vector machines [28] and logistic regression [49]. These weights are used to reflect the relevance of each gene in a multivariate way, and thus allow for the removal of genes with very small weights.

Partially due to the higher computational complexity of wrapper and to a lesser degree embedded approaches, these techniques have not received as much interest as filter proposals. However, an advisable practice is to pre-reduce the search space using a univariate filter method, and only then apply wrapper or embedded methods, hence fitting the computation time to the available resources. Other notable gene selection algorithms [47, 55, 63, 73, 84] are also developed for selection of genes from microarray data.

3.1.2 Statistical Tests

To measure the relevance of a gene, the t-value is widely used in the literature. Assuming that there are two classes of samples in a gene expression data set, the t-value $t(\mathbb{G}_i)$ for gene $\mathbb{G}_i$ is given by:

$$t(\mathbb{G}_i) = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \tag{3.1}$$

where μ_c and σ_c are the mean and the standard deviation of the expression levels of gene $\mathbb{G}_i$ for class c, respectively, and n_c is the number of samples in class c for $c = 1, 2$. When there are multiple classes of samples, the t-value is typically computed for one class versus all the other classes.

For multiple classes of samples, an F-statistic between a gene and the class label can be used to calculate the relevance score of that gene. The F-statistic value of gene $\mathbb{G}_i$ in K classes denoted by $\mathbb{C}$ is defined as follows:

$$F(\mathbb{G}_i, \mathbb{C}) = \left[\sum_{c=1}^{K} n_c (\bar{w}_{ic} - \bar{w}_i)^2 / (K-1) \right] / \sigma^2 \tag{3.2}$$

where $\bar{w}_i$ is the mean of w_{ij} in all samples, $\bar{w}_{ic}$ is the mean of w_{ij} in the cth

class, K is the number of classes, and $\sigma^2 = [\sum_c (n_c - 1)\sigma_c^2]/(n - c)$ is the pooled variance (where n_c and σ_c are the size and the variance of the cth class). Hence, the F-test reduces to the t-test for two class problem with the relation $F = t^2$.

3.1.3 InfoGain: Maximum Relevance Criterion

When selecting a subset of genes from the microarray data, it is critical to minimize the classification error. Thus, the chosen subset of genes should be such that it minimizes the misclassification error. In an unsupervised situation where classifiers are not specified, minimal error requires maximal dependency of the selected subset of genes to target class. This is known as the Maximal Dependency Scheme. Maximum Relevance Criterion is one of the most popular approaches to realize Maximal Dependency [19, 67]. Relevance is generally represented in terms of Mutual Information.

Given two random variables x and y, their mutual information is defined in terms of their probabilistic density functions *p(x), p(y)* and *p(x,y)*:

$$I(x;y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy \tag{3.3}$$

$$\gamma_x(y) = I(x;y) \tag{3.4}$$

In Max-Relevance, the selected genes x_i are required individually, to have highest mutual information $I(x_i, c)$ with the target class c. Sequential selection is performed to select the m best genes, i.e. the genes are all ranked based on their relevance scores, $\gamma_c(x_i)$ and the top m genes are selected. However, the described method of gene selection doesn't account for cases when the microarray data may contain redundant information. Lets consider two genes having a similar gene expression level in the microarray. In that case the two genes would have a similar relevance level with the class of samples concerned. Thus, if both the genes bear a high relevance to the class of

samples, then the top m genes selected may contain redundant information. Therefore, the maximum relevance criterion alone isn't sufficient for gene selection.

3.1.4 mRMR: Minimum Redundancy Maximum Relevance Criterion

The method ranks genes based on their relevance to the class labels, and meanwhile it can also take the redundancy between genes into account. The genes having best trade-off between the highest relevance to the target class labels and minimum redundancy between genes, were considered as "good biomarkers".

Let $\mathbb{C} = \{\mathcal{A}_1, \cdots, \mathcal{A}_i, \cdots, \mathcal{A}_j, \cdots, \mathcal{A}_m\}$ denotes the set of m genes of a given microarray data set and $\mathbb{S}$ is the set of selected genes. Define $\gamma_{\mathcal{A}_i}(\mathbb{D})$ as the relevance of the gene $\mathcal{A}_i$ with respect to the class label $\mathbb{D}$ while $\lambda(\mathcal{A}_i, \mathcal{A}_j)$ as the redundancy between two genes $\mathcal{A}_i$ and $\mathcal{A}_j$. The total relevance of all selected genes is, therefore, given by

$$\mathcal{J}_{\text{relev}} = \sum_{\mathcal{A}_i \in \mathbb{S}} \gamma_{\mathcal{A}_i}(\mathbb{D}) \tag{3.5}$$

while the total redundancy among the selected genes is

$$\mathcal{J}_{\text{redun}} = \sum_{\mathcal{A}_i, \mathcal{A}_j \in \mathbb{S}} \lambda(\mathcal{A}_i, \mathcal{A}_j). \tag{3.6}$$

Therefore, the problem of selecting a set $\mathbb{S}$ of relevant and nonredundant genes from the whole set $\mathbb{C}$ of m genes is equivalent to maximize $\mathcal{J}_{\text{relev}}$ and minimize $\mathcal{J}_{\text{redun}}$, that is, to maximize the objective function $\mathcal{J}$, where

$$\mathcal{J} = \mathcal{J}_{\text{relev}} - \mathcal{J}_{\text{redun}} = \sum_i \gamma_{\mathcal{A}_i}(\mathbb{D}) - \sum_{i,j} \lambda(\mathcal{A}_i, \mathcal{A}_j). \tag{3.7}$$

To solve the above problem, a greedy algorithm is widely used that follows next [19, 67]:

1. Initialize $\mathbb{C} \leftarrow \{\mathcal{A}_1, \cdots, \mathcal{A}_i, \cdots, \mathcal{A}_j, \cdots, \mathcal{A}_m\}, \mathbb{S} \leftarrow \emptyset$.

2. Calculate the relevance $\gamma_{\mathcal{A}_i}(\mathbb{D})$ of each gene $\mathcal{A}_i \in \mathbb{C}$.

3. Select gene $\mathcal{A}_i$ as the most relevant gene that has highest relevance $\gamma_{\mathcal{A}_i}(\mathbb{D})$. In effect, $\mathcal{A}_i \in \mathbb{S}$ and $\mathbb{C} = \mathbb{C} \setminus \mathcal{A}_i$.

4. Repeat the following two steps until the desired number of genes are selected.

5. Calculate the redundancy between selected genes of $\mathbb{S}$ and each of the remaining genes of $\mathcal{A}$.

6. From the remaining genes of $\mathbb{C}$, select gene $\mathcal{A}_j$ that maximizes

$$\gamma_{\mathcal{A}_j}(\mathbb{D}) - \frac{1}{|\mathbb{S}|} \sum_{\mathcal{A}_i \in \mathbb{S}} \lambda(\mathcal{A}_i, \mathcal{A}_j). \tag{3.8}$$

   As a result of that, $\mathcal{A}_j \in \mathbb{S}$ and $\mathbb{C} = \mathbb{C} \setminus \mathcal{A}_j$.

The redundancy measure doesn't take into account the supervised information of classlabels. Thus, the mRMR criterion for gene selection may not always be effective for identification of disease genes.

3.1.5 MRMS: Maximum Relevance Maximum Significance Criterion

The current method uses maximum relevance-maximum significance criterion to select the relevant and significant genes from high dimensional microarray gene expression data sets. The gene set $\mathbb{S}$ is selected using Mutual Information based MRMS method, where mutual information is used to compute the relevance between genes and class of samples, and also the significance of a gene w.r.t. other for a particular class label.

Let $\mathbb{C} = \{\mathcal{A}_1, \cdots, \mathcal{A}_i, \cdots, \mathcal{A}_j, \cdots, \mathcal{A}_m\}$ be the set of m genes of a given microarray gene expression data set and $\mathbb{S}$ is the set of selected genes. Define $\gamma_{\mathcal{A}_i}(\mathbb{D})$ as the relevance of the gene $\mathcal{A}_i$ with respect to the class labels $\mathbb{D}$ while $\sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j)$ as the significance of the gene $\mathcal{A}_j$ with respect to the set $\{\mathcal{A}_i, \mathcal{A}_j\}$. Mutual Information can be used to calculate both relevance and significance of each gene.

Definition The significance of a gene $\mathcal{A}_j$ with respect to another gene $\mathcal{A}_i$ can be defined as follows

$$\sigma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathbb{D},\mathcal{A}_j) = \gamma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathbb{D}) - \gamma_{\mathcal{A}_i}(\mathbb{D}). \tag{3.9}$$

That is, the significance of a gene $\mathcal{A}_j$ is the change in dependency when the gene $\mathcal{A}_j$ is removed from the set $\{\mathcal{A}_i, \mathcal{A}_j\}$. The higher the change in dependency, the more significant the gene $\mathcal{A}_j$ is. If the significance is 0, then the gene $\mathcal{A}_j$ is dispensable.

Hence, the total relevance of all selected genes is

$$\mathcal{J}_{\text{relev}} = \sum_{\mathcal{A}_i \in \mathbb{S}} \gamma_{\mathcal{A}_i}(\mathbb{D}), \tag{3.10}$$

while the total significance among the selected genes is

$$\mathcal{J}_{\text{signf}} = \sum_{\mathcal{A}_i \neq \mathcal{A}_j \in \mathbb{S}} \sigma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathbb{D},\mathcal{A}_j). \tag{3.11}$$

Hence, the problem of selecting a set $\mathbb{S}$ of relevant and significant genes from the whole set $\mathbb{C}$ of m genes is equivalent to maximize both $\mathcal{J}_{\text{relev}}$ and $\mathcal{J}_{\text{signf}}$, that is, to maximize the objective function

$$\mathcal{J} = \mathcal{J}_{\text{relev}} + \beta \mathcal{J}_{\text{signf}}, \tag{3.12}$$

where β is a weight parameter. To solve the above problem, following greedy algorithm is used in the current study:

1. Initialize $\mathbb{C} \leftarrow \{\mathcal{A}_1, \cdots, \mathcal{A}_i, \cdots, \mathcal{A}_j, \cdots, \mathcal{A}_m\}, \mathbb{S} \leftarrow \emptyset$.
2. Calculate the relevance $\gamma_{\mathcal{A}_i}(\mathbb{D})$ of each gene $\mathcal{A}_i \in \mathbb{C}$.
3. Select the gene $\mathcal{A}_i$ as the most relevant gene that has the highest relevance value $\gamma_{\mathcal{A}_i}(\mathbb{D})$. In effect, $\mathcal{A}_i \in \mathbb{S}$ and $\mathbb{C} = \mathbb{C} \setminus \mathcal{A}_i$.
4. Repeat the following two steps until the desired number of genes is selected.

5. Calculate the significance of each of the remaining genes of $\mathbb{C}$ with respect to the selected genes of $\mathbb{S}$ and remove it from $\mathbb{C}$ if it has zero significance value with respect to any one of the selected genes.

6. From the remaining genes of $\mathbb{C}$, select gene $\mathcal{A}_j$ that maximizes the following condition:

$$\gamma_{\mathcal{A}_j}(\mathbb{D}) + \frac{\beta}{|\mathbb{S}|} \sum_{\mathcal{A}_i \in \mathbb{S}} \sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j). \tag{3.13}$$

   As a result of that, $\mathcal{A}_j \in \mathbb{S}$ and $\mathbb{C} = \mathbb{C} \setminus \mathcal{A}_j$.

7. Stop.

3.2 Disease Gene Selection from PPIN Data

3.2.1 Disease Candidate Gene Prioritization using PPI Networks(Method proposed by Chen et al.):

Most of the disease candidate gene identification and prioritization methods depended on functional annotations. However the coverage of gene functional annotations act as a limiting factor. Thus, this method was proposed based entirely on the protein-protein interaction network analyses.

Based on the observation that biologi- cal networks share many properties with Web and social networks, is an attempt to extend the successful graph analysis-based algorithms from computer science research to tackle the disease gene prioritization problem. Literature -based and manually curated protein interactions were used to form the base network, and extended versions of the PageRank algorithm and HITS algorithm, as well as the K-Step Markov method, were applied to prior- itize disease candidate genes in a training-test schema. The prioritization approaches are based on the methods of White and Smyth , whose general framework, consisting of four successive problem formulations, each building on the next.

1. Relative Importance of a node t with respect to a root node r.

2. Rank of importance of a set of nodes T with respect to a root node r.

3. Rank of importance of a set of nodes T with respect to a set of root nodes R.

4. Given G, rank all nodes.

The importance of a gene to the set of root genes is just the average sum of the importance of it to each individual root gene. Thus, the key solution to the above defined problems is to find the importance of the nodes with respect to a root node. Three different algorithms are used for this purpose, namely (a) PageRank with Priors, (b) HITS with Priors, and (c) K-step Markov. Even though network-based methods are generally not as effective as integrated functional annotation-based methods for disease candidate gene prioritization, in a one-to-one comparison, PPIN-based candidate gene prioritization performs better than all other gene features or annotations [12].

3.2.2 The power of protein interaction networks for associating genes with diseases(Method proposed by Navlakha et al.):

The proposed method understands the fact that the association between genetic diseases and their causal genes is an important problem concerning human health. With the recent influx of high-throughput data describing interactions between gene products, the above described associations can be inferred. The authors assessed the utility of physical protein interactions for determining genedisease associations by examining the performance of seven recently developed computational methods (plus several of their variants). The proposed method found that random-walk approaches individually outperform clustering and neighborhood approaches, although most methods make predictions not made by any other method. It combines these methods into a consensus method yields Pareto optimal performance. It also quantified how a diffuse topological distribution of disease-related proteins negatively affects prediction quality and are thus able to identify diseases

especially amenable to network-based predictions and others for which additional information sources are absolutely required [57].

3.2.3 Predicting disease genes using proteinprotein interactions(Method proposed by Oti et al.):

The method was proposed with the objective to investigate whether protein protein interactions can predict genes for genetically heterogeneous diseases. 72 940 proteinprotein interactions between 10 894 human proteins were used to search 432 loci for candidate disease genes representing 383 genetically heterogeneous hereditary diseases. For each disease, the protein interaction partners of its known causative genes were compared with the disease associated loci lacking identified causative genes. Interaction partners located within such loci were considered candidate disease gene predictions. Prediction accuracy was tested using a benchmark set of known disease genes.Almost 300 candidate disease gene predictions were made. Some of these have since been confirmed. On average, 10percent or more are expected to be genuine disease genes, representing a 10-fold enrichment compared with positional information only. Thus, it could be concluded that exploiting proteinprotein interactions can greatly increase the likelihood of finding positional candidate disease genes. When applied on a large scale they can lead to novel candidate gene predictions [61].

3.3 Disease Gene Selection from Microarray Data and PPIN

3.3.1 Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes(Method proposed by Wu et al.):

The method was proposed to investigate the networks in which the genes associated with complex diseases play a role. A new method named, Networked

Gene Prioritizer(NGP) was proposed to prioritize cancer-associated genes. The methods assumes that between compared samples, cancer-associated genes cause the differential expression of their interacting genes by Network Rewiring(NR) and/or Networked Differential Expression(ND). In the mentioned study, the authors preprocess the data by removing all ambigous probe sets and selecting only the ones having the most significant p-values. It preselects a set of hub genes due to some reasons, (1) It is assumed that genes with many interacting partners play important roles in cells, (2) NGP, requires that a candidate gene must have more than 15 interaction neighbors, and (3)the PPI Network is constructed using genes from the microarray and the PPI database used. The PPIs were weighted differently in the NR and ND models. Keeping the more than 15 interacting neighbors constraint in mind,subnets of candidate genes were constructed and these were further trimmed to get subnets of enriched Differentially Expressed (DE) genes. Using Z-score the statistical significance of the trimmed subnet was found and the candidate genes were prioritized according to the sum of z-scores they get in different subnet generation steps [81]. Applications on several breast cancer and lung cancer datasets demonstrated that NGP performs better than the existing methods. It provides stable top ranking genes between independent datasets. The top-ranked genes by NGP are enriched in the cancer-associated pathways.

3.3.2 Ranking Candidate Disease Genes from Gene Expression and Protein Interaction: A Katz-Centrality Based Approach(Method proposed by Zhao et al.):

In this study, the proposed method aims to integrate gene expression level, protein-protein interaction strength and known disease genes. Their are two simple biologically motivated assumptions, which can be stated as- a gene is a good disease-gene candidate if (1) it is differently expressed in cases and controls, and (2) it is close to other disease-gene candidates in its protein interaction network. With the above assumptions in mind the authors proposed a score inspired by Katz centrality. To improve of the performanceof

the method, partial information of known disease genes were also incorporated. This study provides a novel, effective and easy- implemented algorithm for the prioritization of candidate disease genes [87].

3.3.3 dmGWAS: dense module searching for genome-wide association studies in proteinprotein interaction networks(Method proposed by Jia et al.):

The proposed method here presents a dense module searching (DMS) method to identify candidate subnetworks or genes for complex diseases by integrating the association signal from GWAS datasets into the human proteinprotein interaction (PPI) network. The DMS method extensively searches for subnetworks enriched with low P-value genes in GWAS datasets. Compared with pathway-based approaches, this method introduces flexibility in defining a gene set and can effectively utilize local PPI information [34].

The DMS method comprises of the following steps:

1. Score subgraphs : The module is defined as a subgraph within the whole network with a locally maximum proportion of low-P-value genes. To quantitatively evaluate the density of low P-value genes held by a module. Based on this P-value assign a combined Z-score(Z_m) to each module.

2. Normalize the values of Z_m using a random set of genes to determine whether it was higher than expected.

3. Perform permutation-based normalization of Z_m.

4. Define a searching strategy, using every gene in the network as a seed.

3.3.4 Disease Gene Identification by Random Walk on Multigraphs merging Heterogenous Genomic and Phenotypic Data(Method proposed by Li and Li):

This current study aims to merge the seperate list of candidate genes while eliminating the noise and bias which inflates the uncertainity in the data. and then prioritize a set of candidate genes. This work proposes an integration method to merge various genomic networks into a multi- graph which is capable of connecting multiple edges between a pair of nodes. It then operates a random walk on the multigraph to find disease genes. The phenotype data isn't integrated to the multigraph gene network. Instead it is connected, as a subgraph to the multigraph gene network. This approach provides a data platform with strong noise tolerance to prioritize the disease genes. A new idea of random walk is then developed to work on multigraphs using a modified step to calculate the transition matrix. Our method is further enhanced to deal with heterogeneous data types by allowing cross-walk between phenotype and gene networks. Compared on benchmark datasets, our method is shown to be more accurate than the state- of-the-art methods in disease gene identification.

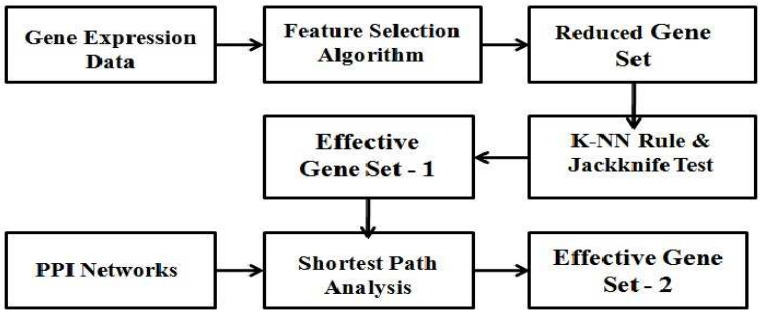

Figure 3.1: Schematic flow diagram of the insilico approach for identification of disease genes

3.3.5 Identification of Disease Genes Using Gene Expression and Protein-Protein Interaction Data :

This study validates the fact that gene expression data and protein-protein interaction data alone does not suffice for identification of novel disease genes. Thus, the method uses both gene expression and PPI data for disease gene identification. The integrated method involves the main operational steps as illustrated in Fig. 3.1. These steps can be outlined as follows:

1. Selection of Differentially Expressed Genes : The first step of the integrated method selects a set $\mathbb{S}$ of differentially expressed genes from the whole gene set $\mathbb{C}$ of the given microarray gene expression data set. The gene set $\mathbb{S}$ can be selected using the different gene selection methods, like MR, mRMR, MRMS, as discussed previously. In general, the microarray data may contain a number of irrelevant and insignificant genes. The presence of such genes may lead to a reduction in useful information. On the other hand, a gene set with high relevance, or high relevance and low redundancy or high relevance and high significance can enhance the predictive capability. The relevance, redundancy and significance parameters can all be computed using Mutual Information in the same way as described in section 3.1 under the MR, mRMR and MRMS headings.

2. Selection of Effective Gene Set I : In second step, a set of effective genes is identified as disease genes. The effective gene set I, as mentioned in Fig. 3.1 and denoted by $\mathbb{S}_{\mathrm{GE}}$, is a subset of $\mathbb{S}$, and defined as the gene set for which the prediction model or classifier attains its maximum classification accuracy. The K-nearest neighbor (K-NN) rule [21] is used here for evaluating the effectiveness of the reduced gene set for classification. The value of K, chosen for the current study, is 1, while the dissimilarity between two samples is calculated as follows:

$$D(x_i, x_j) = 1 - \frac{x_i \cdot x_j}{||x_i|| \cdot ||x_j||} \tag{3.14}$$

where x_i and x_j are two vectors representing two tissue samples, $x_i \cdot x_j$

is their dot product, and $||x_i||$ and $||x_j||$ are their moduli. The smaller the $D(x_i, x_j)$, the more similar the two samples are. To calculate the classification accuracy of the k-NN rule, the jackknife test [68] is used, although both independent data set test and subsampling test can also be used. However, jackknife estimators allow to correct for a bias and its statistical error.

3. Selection of Effective Gene Set II : Finally, the effective gene set II, denoted by $\mathbb{S}_{\text{GE+PPI}}$, is obtained from the PPI data based on the set $\mathbb{S}_{\text{GE}}$, the effective gene set I. The STRING (Search Tool for the Retrieval of Interacting Genes) [76] is an online database resource that provides both experimental as well as predicted interaction information with a confidence score. In general, the graph is a very useful tool for studying complex biological systems as it can provide intuitive insights and the overall structure property, as demonstrated by various studies on a series of important biological topics [2, 4, 13, 14, 15, 16, 17, 88, 89]. In the current work, after selecting the gene set $\mathbb{S}_{\text{GE}}$, a graph $G(V, E)$ is constructed with the PPI data from the STRING using the gene set $\mathbb{S}_{\text{GE}}$. In between each pair of genes, an edge is assigned in the graph. The weight of the edge E in graph G is derived from the confidence score according to the relation $\omega^G = 1000 \times (1 - \omega^0)$, where ω^G is the weight in graph G while ω^0 is the confidence score between two proteins concerned. Accordingly, a functional protein association network with edge weight is generated. In order to identify the shortest path from each of the selected differentially expressed genes of $\mathbb{S}_{\text{GE}}$ to remaining genes of the set $\mathbb{S}_{\text{GE}}$ in the graph, Dijkstra's algorithm [18] is used. Finally, the genes present in the shortest path are picked up and ranked according to their betweenness value. Let this set of genes be $\mathbb{S}_{\text{PPI}}$. The effective gene set II, that is, $\mathbb{S}_{\text{GE+PPI}}$, is the union of sets $\mathbb{S}_{\text{GE}}$ and $\mathbb{S}_{\text{PPI}}$, that is, $\mathbb{S}_{\text{GE+PPI}} = \mathbb{S}_{\text{GE}} \cup \mathbb{S}_{\text{PPI}}$ [53].

Integrating MR Criterion and PPIN Data : Using the maximum relevance criteria for gene selection a set $\mathbb{S}_{\text{GE}_{\text{mr}}}$ is obtained. This gene set obtained is then fed into STRING to generate a new gene set

$\mathbb{S}_{\mathrm{GE}_{\mathrm{mr}}+\mathrm{PPI}}$. This gene set generated incorporates along with itself the disadvantages of using the maximum relevance criterion.

Integrating mRMR Criterion and PPIN Data : In this method a similar approach is followed, making use of the mRMR criterion for generating the effective gene set I. The gene set can be named $\mathbb{S}_{\mathrm{GE}_{\mathrm{mrmr}}}$. Integrating the gene set obtained with the PPI Network data, we generate the effective gene set II, named $\mathbb{S}_{\mathrm{GE}_{\mathrm{mrmr}}+\mathrm{PPI}}$. This method is known to perform better than the previously discussed method. However, the mRMR criterion tends to overestimate the results and attains 100% accuracy with very less number of genes in the effective gene set I. Moreover, the supervised information of class labels still remains untouched in this method [43].

Integrating MRMS Criterion and PPIN Data : This approach uses the MRMS criterion for gene selection, to obtain a gene set $\mathbb{S}_{\mathrm{GE}_{\mathrm{mrms}}}$. Using this gene set with the PPI Network data, the new gene set generated, $\mathbb{S}_{\mathrm{GE}_{\mathrm{mrms}}+\mathrm{PPI}}$, is known to perform much better than the above discussed methods of integration. The fact that the significance of one gene with respect to another is being considered improves the predictive capability of the method [53].

Chapter 4

Proposed Disease Gene Identification Method

4.1 A New Protein-Protein Similarity Measure

In general, the genes, which are associated with the same disorder, tend to share common functional features. The protein products of these genes also have a tendency to interact with each other [6, 26]. Hence, an important characteristic of a disease gene is that its protein product is strongly linked to other disease-gene proteins. It has also been observed that proteins with short distances to each other in the network are more likely to involve in common biological functions [10, 40, 59], and that interactive neighbors are more likely to have identical biological function than non-interactive ones [37, 42]. This is because the query protein and its interactive proteins may form a protein complex to perform a particular function or be involved in a same pathway. Accordingly, a quantitative measure is required that can efficiently compute the similarity between two genes. In this paper, the information of PPI networks is used to calculate the functional similarity.

The PPI networks are commonly represented as graphs, with nodes corresponding to proteins and edges representing PPIs. The weight of the edge in graph depends on experimental as well as predicted interaction information.

Let $\mathcal{N}_i$ be the set of interactive neighbors or successor genes of a candidate gene $\mathcal{A}_i$ and $\omega_{ij} \in [0,1]$ is the weight value of the edge between gene $\mathcal{A}_j \in \mathcal{N}_i$ and candidate gene $\mathcal{A}_i$. The set of successors $\mathcal{N}_i$ of gene $\mathcal{A}_i$ and corresponding weight value ω_{ij} can be obtained from the information of PPI network. Let $\mathcal{N}_{ik}$ be the set of genes, which are successors of both genes $\mathcal{A}_i$ and $\mathcal{A}_k$, that is, $\mathcal{N}_{ik} = \mathcal{N}_i \cap \mathcal{N}_k$. Define $\tilde{\mathcal{N}}_i = \mathcal{N}_i \setminus \mathcal{N}_{ik}$ as the set of genes those are successors of gene $\mathcal{A}_i$ but not of gene $\mathcal{A}_k$. The functional similarity between two genes $\mathcal{A}_i$ and $\mathcal{A}_k$, having sets of successor genes $\mathcal{N}_i$ and $\mathcal{N}_k$, respectively, is as follows:

$$\mathcal{S}(\mathcal{A}_i, \mathcal{A}_k) = \frac{\sum\limits_{\mathcal{A}_j \in \mathcal{N}_{ik}} \min\{\omega_{ij}, \omega_{kj}\}}{\sum\limits_{\mathcal{A}_j \in \tilde{\mathcal{N}}_i} \omega_{ij} + \sum\limits_{\mathcal{A}_j \in \mathcal{N}_{ik}} \max\{\omega_{ij}, \omega_{kj}\} + \sum\limits_{\mathcal{A}_j \in \tilde{\mathcal{N}}_k} \omega_{kj}}. \tag{4.1}$$

Hence, if the interactive neighbors and the corresponding edge weights of two genes are same, then the functional similarity between these two genes is high. On the other hand, two genes are functionally dissimilar if they have no common interactive neighbors.

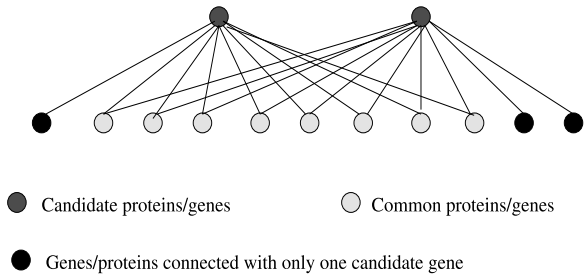

Figure 4.1: An example of protein-protein interaction network

The following properties can be stated about the measure:

1. $0 \leq \mathcal{S}(\mathcal{A}_i, \mathcal{A}_k) \leq 1$.
2. $\mathcal{S}(\mathcal{A}_i, \mathcal{A}_k) = 1$ if and only if two sets $\mathcal{N}_i$ and $\mathcal{N}_k$ contain exactly same set of successor genes, that is, $\mathcal{N}_{ik} = \mathcal{N}_i = \mathcal{N}_k$, and weight value $\omega_{ij} = \omega_{kj}, \forall \mathcal{A}_j \in \mathcal{N}_{ik}$.
3. $\mathcal{S}(\mathcal{A}_i, \mathcal{A}_k) = 0$ if and only if $\mathcal{N}_{ik} = \emptyset$.

4. $\mathcal{S}(\mathcal{A}_i, \mathcal{A}_k) = \mathcal{S}(\mathcal{A}_k, \mathcal{A}_i)$ (symmetric).

In this regard, it should be noted that if the weight value $\omega_{ij} \in \{0, 1\}$, then the proposed similarity measure reduces to

$$\mathcal{S}(\mathcal{A}_i, \mathcal{A}_k) = \frac{|\mathcal{N}_i \cap \mathcal{N}_k|}{|\mathcal{N}_i \cup \mathcal{N}_k|} \tag{4.2}$$

which is Jaccard index $J(\mathcal{A}_i, \mathcal{A}_k)$.

4.2 Proposed Disease Gene Selection Algorithm

Recent advancement and wide use of high-throughput biotechnologies have been producing huge amount of gene expression profiles data, which have been widely used in different studies to understand the function of disease genes. If a set of genes shows a consistent pattern of different expression levels in sick subjects and a control group, then that gene set is likely a strong candidate of playing a pathogenic role. The difference of gene expression levels in particular cell types can be studied to get an idea about the propensity of a disease. On the other hand, the genes associated with the same disease tend to share common functional features. Also, the protein products of disease genes have a tendency to interact with other disease-gene proteins.

In this regard, the paper presents a new gene selection algorithm, integrating judiciously the gene expression and PPI data, to identify pleiotropic genes involved in the physiological cellular processes of the disease. The proposed method assumes that the protein products of disease genes tend to be close to differentially expressed genes in the protein interaction network. Hence, the proposed gene selection algorithm selects a set $\mathbb{S}$ of disease genes from the whole gene set $\mathbb{C}$ of the given microarray gene expression data set by maximizing both relevance and functional similarity of genes present in $\mathbb{S}$. Let $\mathbb{C} = \{\mathcal{A}_1, \cdots, \mathcal{A}_i, \cdots, \mathcal{A}_j, \cdots, \mathcal{A}_m\}$ be the set of m genes of a given microarray gene expression data set and $\mathbb{S}$ is the set of selected genes. Define

$\gamma_{\mathcal{A}_i}(\mathbb{D})$ as the relevance of the gene $\mathcal{A}_i$ with respect to the class labels $\mathbb{D}$ while $\mathcal{S}(\mathcal{A}_i, \mathcal{A}_j)$ as the functional similarity between two genes $\mathcal{A}_i$ and $\mathcal{A}_j$. Hence, the total relevance of all selected genes is

$$\mathcal{J}_{\text{relevance}} = \sum_{\mathcal{A}_i \in \mathbb{S}} \gamma_{\mathcal{A}_i}(\mathbb{D}), \tag{4.3}$$

while the total functional similarity among the selected genes is

$$\mathcal{J}_{\text{similarity}} = \sum_{\mathcal{A}_i \neq \mathcal{A}_j \in \mathbb{S}} \mathcal{S}(\mathcal{A}_i, \mathcal{A}_j). \tag{4.4}$$

Hence, the problem of selecting a set $\mathbb{S}$ of relevant and functionally similar genes from the whole set $\mathbb{C}$ of m genes is equivalent to maximizing both $\mathcal{J}_{\text{relevance}}$ and $\mathcal{J}_{\text{similarity}}$, that is, to maximize the objective function

$$\mathcal{J} = a\mathcal{J}_{\text{relevance}} + (1 - a)\mathcal{J}_{\text{similarity}}, \tag{4.5}$$

where a is a weight parameter. To solve the above problem, following greedy algorithm is used in the current study:

1. Initialize $\mathbb{C} \leftarrow \{\mathcal{A}_1, \cdots, \mathcal{A}_i, \cdots, \mathcal{A}_j, \cdots, \mathcal{A}_m\}, \mathbb{S} \leftarrow \emptyset$.

2. Calculate the relevance $\gamma_{\mathcal{A}_i}(\mathbb{D})$ of each gene $\mathcal{A}_i \in \mathbb{C}$.

3. Select the gene $\mathcal{A}_i$ as the most relevant gene that has the highest relevance value $\gamma_{\mathcal{A}_i}(\mathbb{D})$. In effect, $\mathcal{A}_i \in \mathbb{S}$ and $\mathbb{C} = \mathbb{C} \setminus \mathcal{A}_i$.

4. Repeat the following two steps until the desired number of genes is selected.

5. Calculate the functional similarity between each of the remaining genes of $\mathbb{C}$ with respect to the selected genes of $\mathbb{S}$ and remove it from $\mathbb{C}$ if it has zero functional similarity value with respect to any one of the selected genes.

6. From the remaining genes of $\mathbb{C}$, select gene $\mathcal{A}_j$ that maximizes the

following condition:

$$a\gamma_{\mathcal{A}_j}(\mathbb{D}) + \frac{(1-a)}{|\mathbb{S}|}\sum_{\mathcal{A}_i \in \mathbb{S}} \mathcal{S}(\mathcal{A}_i, \mathcal{A}_j). \tag{4.6}$$

As a result of that, $\mathcal{A}_j \in \mathbb{S}$ and $\mathbb{C} = \mathbb{C} \setminus \mathcal{A}_j$.

7. Stop.

The mutual information [65] can be used to calculate the relevance of a gene with respect to class labels, while the proposed similarity measure, based on the information of PPI data, can be used for computing functional similarity between two genes. However, in microarray gene expression data sets, the class labels of samples are represented by discrete symbols, while the expression values of genes are continuous. Hence, to measure the gene-class relevance of a gene with respect to class labels using mutual information, the continuous expression values of a gene are divided into several discrete partitions. The *a prior* (marginal) probabilities and their joint probabilities are then calculated to compute the gene-class relevance using the definitions for discrete cases. In this paper, the discretization method reported in [20, 65] is employed to discretize the continuous gene expression values. The expression values of a gene are discretized using mean μ and standard deviation σ computed over n expression values of that gene: any value larger than $(\mu + \sigma/2)$ is transformed to state 1; any value between $(\mu - \sigma/2)$ and $(\mu + \sigma/2)$ is transformed to state 0; any value smaller than $(\mu - \sigma/2)$ is transformed to state -1. These three states correspond to the over-expression, baseline, and under-expression of genes. On the other hand, the STRING (Search Tool for the Retrieval of Interacting Genes) is an online database resource that provides both experimental as well as predicted PPI information, along with a confidence score. In the current work, STRING is used for computing functional similarity between two genes considering confidence score as the weight value.

4.3 Complexity Analysis

The Mutual Information based proposed gene selection algorithm has low computational complexity compared to the number of genes in the original microarray gene expression data set. Its computational complexity may be established as follows:

1. The computation of the relevance of m genes is carried out in step 2 of the proposed algorithm, which has a time complexity of $\mathcal{O}(m)$.

2. The selection of most relevant gene from the set of $\Updownarrow$ genes, which is carried out in step 3, has also a complexity $\mathcal{O}(m)$.

3. Since there is only one loop in step 4 of the proposed gene selection method, which needs to be executed $(d-1)$ times, where d is the desired number of genes to be selected.

   (a) The computation of functional similarity of a candidate gene with respect to a gene in the already selected set of genes takes $\mathcal{O}(n_0^2)$ time, where n_0 is the average number of neighbors to a protein in the protein-protein interaction network. If $\tilde{m}$ represents the cardinality of the already-selected gene set, the total complexity to compute functional similarity of $(m-\tilde{m})$ candidate genes, which is carried out in step 5, is $\mathcal{O}((m-\tilde{m})n_0^2)$

   (b) The selection of a gene from $(m-\tilde{m})$ candidate genes by maximizing both relevance functional similarity, which is carried out in step 6, has also a complexity $\mathcal{O}(m-\tilde{m})$

Hence, the total complexity to execute the loop $(d-1)$ times is $\mathcal{O}((d-1)(((m-\tilde{m})n_0^2)+(m-\tilde{m}))) = O(d(m-\tilde{m})n_0^2)$.

In effect, the selection of a set of d relevant and functionally similar genes from the whole set of m genes using the proposed method has an overall computational complexity of $\mathcal{O}(m) + O(m) + O(d(m-\tilde{m})n_0^2) = O(mn_0^2)$ since $d, \tilde{m} << m$.

Chapter 5

Experimental Results and Discussion

This section presents the performance of the proposed maximum relevance-maximum functional similarity (MRMFS) criterion based proposed gene selection algorithm, along with a comparison with other related methods. The algorithms compared are t-test, MR (maximum relevance), mRMR [20], MRMS [52], MR+PPIN [65], mRMR+PPIN [43], and MRMS+PPIN [65]. The mutual information is used to compute the relevance, redundancy, and significance of the genes. The value of a in (4.6) is set to 0.5.

5.1 Description of data Sets

5.1.1 Gene Expression Data

In this study, the gene expression data from the colorectal cancer study of Hinoue et al. [29] is used. The gene expression profiling of 26 colorectal tumors and matched histologically normal adjacent colonic tissue samples were retrieve from the NCBI Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) with the accession number of GSE25070. The number of genes and samples in this data set are 24526 and 52, respectively. The data set is pre-processed by standardizing each sample to zero mean and unit variance [43].

The performance of different methods is compared with respect to the

degree of overlapping with three gene lists, namely, LIST-1, LIST-2, and LIST-3. The LIST-1 contains 742 cancer related genes, which are collected from the Cancer Gene Census of the Sanger Centre, Atlas of Genetics and Cytogenetic in Oncology [31], and Human Protein Reference Database [38]. On the other hand, both LIST-2 and LIST-3 consist of colorectal cancer related genes. While the LIST-2 is retrieved from the study of Sabatas-Bellver et al. [71], the LIST-3 is prepared from the work of Nagaraj and Reverter [56]. While LIST-2 contains 438 colorectal cancer genes, LIST-3 consists of 134 colorectal cancer genes.

5.1.2 Protein-Protein Interaction Network Data Used

The initial weighted PPI network was retrieved from STRING which is a large database of known and predicted protein interactions. Proteins in the interaction network were represented with nodes, while the interaction between any two proteins therein was represented with an edge. These interactions contain direct (physical) and indirect (functional) interactions, derived from numerous sources such as experimental repositories, computational prediction methods. In the network, each edge is marked with a score to quantify the interaction confidence, i.e., the likelihood that an interaction may occur [43].

5.2 Comparative Performance Analysis Between different Gene Selection Methods

5.2.1 Degree of Overlapping with Known Disease Genes

This section presents the comparative performance analysis of different gene selection algorithms with respect to the degree of overlapping with the three gene lists. The algorithms compared are t-test, MR, mRMR [20], MRMS [52], and the proposed MRMFS. Results are reported for first twenty genes selected by different algorithms.

t-Test		MR		mRMR		MRMS		Proposed	
Gene	Y/N	Gene	Y/N	Gene	Y/N	Gene	Y/N	Gene	Y/N
GUCA2B	y	GUCA2B	y	GUCA2B	y	GUCA2B	y	GUCA2B	y
ADH1B	y	BEST2	n	PI16	n	BCHE	y	GUCA2A	y
SCARA5	y	TMIGD1	n	CDH3	y	CLDN8	y	BEST2	n
ESM1	n	CLDN8	y	SPIB	y	PI16	n	CLCA4	y
TSPAN7	n	PI16	n	BEST2	n	BEST2	n	SCNN1B	y
CA7	y	SCNN1B	y	HMGCLL1	n	TMIGD1	n	NR3C2	y
LGI1	n	CLCA4	y	CILP	n	CILP	n	CA4	y
CEMIP	n	ADH1B	y	NR3C2	y	CLCA4	y	CA1	y
GLTP	n	CA1	y	ADH1B	y	ADH1B	y	ELANE	n
CLDN1	y	CA4	y	BOP1	n	SCNN1B	y	AQP8	y
TMIGD1	n	SCARA5	y	ECI2	n	ECI2	n	GCG	y
ACKR2	n	GNG7	n	CXCL8	n	CA1	y	PLCD1	n
NR3C2	y	NR3C2	y	CLCA4	y	CXCL8	n	CFD	n
PLAC9	y	ECI2	n	TEP1	n	TMEM37	n	C7	y
PCOLCE2	n	CXCL8	n	LRP8	n	GNG7	n	BGN	y
MMP7	y	CILP	n	GCG	y	CA4	y	CDK4	y
CLEC3B	y	TMEM37	n	WISP2	n	AFF3	y	PRPH	n
BEST4	n	CLEC3B	y	TMIGD1	n	NR3C2	y	TGFBI	y
AQP8	y	ELANE	n	CFD	n	SCARA5	y	KLF4	n
RUNDC3B	n	HEPACAM2	n	C16ORF62	n	WISP2	n	MMP3	y

Table 5.1: Twenty Top-Ranked Genes and Overlapping With Known Disease Genes

Table 5.2.1 presents the lists of genes selected by different gene selection algorithms, along with their degree of overlapping with any one of the three cancer gene lists. From the results reported in Table 5.2.1, it can be seen that the proposed method provides better results than that of other methods with respect to degree of overlapping with known gene lists. Out of 20 selected genes, 14 genes selected by the proposed algorithm overlap with known disease genes, while t-test, MR, mRMR, and MRMS algorithms can identify 10, 10, 7, and 11 disease genes.

5.3 Comparative Performance Analysis Between Different Integrated Methods of Disease Gene Identification

The performance of the proposed algorithm is compared with two algorithms, namely, MR+PPIN [65] and mRMR+PPIN [43], which combine gene expression and PPIN data for selection of disease genes. The results are reported in Table 5.2 considering 41 genes as both MR+PPIN and mRMR+PPIN methods consider 41 genes for their analysis. Table 5.2 also presents the statistical significance test of the gene sets selected by the MR+PPIN, mRMR+PPIN, and proposed methods with respect to the genes of LIST-1, LIST-2, and LIST-3. Using the Fisher's exact test, statistical analysis of the overlapped genes is performed.

Table 5.2: Degree of Overlapping and Fisher's Exact Test

Methods/ Algorithms	LIST-1		LIST-2		LIST-3		LIST 2-3
	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value	
MR+PPIN	9	2.84E-05	7	2.10E-05	5	5.01E-06	10
mRMR+PPIN	8	1.91E-04	4	1.06E-02	3	2.02E-03	5
Proposed	5	2.33E-02	16	2.20E-16	8	1.29E-10	19

Out of total 41 genes selected by the proposed method, 16 and 8 genes are related to colorectal cancer with respect to the LIST-2 and LIST-3, respectively, while only 7 and 5 genes obtained using MR+PPIN are colon cancer related genes. On the other hand, only 4 and 3 genes selected using mRMR+PPIN are related to colon cancer with respect to two lists. Hence, the Fisher's exact test for the proposed method generates lower p-values for both LIST-2 and LIST-3, which are significantly better than the p-values obtained by other two methods. However, the degree of overlapping by the proposed algorithm with cancer related genes of LIST-1 is lower than that by existing methods. The last column of Table 5.2 depicts the degree of overlapping with respect to the two colorectal cancer gene lists. While the proposed

method can identify 19 colorectal cancer related genes, only 10 and 5 disease genes are identified by the MR+PPIN and mRMR+PPIN methods.

The performance of the proposed algorithm is now compared with the MRMS+PPIN algorithm [53],which combine gene expression and PPIN data for selection of disease genes. The results are reported in Table 5.3 considering 97 genes as the MRMS+PPIN methods consider 97 genes for their analysis. Table 5.3 also presents the statistical significance test of the gene sets selected by the MRMS+PPIN, and proposed method with respect to the genes of LIST-1, LIST-2, LIST-3 and LIST-2, LIST-3 taken together. Using the Fisher's exact test, statistical analysis of the overlapped genes is performed.

Out of total 97 genes selected by the proposed method, 28 and 11 genes are related to colorectal cancer with respect to the LIST-2 and LIST-3, respectively, while only 15 and 9 genes obtained using MRMS+PPIN are colon cancer related genes. Hence, the Fisher's exact test for the proposed method generates lower p-values for both LIST-2 and LIST-3, which are significantly better than the p-values obtained by other two methods. However, the degree of overlapping by the proposed algorithm with cancer related genes of LIST-1 is lower than that of MRMS+PPIN algorithm.

Table 5.3: Degree of Overlapping and Fisher's Exact Test

Methods/ Algorithms	LIST-1		LIST-2		LIST-3		LIST 2-3
	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value	
MRMS+PPIN	22	2.792E-11	15	1.758E-09	9	8.338E-09	19
Proposed	14	3.169E-05	28	2.20E-16	11	1.895E-11	31

The last column of Table 5.3 again depicts the degree of overlapping with respect to the two colorectal cancer gene lists. While the proposed method can identify 31 colorectal cancer related genes, only 19 disease genes are identified by the MRMS+PPIN method.

5.3.1 Graph of Resultant PPI Networks

The PPI network is generated for each gene selected by three gene selection algorithms, namely, MR, mRMR, MRMS and the proposed method. These networks are generated using the STRING database. The level of interaction between the selected set $\mathbb{S}_{\text{GE}}$ of genes and the proteins of the STRING database is measured by their confidence score. For the MRMS method, among the 20 genes of $\mathbb{S}_{\text{GE}}$, one gene, namely, **TMIGD**, does not have any interaction with any other genes. The shortest path analysis is conducted on this merged PPI network.

Fig. 5.1 and 5.2 shows the PPI network for 8 and 6 genes obtained by the MR and mRMR method respectively, along with their confidence scores. Fig. 5.3 shows the PPI network for 20 genes obtained by the MRMS method, along with their confidence scores. The nodes marked yellow represent the genes of $\mathbb{S}_{\text{GE}}$ set identified by the MRMS method, while other genes of $\mathbb{S}_{\text{PPI}}$ are existing in the shortest paths. The values on the edges represent the edge weights to quantify the interaction confidence. The smaller value indicates the stronger interaction between the two nodes.

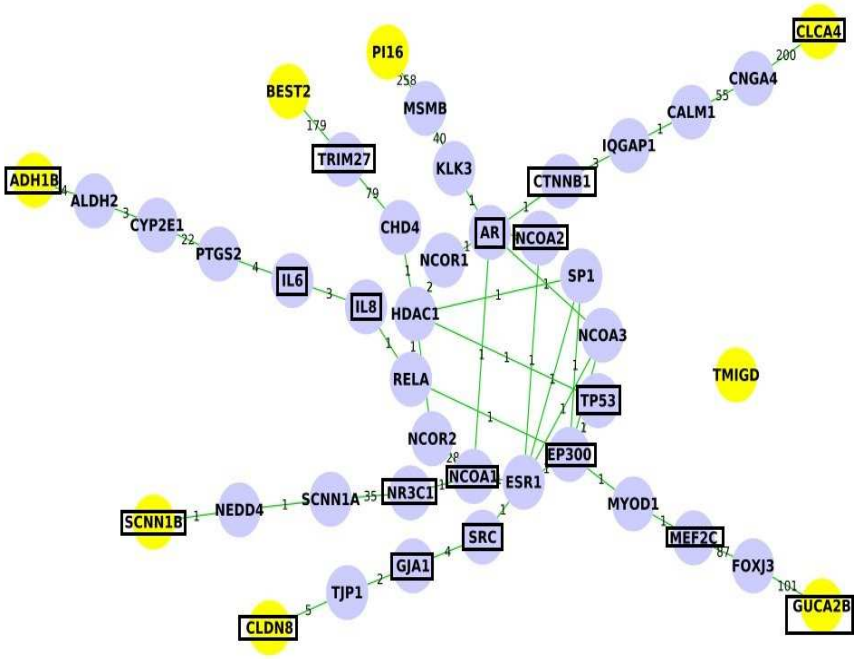

Figure 5.1: PPI network for 8 genes obtained by the MR method, along with their confidence scores

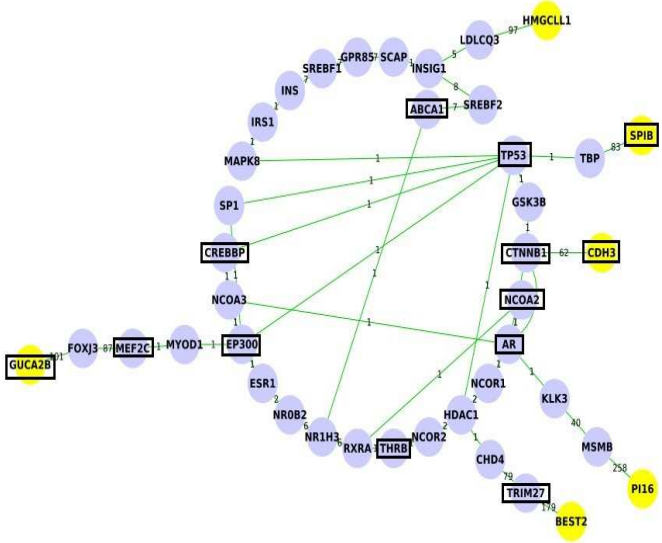

Figure 5.2: PPI network for 6 genes obtained by the mRMR method, along with their confidence scores

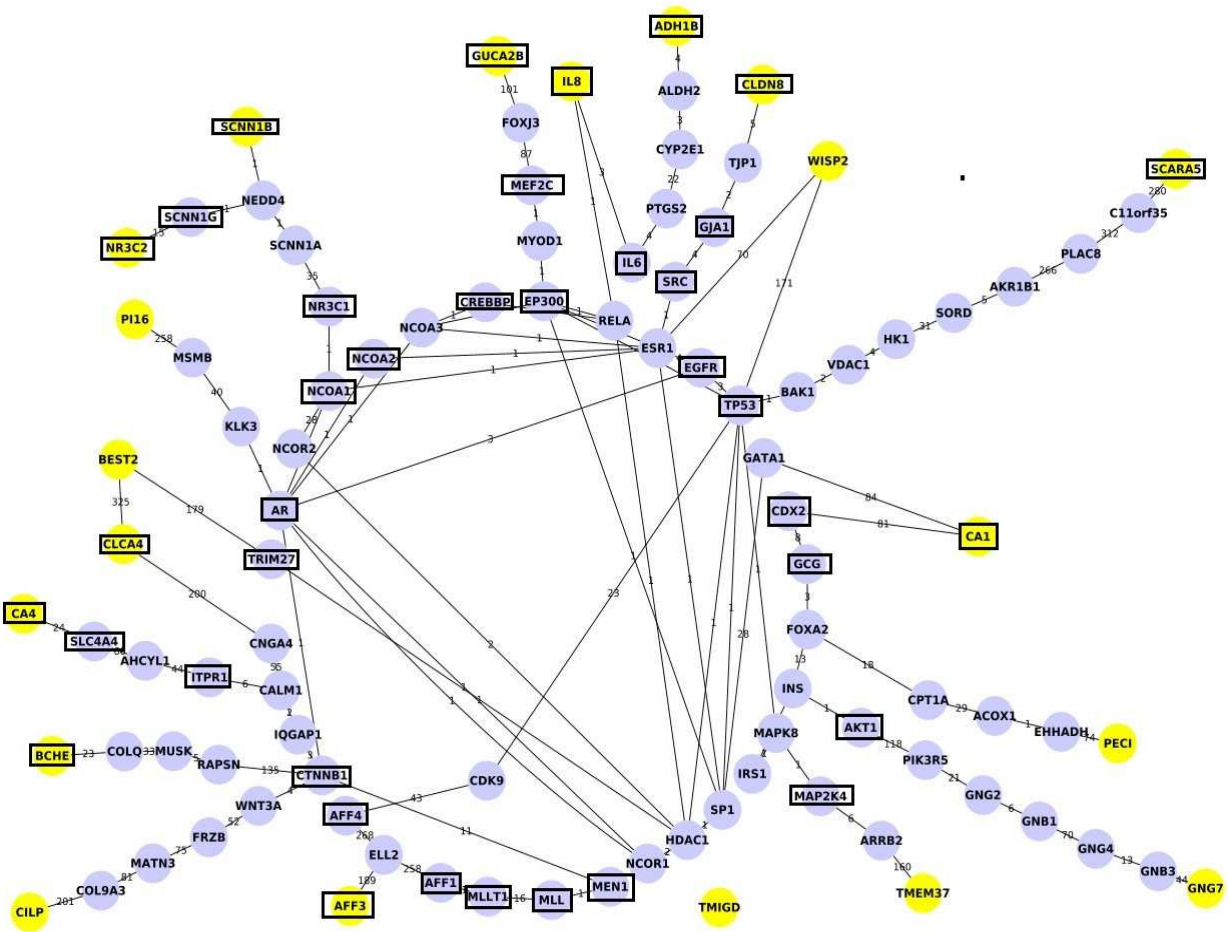

Figure 5.3: PPI network for 20 genes obtained by the MRMS method, along with their confidence scores

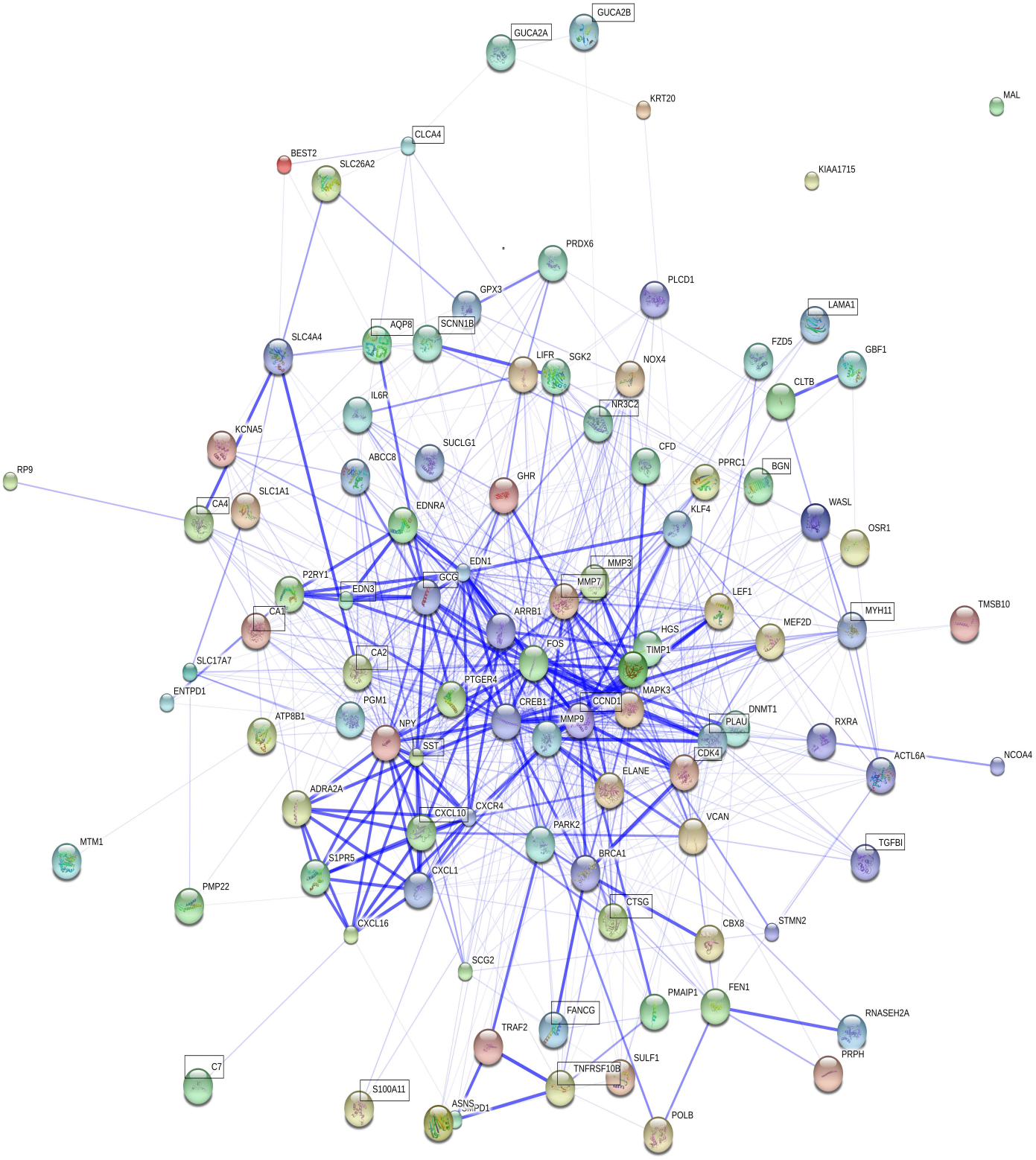

Figure 5.4: PPI network for 100 genes obtained by the MRMFS method, along with their confidence scores

5.3.2 KEGG Pathway Analysis

The hundred genes selected by the proposed method are further analyzed using the functional annotation tool of David. The enriched p-value was corrected to control family-wide false discovery rate under certain threshold (for example,<0.05) with Benjamin multiple testing correction method. Table 5.4 represents the KEGG pathway enrichment analysis of the gene set obtained by the proposed algorithm. From the table, it is seen that most of the networks are associated with cancer. Various processes, those are associated with colon cancer like p53 signaling pathway and colorectal cancer, are also observed in the result. Moreover, the gene set is found to be highly associated with colorectal cancer disease according to the OMIM disease database as analyzed by the functional annotation tool of David.

Table 5.4: KEGG Enrichment Analysis

KEGG ID	Term	Count	%	P-Value	Benjamin
05216	Thyroid cancer	5	0.42955	3.33E-04	3.37E-02
00910	Nitrogen metabolism	4	0.34364	2.33E-03	1.14E-01
05200	Pathways in cancer	11	0.94502	4.53E-03	1.44E-01
05219	Bladder cancer	4	0.34364	1.29E-02	2.85E-01
05222	Small cell lung cancer	5	0.42955	1.67E-02	2.94E-01
05210	Colorectal cancer	5	0.42955	1.67E-02	2.94E-01
04062	Chemokine signaling pathway	7	0.60137	2.20E-02	3.17E-01
05223	Non-small cell lung cancer	4	0.34364	2.53E-02	3.14E-01
04916	Melanogenesis	5	0.42955	2.87E-02	3.12E-01
04060	Cytokine-cytokine receptor interaction	8	0.68729	3.32E-02	3.21E-01
04115	p53 signaling pathway	4	0.34364	4.56E-02	3.81E-01

Chapter 6

Conclusion and Future Directions

The main contribution of the paper is to present a new gene selection algorithm to identify disease genes. The proposed algorithm integrates judiciously the information of gene expression profiles and protein-protein interaction networks. It selects a set of genes from microarray data as disease genes by maximizing the relevance and functional similarity of the selected genes. A new similarity measure is introduced to compute the functional similarity between two genes. It is based on the information of protein-protein interaction networks. The performance of the proposed algorithm, along with a comparison with other related methods, is demonstrated on colorectal cancer data set. Extensive experimental study on colorectal cancer establishes the fact that the genes identified by the proposed method have more colorectal cancer genes than the genes identified by the existing gene selection algorithms. All these results indicate that the proposed method is quite promising and may become a useful tool for identifying disease genes.

The proposed approach has quite a lot of future directions to explore. The study presented makes use of the Jaccard Index to compute the functional similarity, in a similar manner various other similarity indexes stated in literature can be used. Their performances can be compared and an optimal similarity criteria can be formulated.

Bibliography

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. 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 Sciences, USA*, 96(12):6745–6750, 1999.

[2] I. W. Althaus, A. J. Gonzales, J. J. Chou, D. L. Romero, M. R. Deibel, K. C. Chou, F. J. Kezdy, L. Resnick, M. E. Busso, and A. G. So. The Quinoline U-78036 is a Potent Inhibitor of HIV-1 Reverse Transcriptase. *Journal of Biological Chemistry*, 268(20):14875–14880, 1993.

[3] David Altshuler, Mark J. Daly, and Eric S. Lander. Genetic Mapping in Human Disease. *Science*, 322(5903):881–888, 2008.

[4] J. Andraos. Kinetic Plasticity and the Determination of Product Ratios for Kinetic Schemes Leading to Multiple Products Without Rate Laws - New Methods Based on Directed Graphs. *Canadian Journal of Chemistry*, 86(4):342–357, 2008.

[5] P. Baldi and A. D. Long. A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t-test and Statistical Inferences of Gene Changes. *Bioinformatics*, 17(6):509–519, 2001.

[6] Fredrik Barrenas, Sreenivas Chavali, Petter Holme, Reza Mobini, and Mikael Benson. Network Properties of Complex Human Disease Genes Identified through Genome-Wide Association Studies. *PLoS ONE*, 4(11):e8090, 2009.

[7] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue Classification with Gene Expression Profiles. *Journal of Computational Biology*, 7(3/4):559–584, 2000.

[8] R. Blanco, P. Larranaga, I. Inza, and B. Sierra. Gene Selection for Cancer Classification Using Wrapper Approaches. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(8):1373–1390, 2004.

[9] T. Bø and I. Jonassen. New Feature Subset Selection Procedures for Classification of Expression Profiles. *Genome Biology*, 3(4), 2002.

[10] P. Bogdanov and A.K. Singh. Molecular Function Prediction Using Neighborhood Features. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(2):208–217, 2010.

[11] Yu-Dong Cai, Tao Huang, Kai-Yan Feng, Lele Hu, and Lu Xie. A Unified 35-Gene Signature for both Subtype Classification and Survival Prediction in Diffuse Large B-Cell Lymphomas. *PLoS ONE*, 5(9):e12726, 2010.

[12] Jing Chen, Bruce Aronow, and Anil Jegga. Disease Candidate Gene Identification and Prioritization Using Protein Interaction Networks. *BMC Bioinformatics*, 10(1):73, 2009.

[13] K. C. Chou. Applications of Graph Theory to Enzyme Kinetics and Protein Folding Kinetics: Steady and Non-Steady-State Systems. *Biophysical Chemistry*, 35(1):1–24, 1990.

[14] K. C. Chou. Graphic Rule for Non-Steady-State Enzyme Kinetics and Protein Folding Kinetics. *Journal of Mathematical Chemistry*, 12(1):97–108, 1993.

[15] K. C. Chou. Graphic Rule for Drug Metabolism Systems. *Current Drug Metabolism*, 11:369–378, 2010.

[16] K. C. Chou and S. Forsen. Graphical Rules for Enzyme-Catalysed Rate Laws. *Biochemical Journal*, 187:829–835, 1980.

[17] K. C. Chou, F. J. Kezdy, and F. Reusser. Kinetics of Processive Nucleic Acid Polymerases and Nucleases. *Analytical Biochemistry*, 221(2):217–230, 1994.

[18] E. W. Dijkstra. A Note on Two Problems in Connexion with Graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

[19] C. Ding and H. Peng. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. In *Proceedings of the International Conference on Computational Systems Bioinformatics*, pages 523–528, 2003.

[20] C. Ding and H. Peng. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *Journal of Bioinformatics and Computational Biology*, 3(2):185–205, 2005.

[21] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification and Scene Analysis*. John Wiley and Sons, New York, 1999.

[22] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, 97(457):77–87, 2002.

[23] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001.

[24] R. Fox and M. Dimmic. A Two-Sample Bayesian t-test for Microarray Data. *BMC Bioinformatics*, 7(1):126, 2006.

[25] O. Gevaert, F. D. Smet, D. Timmerman, Y. Moreau, and B. D. Moor. Predicting the Prognosis of Breast Cancer by Integrating Clinical and Microarray Data with Bayesian Networks. *Bioinformatics*, 22(14):e184–e190, 2006.

[26] Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Lszl Barabsi. The Human Disease Network. *Proceedings of the National Academy of Sciences, USA*, 104(21):8685–8690, 2007.

[27] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–537, 1999.

[28] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning*, 46(1-3):389–422, 2002.

[29] T. Hinoue, D. J. Weisenberger, C. P.E. Lange, H. Shen, H. M. Byun, D. Van Den Berg, S. Malik, F. Pan, H. Noushmehr, C. M. van Dijk, R. A. E. M. Tollenaar, and P. W. Laird. Genome-Scale Analysis of Aberrant DNA Methylation in Colorectal Cancer. *Genome Research*, 22(2):271–282, 2012.

[30] Tao Huang, Lei Chen, Yu-Dong Cai, and Kuo-Chen Chou. Classification and Analysis of Regulatory Pathways Using Graph Property, Biochemical and Physicochemical Property, and Functional Property. *PLoS ONE*, 6(9):e25297, 2011.

[31] J. L. Huret, P. Dessen, and A. Bernheim. Atlas of Genetics and Cytogenetics in Oncology and Haematology, Year 2003. *Nucleic Acids Research*, 31(1):272–274, 2003.

[32] I. Inza, P. Larranaga, R. Blanco, and A. J. Cerrolaza. Filter Versus Wrapper Gene Selection Approaches in DNA Microarray Domains. *Artificial Intelligence in Medicine*, 31(2):91–103, 2004.

[33] P. Jafari and F. Azuaje. An Assessment of Recently Published Gene Expression Data Analyses: Reporting Experimental Design and Statistical Factors. *BMC Medical Informatics and Decision Making*, 6(1):27, 2006.

[34] Peilin Jia, Siyuan Zheng, Jirong Long, Wei Zheng, and Zhongming Zhao. dmGWAS: Dense Module Searching for Genome-Wide Associ-

ation Studies in Protein-Protein Interaction Networks. *Bioinformatics*, 27(1):95–102, 2011.

[35] H. Jiang, Y. Deng, H. S. Chen, L. Tao, Q. Sha, J. Chen, C. J. Tsai, and S. Zhang. Joint Analysis of Two Microarray Gene-Expression Data Sets to Select Lung Adenocarcinoma Marker Genes. *BMC Bioinformatics*, 5(1):81, 2004.

[36] T. Jirapech-Umpai and S. Aitken. Feature Selection and Classification for Microarray Data Analysis: Evolutionary Methods for Identifying Predictive Genes. *BMC Bioinformatics*, 6(1):148, 2005.

[37] Ulas Karaoz, T. M. Murali, Stan Letovsky, Yu Zheng, Chunming Ding, Charles R. Cantor, and Simon Kasif. Whole-Genome Annotation by Using Evidence Integration in Functional-Linkage Networks. *Proceedings of the National Academy of Sciences, USA*, 101(9):2888–2893, 2004.

[38] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human Protein Reference Database-2009 Update. *Nucleic Acids Research*, 37(suppl 1):D767–D772, 2009.

[39] S. Kohler, S. Bauer, D. Horn, and P. N. Robinson. Walking the Interactome for Prioritization of Candidate Disease Genes. *The American Journal of Human Genetics*, 82(4):949 – 958, 2008.

[40] Y. A. I. Kourmpetis, A. D. J. van Dijk, M. C. A. M. Bink, R. C. H. J. van Ham, and C. J. F. ter Braak. Bayesian Markov Random Field Analysis for Protein Function Prediction Based on Network Data. *PLoS ONE*, 5(2):e9293, 2010.

[41] J. W. Lee, J. B. Lee, M. Park, and S. H. Song. An Extensive Comparison of Recent Classification Tools Applied to Microarray Data. *Computational Statistics and Data Analysis*, 48(4):869–885, 2005.

[42] Stanley Letovsky and Simon Kasif. Predicting Protein Function from Protein/Protein Interaction Data: A Probabilistic Approach. *Bioinformatics*, 19(suppl 1):i197–i204, 2003.

[43] Bi-Qing Li, Tao Huang, Lei Liu, Yu-Dong Cai, and Kuo-Chen Chou. Identification of Colorectal Cancer Related Genes with mRMR and Shortest Path in Protein-Protein Interaction Network. *PLoS ONE*, 7(4):e33393, 2012.

[44] L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen. Gene Selection for Sample Classification Based on Gene Expression Data: Study of Sensitivity to Choice of Parameters of the GA/KNN Method. *Bioinformatics*, 17(12):1131–1142, 2001.

[45] T. Li, C. Zhang, and M. Ogihara. A Comparative Study of Feature Selection and Multiclass Classification Methods for Tissue Classification Based on Gene Expression. *Bioinformatics*, 20(15):2429–2437, 2004.

[46] Yongjin Li and Jinyan Li. Disease Gene Identification by Random Walk on Multigraphs Merging Heterogeneous Genomic and Phenotype Data. *BMC Genomics*, 13(Suppl 7):S27, 2012.

[47] Q. Liu, A. Sung, Z. Chen, J. Liu, L. Chen, M. Qiao, Z. Wang, X. Huang, and Y. Deng. Gene Selection and Classification for Cancer Microarray Data Based on Machine Learning and Similarity Measures. *BMC Genomics*, 12(Suppl 5):S1, 2011.

[48] J. Lyons-Weiler, S. Patel, M. Becich, and T. Godfrey. Tests for Finding Complex Patterns of Differential Expression in Cancers: Towards Individualized Medicine. *BMC Bioinformatics*, 5(1):110, 2004.

[49] S. Ma and J. Huang. Regularized ROC Method for Disease Classification and Biomarker Selection with Microarray Data. *Bioinformatics*, 21(24):4356–4362, 2005.

[50] P. Maji. f-Information Measures for Efficient Selection of Discriminative Genes from Microarray Data. *IEEE Transactions on Biomedical Engineering*, 56(4):1063–1069, 2009.

[51] P. Maji and S. K. Pal. Fuzzy-Rough Sets for Information Measures and Selection of Relevant Genes from Microarray Data. *IEEE Transactions on System, Man, and Cybernetics, Part B: Cybernetics*, 40(3):741–752, 2010.

[52] P. Maji and S. Paul. Rough Set Based Maximum Relevance-Maximum Significance Criterion and Gene Selection from Microarray Data. *International Journal of Approximate Reasoning*, 52(3):408–426, 2011.

[53] P. Maji and S. Paul. *Scalable Pattern Recognition Algorithms: Applications in Computational Biology and Bioinformatics*, page 304. Springer-Verlag, London, April 2014.

[54] H. Mamitsuka. Selecting Features in Microarray Classification Using ROC Curves. *Pattern Recognition*, 39(12):2393–2404, 2006.

[55] S. Miyano, S. Imoto, and A. Sharma. A Top-r Feature Selection Algorithm for Microarray Gene Expression Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(3):754–764, 2012.

[56] S. Nagaraj and A. Reverter. A Boolean-Based Systems Biology Approach to Predict Novel Genes Associated with Cancer: Application to Colorectal Cancer. *BMC Systems Biology*, 5(1):35, 2011.

[57] Saket Navlakha and Carl Kingsford. The Power of Protein Interaction Networks for Associating Genes with Diseases. *Bioinformatics*, 26(8):1057–1063, 2010.

[58] M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K. W. Tsui. On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. *Journal of Computational Biology*, 8(1):37–52, 2001.

[59] Ka-Lok Ng, Jin-Shuei Ciou, and Chien-Hung Huang. Prediction of Protein Functions Based on Function-Function Correlation Relations. *Computers in Biology and Medicine*, 40(3):300–305, 2010.

[60] C. H. Ooi and P. Tan. Genetic Algorithms Applied to Multi-Class Prediction for the Analysis of Gene Expression Data. *Bioinformatics*, 19(1):37–44, 2003.

[61] M Oti, B Snel, M A Huynen, and H G Brunner. Predicting Disease Genes Using Protein-Protein Interactions. *Journal of Medical Genetics*, 43(8):691–698, 2006.

[62] W. Pan. On the Use of Permutation in and the Performance of a Class of Nonparametric Methods to Detect Differential Gene Expression. *Bioinformatics*, 19(11):1333–1340, 2003.

[63] H. Pang, S. L. George, K. Hui, and T. Tong. Gene Selection Using Iterative Feature Elimination Random Forests for Survival Outcomes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(5):1422–1431, 2012.

[64] P. J. Park, M. Pagano, and M. Bonetti. A Nonparametric Scoring Algorithm for Identifying Informative Genes from Microarray Data. In *Proceedings of Pacific Symposium on Biocomputing*, pages 52–63, 2001.

[65] S. Paul and P. Maji. Gene Expression and Protein-Protein Interaction Data for Identification of Colon Cancer Related Genes Using f-Information Measures. *Natural Computing*, doi:10.1007/s11047-015-9485-6, 2015.

[66] P. Pavlidis and P. Poirazi. Individualized Markers Optimize Class Prediction of Microarray Data. *BMC Bioinformatics*, 7(1):345, 2006.

[67] H. Peng, F. Long, and C. Ding. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[68] M. H. Quenouille. Approximate Tests of Correlation in Time-Series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):68–84, 1949.

[69] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, de R. M. Van, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown. Systematic Variation in Gene Expression Patterns in Human Cancer Cell Lines. *Nature Genetics*, 24(3):227–235, 2000.

[70] R. Ruiz, J. C. Riquelme, and J. S. A. Ruiz. Incremental Wrapper-Based Gene Selection from Microarray Data for Cancer Classification. *Pattern Recognition*, 39(12):2383–2392, 2006.

[71] J. Sabates-Bellver, L. G. Van der Flier, M. de Palo, E. Cattaneo, C. Maake, H. Rehrauer, E. Laczko, M. A. Kurowski, J. M. Bujnicki, M. Menigatti, J. Luz, T. V. Ranalli, V. Gomes, A. Pastorelli, R. Faggiani, M. Anti, J. Jiricny, H. Clevers, and G. Marra. Transcriptome Profile of Human Colorectal Adenomas. *Molecular Cancer Research*, 5(12):1263–1275, 2007.

[72] Y. Saeys, I. Inza, and P. Larraaga. A Review of Feature Selection Techniques in Bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[73] M. Shah, M. Marchand, and J. Corbeil. Feature Selection with Conjunctions of Decision Stumps and Learning from Microarray Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):174–186, 2012.

[74] R. L. Somorjai, B. Dolenko, and R. Baumgartner. Class Prediction and Discovery Using Gene Microarray and Proteomics Mass Spectroscopy

Data: Curses, Caveats, Cautions. *Bioinformatics*, 19(12):1484–1491, 2003.

[75] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A Comprehensive Evaluation of Multicategory Classification Methods for Microarray Gene Expression Cancer Diagnosis. *Bioinformatics*, 21(5):631–643, 2005.

[76] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering. The STRING Database in 2011: Functional Interaction Networks of Proteins, Globally Integrated and Scored. *Nucleic Acids Research*, 39(suppl 1):D561–D568, 2011.

[77] J. G. Thomas, J. M. Olson, S. J. Tapscott, and L. P. Zhao. An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles. *Genome Research*, 11(7):1227–1236, 2001.

[78] V. Tusher, R. Tibshirani, and G. Chu. Significance Analysis of Microarrays Applied to the Ionizing Radiation Response. *Proceedings of the National Academy of Sciences, USA*, 98:5116–5121, 2001.

[79] R. D. Uriarte and S. A. de Andres. Gene Selection and Classification of Microarray Data Using Random Forest. *BMC Bioinformatics*, 7(1):3, 2006.

[80] Y. Wang, I. V. Tetko, M. A. Hall, E. Frank, A. Facius, K. F. X. Mayer, and H. W. Mewes. Gene Selection from Microarray Data for Cancer Classification - A Machine Learning Approach. *Computational Biology and Chemistry*, 29(1):37–46, 2005.

[81] Chao Wu, Jun Zhu, and Xuegong Zhang. Integrating Gene Expression and Protein-Protein Interaction Network to Prioritize Cancer-Associated Genes. *BMC Bioinformatics*, 13(1):182, 2012.

[82] E. P. Xing, M. I. Jordan, and R. M. Karp. Feature Selection for High-Dimensional Genomic Microarray Data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 601–608, 2001.

[83] M. Xiong, X. Fang, and J. Zhao. Biomarker Identification by Feature Wrappers. *Genome Research*, 11(11):1878–1887, 2001.

[84] F. Yang and K. Z. Mao. Robust Feature Selection for Microarray Data Based on Multicriterion Fusion. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(4):1080–1092, 2011.

[85] E. J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C. H. Pui, W. E. Evans, C. Naeve, L. Wong, and J. R. Downing. Classification, Subtype Discovery, and Prediction of Outcome in Pediatric Acute Lymphoblastic Leukemia by Gene Expression Profiling. *Cancer Cell*, 1(2):133–143, 2002.

[86] K. Yeung and R. Bumgarner. Multiclass Classification of Microarray Data with Repeated Measurements: Application to Cancer. *Genome Biology*, 4(12):R83, 2003.

[87] Jing Zhao, Ting-Hong Yang, Yongxu Huang, and Petter Holme. Ranking Candidate Disease Genes from Gene Expression and Protein Interaction: A Katz-Centrality Based Approach. *PLoS ONE*, 6(9):e24306, 2011.

[88] G. P. Zhou. The Disposition of The LZCC Protein Residues in Wenxiang Diagram Provides New Insights into The Protein-Protein Interaction Mechanism. *Journal of Theoretical Biology*, 284(1):142–148, 2011.

[89] G. P. Zhou and M. H. Deng. An Extension of Chou's Graphic Rules for Deriving Enzyme Kinetic Equations to Systems Involving Parallel Reaction Pathways. *Biochemical Journal*, 222:169–176, 1984.