AN APPROACH TO IDENTIFY THE DISEASE-GENE ASSOCIATION THROUGH BIOLOGICAL NETWORKS TOPOLOGICAL FEATURES

BY

SADIQ ABDULMOHSIN AL-INSAIF

A Thesis Presented to the
DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

COMPUTER SCIENCE

May, 2014

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN- 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by **SADIQ AL-INSAIF** under the direction of his thesis advisor and approved by his thesis committee, has been presented and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN COMPUTER SCIENCE.**

Md. Rafiul Hassan 22/6/2014

Dr. Md Rafiul Hassan
(Advisor)

Dr. Emad Ramadan
(Co-Advisor)

Dr. Moataz Ahmed
(Member)

M Alshayeb

Dr. Mohammad Alshayeb
(Member)

Tarek Hew

Dr. Tarek El-Basuny
(Member)

22/14

Dr. Adel Ahmed
Department Chairman

Dr. Salam A. Zummo
Dean of Graduate Studies

26/6/14

Date

To my beloved family

ACKNOWLEDGMENTS

Praise be to Allah who guided me and gave me patient to complete my studies. Peace and blessings be upon His messenger Muhammad.

I would like to express my great appreciation to my advisors, Dr. Md Rafiul Hassan and Dr. Emad Ramadan for their continuous guidance to blossom this thesis work. I would like to thank my committee members Dr. Moataz Ahmed, Dr. Mohammad Alshayeb, and Dr. Tarek El-Basuny for showing interest in my work and their great insightful comments to improve my write-up and experimental set-up.

Lastly, I would like to thank my father and mother for their continuous support and patient to complete my graduate studies. I thank my brothers and sisters for their continuous concern about me and my success. I thank everyone who prayed for me and wish me success in my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

MLP:	**Multilayer Perceptron Neural Networks**
PCCs:	**Pearson correlation coefficients**
ACC:	**Accuracy**
ANFIS:	**Adaptive Neuro-Fuzzy Inference System**
ANN:	**Artificial Neural Networks**
BMD:	**bone mineral density**
CATAPULT:	**Combining dATa Across species using Positive-Unlabeled Learning Techniques**
CFS:	**Correlation based Feature Selection**
CV:	**cross-validation**
DNA	**desoxyribonucleicacid**
DBC:	**Density Based Clustering**
DBN:	**Dynamic Bayesian Network**
DTB:	**Decision Tree Bagger**
EM:	**Expectation Maximization**
F:	**F-measure**
FCM:	**fuzzy-C-means algorithm**
FI:	**Functional Interaction**
FN:	**false negatives**
FP:	**false positives**
GAD:	**genetic association database**
GBM:	**glioblastomamultiforme**

Gm:	**Geometric mean**
GMM:	**Gaussian Mixture Model**
GO:	**Gene Ontology**
HMM:	**Hidden Markov Model**
HPRD:	**Human Protein Reference Database**
LM:	**Levenberg–Marquardt**
MCL:	**Markov clustering**
MEFIT:	**Microarray Experiment Functional Integration Technology**
MIPS:	**Munich Information Center for Protein Sequence**
MIT:	**mutual information test**
MP:	**Mammalian Phenotype**
NBC:	**naïve Bayes classifier**
Net-Cox:	**network-based Cox regression**
OCC:	**one-class classification**
OMIM:	**Online Mendelian Inheritance in Man**
PPIBM:	**Protein-Protein Interaction based Bayesian Models**
PPIs:	**protein-protein interactions**
PPV:	**Positive predictive value**
PRINCE:	**PRIoritizatioN and Complex Elucidation**
RBF:	**Radial Basis Function**
RNA	**ribonucleicacid**
ROC:	**receiver operation characteristic**
RUSBoost:	**Random Under Sampling Boost**

SGD:	**Saccharomyces Genome Database**
SMOTE:	**Synthetic Minority Over-sampling TEchnique**
SN:	**Sensitivity**
SNP:	**Single-nucleotide polymorphism**
SOM:	**Kohonen’s self-organizing maps**
SP:	**Specificity**
SVM:	**Support Vector Machines**
SVM-RFE:	**support vector machine recursive feature elimination**
TMs:	**transcriptional modules**
TN:	**true negatives**
TP:	**true positives**

ABSTRACT

Full Name	:	Sadiq Abdulmohsin Makki Al-Insaif
Thesis Title	:	An Approach to Identify the Disease-Gene Association Through Biological Networks Topological Features
Major Field	:	Computer Science
Date of Degree	:	May/2014

Background: biologists generate massive datasets which are difficult to interpret and make use of for further diseases analysis. Hence, computational analysis techniques are required towards systems biology. There is an urgent need to investigate a system rather than evaluating individuals solely. This could be experimentally accomplished through studying individuals and then how they are interacting with others, such a concept is known as networks.

Approach: Complex biological networks constitute of thousands of nodes and thousands of interactions which could be represented as a graph. So, a graph theory could be utilized to rank nodes (i.e. genes) of the biological networks as individuals and as whole to retrieve the hidden properties and knowledge. Furthermore, this study applies different ranking algorithms (i.e. 14 ranking algorithms are used) and fed as features to two classification models which are decision tree bagger (DTB) and random under sampling boost (RUSBoost) to investigate the breast cancer genes. More importantly, to overcome the problem of the skewed datasets a synthetic minority oversampling technique (SMOTE) is adopted in order to transform an imbalanced dataset to a balanced one. Finally, we propose our modified co-cross validation technique to compute four-evaluation metrics (Accuracy, F-Measure, AUC, and Geometric Mean).

Results: We have extracted fourteen features from publicly co-expression network, protein interaction network, and functional interaction network. Then, we use classification models to investigate the phenotype-gene association in breast cancer genes. Prior classification, a SMOTE sampling technique is utilized in all biological networks. By using a Hill-climbing feature selection approach and Pearson correlation coefficients, a subset features have been identified as important features to predict phenotype-gene association. Specifically, 'within module z-score' is identified as the most influential feature to predict breast cancer genes.

Conclusion: We are able to identify sub-set of topological features that are significant to identify the breast cancer genes. We have adopted 'within module z-score' topological features to apply in-depth analysis which helped us to identify three genes that are subtle to breast cancer genes (i.e. ‘*ZNF22*’, ‘*CUL2*’, and ‘*GTF2F2*’).

Keywords: Biological network; Machine learning; Decision tree bagger; SMOTE; Breast cancer; Gene ranking.

ملخص الرسالة

الاسم الكامل: صادق عبد المحسن مكي آل انصيف

عنوان الرسالة: نهج لتحديد العلاقة مابين الأمراض و الجينات من خلال الخصائص المستخرجة من الشبكات البيولوجية

التخصص: علوم حاسب

تاريخ الدرجة العلمية: مايو/2014

المقدمة: علماء الأحياء يقوموا بإنتاج كم هائل من البيانات التي يصعب تفسيرها والإستفادة منها لتحليل الأمراض. وبالتالي، يلزم التوجه إلى إستخدام تقنيات التحليل الحسابي نحو أنظمة الشبكات البيولوجية لدراسة الأنظمة الحيوية بدلا من تقييم الأفراد فقط. وهذا يمكن أن يتحقق تجريبيا من خلال دراسة الأفراد ومن ثم كيف إن الأفراد تتفاعل مع الآخرين، ومن المعروف مثل هذا المفهوم يسمى بالشبكات.

النهج: الشبكات البيولوجية المعقدة تتشكل من الآلاف من العقد والآلاف من التفاعلات التي يمكن أن توضح بإستخدام الشبكات. لذلك، يمكن استخدام نظريات الرسم البياني لترتيب وتقييم العقد (أي الجينات) من الشبكات البيولوجية للحصول على الخصائص والمعرفة الخفية. علاوة على ذلك، هذا البحث يطبق 14 من خوارزميات الترتيب والتقييم للجينات الوراثية, حيث تُغذى إلى نموذجين من تعليم الآلة لدراسة الجينات المتعله بسرطان الثدي. الأهم من ذلك، للتغلب على مشكلة البيانات المنحرفة تم اعتماد اسلوب صناعة عينات اصطناعية لزيادة العينات الأقلية والتخلص من العينات المفرطة بشكل عشوائي من أجل تحويل مجموعة البيانات غير المتوازنة إلى بيانات متوازنه. أخيرا، تم اقترح اسلوب جديد في هذه الدراسة لتقسييم البيانات ومن ثم تعليم نمذجي الآلة, وتم التحقق من قوة وصحة النمذجين من خلال حساب أربعة مقاييس.

النتائج: لقد تم استخرج أربعة عشر خاصية على حدة كلاً من: شبكة التعبير، شبكة تفاعل البروتين، و شبكة التفاعل الوظيفي. ثم، تم استخدام هذه الخصائص من خلال نموذجين من تعليم الآله لكتشاف الأنماط المتعلقة بجينات سرطان الثدي. الجدير بالذكر، قبل تعليم نموذجي الآلة تم استخدام تقنية موازنة البيانات للتغلب على مشكلة البيانات المنحرفة وقد تم فعل ذلك في جميع الشبكات البيولوجية. باستخدام تقنيات مختلفة لتقييم الخصائص المستخرجة من الشبكات

الحيوية, تم التوصل إلى ان "الدرجة المعيارية" أكثر الخصائص أهمية لكتشاف الأنماط المتعلقة بجينات سرطان الثدي.

الخاتمة: في هذه الدراسة تم التعرف على مجموعة فرعية من الميزات الطوبوغرافية التي تعتبر هامة لتحديد جينات سرطان الثدي. وقد اعتمدنا بوجه الخصوص على خاصية 'الدرجة المعيارية' لتطبيق تحليل متعمق, والتي ساعدتنا على تحديد ثلاثة جينات قد تكون بشكل كبير متعلقة بمرض سرطان الثدي.

CHAPTER 1

INTRODUCTION

1.1. Overview

The ultimate goal of biologists is to understand the genetic bases of human diseases. Hence, the disease-gene association will prevent and provide proper therapeutic of diseases. For instance, Online Mendelian Inheritance in Man [1], the genetic association database [2], and other databases store phenotypic genes. These databases provide the "seed" genes of a certain disease, which might help to investigate further candidate genes. However, to understand and study new disease-genes association is a challenging task which requires laborious experiments. Moreover, it is time-consuming for a researcher to genotype and phenotype large population to determine which sequence features are most related to specific phenotype. Therefore, different bio-statistical techniques have been proposed to predict disease-gene association. For instance, linkage mapping is a statistical approach that has been successful to find the genomic regions that are related to a disease [3]. It is widely known technique that might examine genetically well-characterized populations, such as to identify quantitative trait loci (QTL) that contain causal mutations [4]. However, the identified regions typically reveal hundreds of candidate disease-causing genes [3] which means to identify the actual disease genes among the revealed candidate genes empirically it will still take a huge amount of time and effort.

Although quantitative genetics methods had promising results, these approaches have problematic limitations. For instance, QTL technique could suffer from sampling biases [4]. In addition, several linkages mapping QTL studies lack the statistical power to reduce the defined causal loci, which go through the entire chromosomes population that contain hundreds of candidate gens [5]. Therefore, it is necessary to provide an alternative and complementary approach that could substitute quantitative genetics techniques. Such new techniques must be independent and bias-free inherent in quantitative genetics. Furthermore, a combination of graph theory (i.e. biology network) and computational intelligence techniques could be applied to model co-expression, PPIs, and functional genomics networks of the activities of genes (e.g. expression, interaction, etc.). These techniques could identify candidate genes that probably will be involved in a phenotypic outcome, thus, that would lead to new discoveries in phenotype-gene associations.

The study of complex biological networks is a significant emerging field in a wide variety of disciplines, ranging from computer science, physics, and sociology, to biology. In the field of biomedicine and biology, there are a potential of research applications for network analysis. These include, drug development, personalized medicine, determining a gene's function, or phenotype-gene association studies. Furthermore, large and complex biological networks could be often represented using nodes and edges (i.e. nodes represent genes and edges represent interactions among genes). Such representation could be used to reveal significant hidden knowledge for further data analysis and pattern extraction. The graph theory plays an important role in the interactive analysis to build up the question of "which genes are most important relatively to a particular gene or set of genes". However, biological networks are too complex and cannot be analyzed easily. Thus, these biological networks could be understood through the utilization of graph theory analysis approaches. The graph theory is well-known approach that has been inspired from the study of the social network problems [6] and could be utilized to study biological networks as well. More interestingly, graph theory could be used to compute the topological properties and extract hidden knowledge of complex biological networks. In this study, fourteen different raking properties are computed (i.e. degree, closeness, betweeness, eigenvectors, within module z-score, k-step Markov, sub-graph, clustering coefficient, flow coefficient, Katz status, coreness, structural holes, proximity and Bary-center score). These topological properties have been extracted from co-expression, BioGrid PPIs, and functional protein interaction networks to rank the importance of individual genes relatively to others. Further, these aforementioned topological features are assessed in a hill-climbing feature selection to reveal the most significant sub-set topological features that provide best prediction rate.

In this study, the extracted topological features out of complex biological networks are utilized as features to build automated models using machine learning. The built models could be used to associate different kind of phenotypes (i.e. this study is investigating breast cancer genes). Prior the classification and due to the sever imbalance datasets; a SMOTE sampling technique is applied. Then, the balanced datasets are fed to two different classification models: Decision Tree Bagger (DTB) and RUSBoost. The two classifiers are implemented to predict breast cancer and non-breast cancer genes.

1.2. Motivation

Prior to the development of high-throughput technologies, it was difficult to diagnose diseases. The diagnostic process was based heavily on the identification of symptoms. Obviously, these symptoms are common among different types of diseases. Interestingly, the completion of human genome project in April, 2003 (http://www.genome.gov/11006943) opened the horizon to the informatics era. The genomic information abundance brought significant improvement in the way for diseases association. More specifically, a certain gene could be associated with a disease on interest. For instance, remarkable results achieved by proteomics (i.e. the large-scale of protein-protein interactions network) helped us to have a better understanding of protein functions as proteins are the essence of metabolic pathways of cells.

The disease-gene association in wet-labs is a challenging process for biologists. Thousands of genes need to be analyzed, which is not feasible to be conducted in labs. Thus, the high-throughput technologies are utilized to narrow potential candidate genes for a disease of interest. A combination of graph theory (i.e. biology network) and computational intelligence techniques from machine learning tools can help us to better analyze and identify gene functions. Functional networks and machine learning emerging techniques could reveal significant phenotypic genes in remarkable manner of time.

1.3. Objectives and Goals

The main objective and goals of the proposed study is to develop new models capable of predicting genes, and validate pre-exiting genes that are associated or related to a certain phenotype using biological networks data with the latest available Artificial Intelligence (AI) and data mining techniques. These objectives will be achieved through the following millstones:

1. Providing gene function predictions to prioritize genes most likely to be relative or associated with breast cancer through biological networks analysis.
2. Utilizing a protein-protein interactions network and gene expression network in order to investigate breast cancer phenotypic genes.
3. To utilize integrated complex network (functional network) and identify disease specific genes.
4. To apply a network approach and machine learning methods in order to identify breast cancer phenotype.
5. To provide insight of how well topological properties of different biological networks could help identifying biomarkers utilizing, functional genomics, protein-protein interactions network and co-expression datasets.

1.4. Contribution

This thesis work provides an efficient and reliable approach with high accuracy rate to solve the problem of disease-gene association. The thesis contribution could be summarized as follows:

1. Implementation of fourteen different centrality measures and ranking algorithms. These ranking algorithms could be categorized as follows: local and global centrality measures. More specifically, we predict the phenotypic genes based on a collection of topological features, rather than using a single topological feature at a time.
2. Applying various features selection: 1) Pearson correlation coefficients (PCCs) and 2) hill-climbing feature selection to identify the most significant topological features that provide best prediction rate.
3. Building different classification models that allow biologist to predict breast cancer genes.
4. Proposing our co-cross validation scheme to provide a generalized classification performance.
5. Implementation of a SMOTE sampling technique to balance the datasets. In return, enhance the classification performance.
6. Utilizing three publicly datasets (functional genomics network, protein-protein interactions network, and co-expression network) to extract different topological features.

1.5. Outline

The rest of the thesis is organized as follows: Chapter 2 provides a background to the biological, graph theory terminologies, and the proposed ranking and scoring algorithm for the genes in the biological networks.

Chapter 3 provides review of previously published work related to disease-gene association.

Chapter 4 presents how to overcome the problem of unbalanced datasets, the randomization of the biological networks, and more importantly, the implemented classifiers in this study.

Chapter 5 presents the experimental datasets and a description of how the classification models would be evaluated.

Chapter 6 presents the Experimental Results and discussion of the proposed computed scoring algorithms. Finally Chapter 7 concludes the thesis work and highlights the future work directions.

CHAPTER 2

BACKGROUND

2.1. Biological Terminology

Cell biology is the essential component of living organisms, which is the structural, functional and biological unit of such organisms. More specifically, a cell contains further smaller ingredients of DNA, RNA, proteins, and metabolites, which are a part of a tissue. Collections of different tissue form the organs of an organism. More importantly, the organisms are evolving, which results to form relationships among elements in the organism. The relationship among these elements is a place for interest to study. These relationships could be presented as a graph (network), in which the organism's elements are the vertices (proteins, genes, DNA, RNA), and the relationships among these elements could be presented as edges (interactions among the elements). Very well-known types of biological networks at the molecular level are known as gene regulation networks, signal transduction networks, protein interaction networks, and metabolic networks. An example of a biological network is shown in Fig. 1.

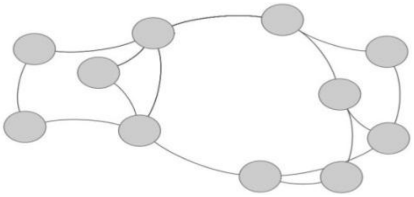


Fig. 1. Example of a biological network that is generated randomly.

2.1.1 Basic Biological Ingredient

The desoxyribonucleic acid (DNA) (see Fig. 2.) stores the information of an organism (Fig. 3.). DNA is viewed as a double coiled ladder (double helix), which consists of two sugar phosphates in the backbones. The backbones are enclosed by pairs of nucleotide bases (i.e. adenine, cytosine, guanine, and thymine (A, C, G, and T). It is worth noting that the nucleotide A pairs only with T, and C pairs only with G. On other hand, ribonucleic acid (RNA) could be viewed as a single helix (Fig. 2.), which the backbone consist of Phosphate and ribose sugar, and backbone consists of the following four bases: adenine, guanine, uracil, and cytosine (A, G, U, and C)

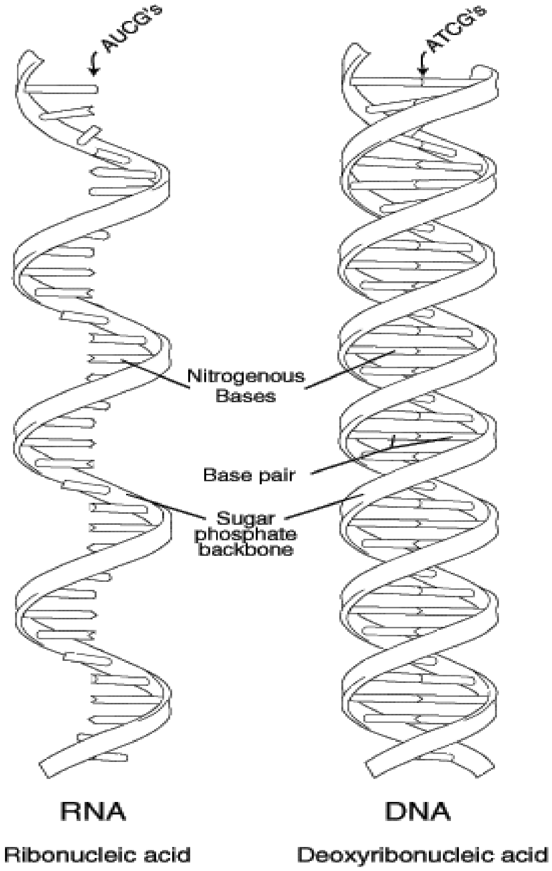


Fig. 2. DNA and RNA structure.

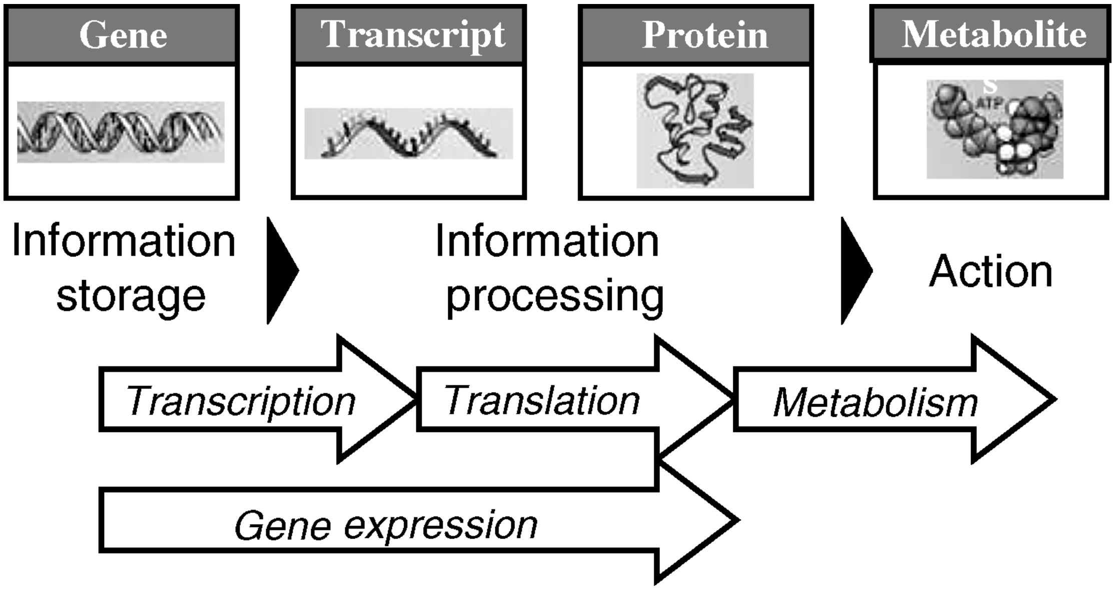


Fig. 3. Information processing staring from genes until research metabolites in cells. The picture is obtained from the book by Junker, Björn H., and Falk Schreiber [7].

The process of information transition from DNA to proteins is known as *gene expression* (Fig. 3.). This process could be explained in two main folds:

1. *Transcription:* is a sophisticated, highly regulated process. A protein complex contains RNA polymerase opens the DNA helix, reads one strand and generates the corresponding RNA like a scheme. The corresponding generated RNA is named as transcript (Fig. 3.)
2. *Translation*: is the final stage of protein generation from generated RNA. RNA goes through a process which is called slicing, such that the *introns* (noncoding regions) are removed and only the exons (coding regions) stays. In translation, amino acid chains are generated from the (spliced) RNA by the ribosomes. The RNA information is read in triplets (*codons*), such that there are $4^3 = 64$

combinations. These combinations are utilized to code 20 amino acids (i.e.one codon might represent more than one amino acid), additionally one start codon and three stop codons.

2.1.2 Cell Biology

Organism could have one of two cells types. *Prokaryotes*, e.g., bacteria, which are only single cells and cannot be subdivided. *Prokaryotes* genomes (i.e. the whole of the genes) are formed as a single circular chromosome (i.e. the structure of DNA, RNA, protein that could be found in the cells). On the other hand, *eukaryotes* cells are more sophisticated (Fig. 4.).

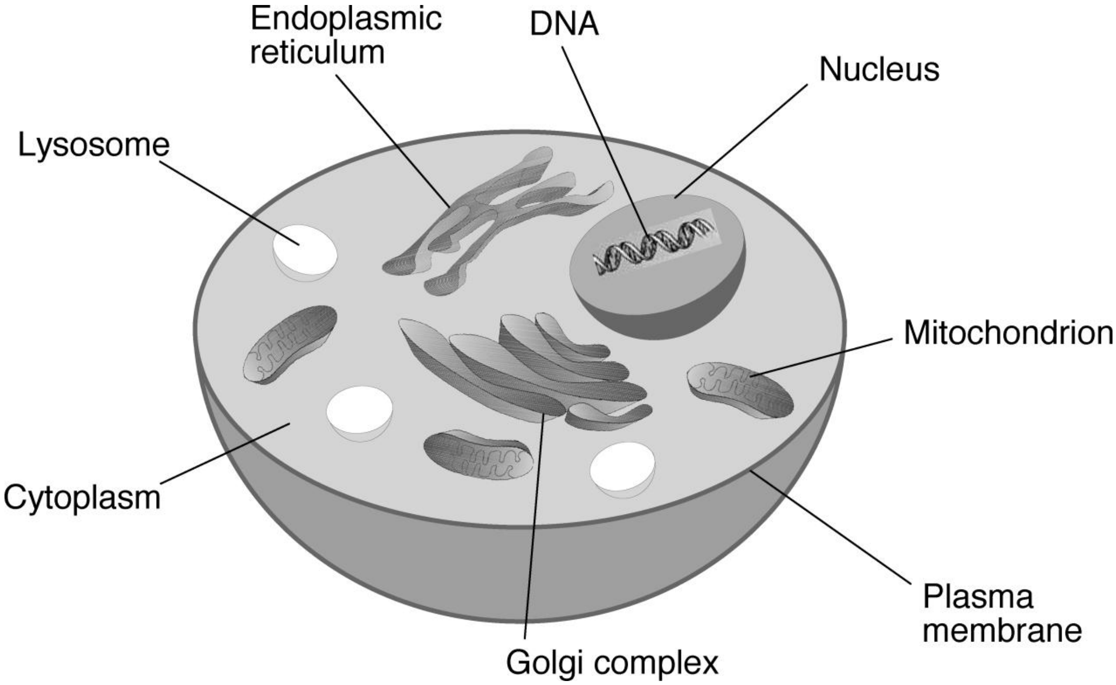


Fig. 4. Elucidation of an animal cell with some organelles. The picture is obtained from the book by Junker, Björn H., and Falk Schreiber [7].

The genome of *eukaryotes* cells is structured on many chromosomes inside the nucleus. Each chromosome contains two *chromatids* (Fig. 5.). It means on the gene level, a eukaryotic cell has at least two copies of every gene. In addition, majority of organism cells contains two sets of chromosomes, one from each parent.

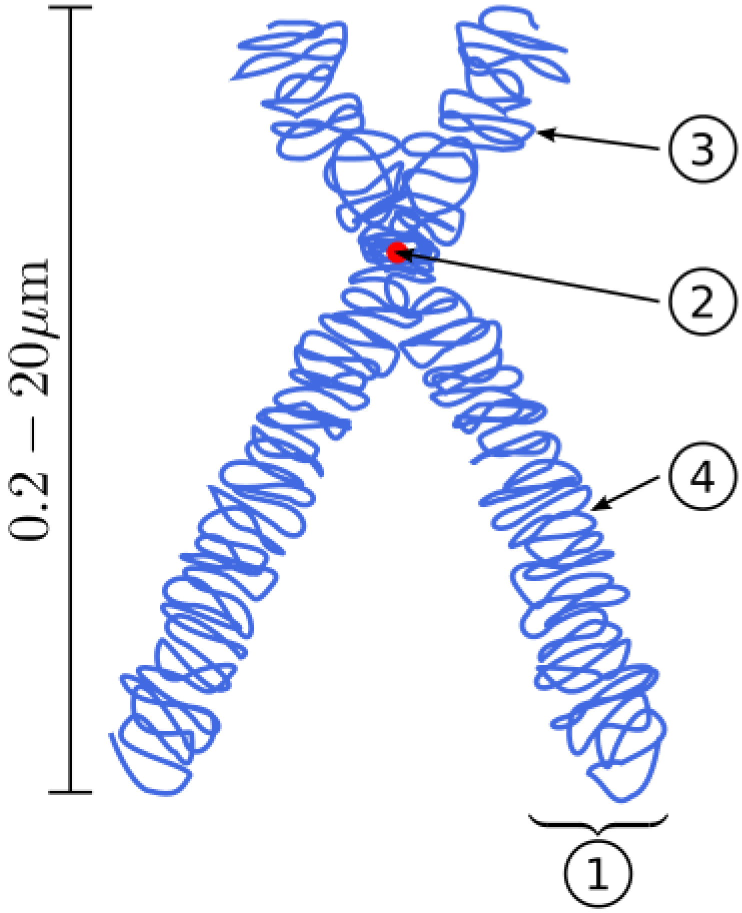


Fig. 5. Illustration of eukaryotes chromosome: 1) chromatid – one of the two identical parts of chromosome, 2) Centromere – where the two chromatids intersect, 3) short arm, 4) long arm.

2.1.3 System Biology

The field of biology is evolving which will shift into accurate science with the availability of high-throughput technologies and bioinformatics techniques. The massive and wealthy datasets that are generated out of high-through technologies shifted the attention rather than dealing with an individual genes, to deal with a collection of genes. This field of biology is called "systems biology: a new field in biology that meant to understand and make sense of a biological system at a system-level". Interestingly, system biology will enable the understanding of totality of biological systems by visualization, modeling, and the prediction of the behavior of a system's components, sub-components, and interaction within components.

As promising as the field sounds, but the abundance of the –omics datasets that are generated from high-throughput technologies make it difficult to make sense of these datasets. To model these datasets and might generate hypothesis to be studied, network (graph) analysis could be used, as it will be described in the following section.

2.2. Graph Theoretic Terminology

This section introduces the fundamental concepts of graph theory that will be used in this thesis work. The mathematical formulation will be described as provided in the literature [7-9].

2.2.1. Undirected Graph

A graph $G$is usually defined as a pair of (V, E), such that V contains the set of vertices, which represent the nodes and E contains the set of edges, representing the interaction between the nodes. Furthermore, E could be defined as follows $E = \{i, j \mid i, j \in V\}$, which is a connection between the two nodes i and j. So, we could say that the two nodes i and j are neighbors. In addition, the two nodes could have a multiedge connection, such that they are two or more edges having the same endpoints. A multi-edges connection is an important feature, such that the two nodes are connected by more than on link. In such case, each connection might convey different type of information.

2.2.2. Directed Graph

A directed graph could be defined as follows, $G = (V, E, f)$, such that the function f maps each member of E to an ordered pair of nodes in V. The ordered pairs are directed edges, known as arcs or arrows. Furthermore, the directed graph is asymmetric, such that if an edge $E = (i, j)$ has a direction from $i$to j, then it is not necessary there be an edge having a direction from j to i.

2.2.3. Weighted Graph

The weighted graph could be defined as $G = (V, E)$, where V is the set of nodes and E is the set of edges between the vertices $E = \{(u, v) \mid u, v \in V\}$. In this case there will be a weight function to assign a score for every edge: $w: E \rightarrow R$, where R denotes the set of all real numbers. Usually the weight w_{ij} that is assigned to an edge has a meaning of how relevant two nodes i and j are, so, the higher the weight, the higher the relevance might be.

2.2.4. Data Structure

The main data structure that has been used to store the network graph representations is *adjacency matrix.* For a given graph$G = (V, E)$, then the adjacency matrix representation size will be $|V|x|V| = nxn$. However, matrix A contains entries either of 1's or 0's, such that $a_{ij} = 1$ if i and $j \in V$ or $a_{ij} = 0$ otherwise. Such that,

$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$, where $n = |V|$. In the case of a weighted graph, the entry $a_{ij} = w_{ij}$if i and $j \in V$ or $a_{ij} = 0$ otherwise. Furthermore, for an undirected graph, the adjacency matrix is symmetric for the reason that $a_{ij} = a_{ji}$. However, the previously mentioned rule is not applicable to directed graphs, for the reason that the upper and lower triangular parts of the matrix shows the direction of the edges. Moreover, the space complexity of the adjacency matrix is $\Theta(|\mathrm{V}|^2)$.Thus, it is highly preferred to be used with dense graphs rather than the sparse graphs. In addition, for an all-against-all symmetric data set, then either the upper or the lower triangular part of the matrix is needed. This will then reduce the amount of space complexity to be $\Theta(|\mathrm{V}|)$. However, such data structure will be well suited with the cluttered networks that have high density connections between the elements. Finally, for the case of the fully connected graph, such that all nodes are connected to each other, then an adjacency matrix is highly suggested.

2.2.5. Graph (Network) Properties

Graph theory could play important role in assessing each node of the network, and could retrieve valuable properties and features that could provide insight about the internal structure of biological networks. In the following, a brief description of the fundamental properties that could be used to analyze a network is as follows:

2.2.5.1. Graph Density

The ***graph density*** helps to reveal the likelihood of a graph, whether it is spare or dense (i.e. depends of the number of links for every node set). This could be computed as follows: $density = \frac{2\,|E|}{|V|(|V|-1)}$. Moreover, for a ***spare graph*** would be a graph meets the following parameters, $|E| = O(|V|^k)$and $2 > k\ > 1$, or otherwise when $|E|$ " $|V|$.On the other hand, a ***dense graph*** would meet the following parameter (i.e. a graph where $|E|$ " $|V|^2$).

2.2.5.2. Average Shortest Path

The average path length of a given network is defined as the average number of edges between nodes. The average path length must be traversed in the shortest path between any two nodes. Furthermore, it can be computed as $\delta = \frac{2}{N\,(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{min}(i,j)$ where $\delta_{min}(i,j)$is the minimum distance between nodes i and j.

It is also worth mentioning that the most well-known two algorithms for computing the shortest paths are Dijkstra's greedy algorithm [10] and Floyd's dynamic algorithm [11].

2.2.5.3. Diameter

The ***graph diameter*** could be found by computing the longest shortest path within a graph. Thus, the diameter is defined $D = {}^{max}_{i,j}\delta_{min}(i,j)$.

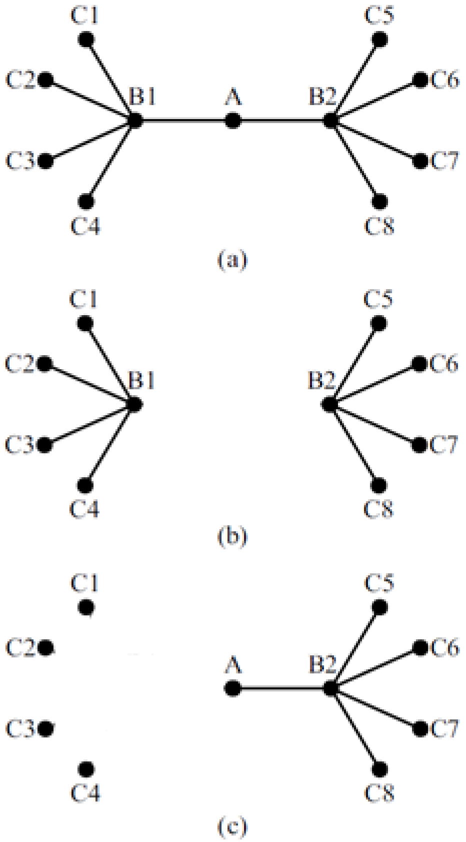


Fig. 6. (a) An original graph to illustrate the purpose for different centrality measures. (b) If we remove A, the network will be disconnected into two sub-graphs. Then, is node A an important vertex?. (c) If node B1 is removed then the graph will be scatted into disconnected components. Therefore, an identification of important node is non-deterministic. The picture is obtained from the book by Junker, Björn H., and Falk Schreiber [7].

2.3. Centrality Measures and Node Ranking

This section provides a mathematical description of different raking theories [7-9, 12], such that how is a certain node could be raked or sorted according to its position in the biological network. More importantly, the obtained score for every vertex is meaningless, unless it is associated with the properly asked question. For example, in the biological networks, how would the detection of central or intermediate nodes affect the topology of the network. It is important to state a goal of finding the central nodes. However, a question could be to find the molecules that are really central, but more importantly, they are participating in a biological pathway and have a crucial role in signal transduction. Furthermore, a centrality measure can only capture a portion of the important elements within a network as illustrated in Fig. 6. Therefore, different centrality measures are developed with different interpretation [13-16]. More specifically, in this thesis we compute fourteen centrality measures for every gene in the biological network.

1. Degree (DC) centrality:

DC reveals that the highly significant nodes participate in a large number of interactions. For a given node i, the degree centrality is defined as $DC(i)$ = the number incident edges. For directed graphs, an in-degree $DC_{in}(i)$ and an out-degree $DC_{out}(i)$ can be defined, such that, $DC_{in}(i)$, represents the number of arcs going toward a node i, but $DC_{out}(i)$, represents the number of arcs going out of a node i. Furthermore, those nodes which are highly connected are called ***hubs***. The scale free networks tend to have hubs. The removal of such highly connected nodes will have an effect upon network topology. The biological networks are robust against random removal of hubs. On the other hand, their removal may cause system failure [17, 18].

2. Closeness Centrality ($\boldsymbol{ClosC}$):

$\boldsymbol{ClosC}$ elucidates the significant nodes that can communicate quickly with other nodes of the network. For a vertex $v \in V$ in graph $G = (V, E)$, it is defined by the following equation:

$$ClosC(i) = \frac{1}{\sum_{t \in V}^{|V|} dist(i, j)} \tag{1}$$

Where$dist(i, j)$ is the length of the shortest path between two vertices i and j. For a node with a high $ClosC$ will require a minimum number of steps to pass information originating from itself to other nodes in the network. It also indicates how a change in a node affects other nodes in the network.

3. Betweeness Centrality ($\boldsymbol{BC}$):

$\boldsymbol{BC}$ indicates the ease by which certain vertices can monitor communications between other vertices. Without the existence of such nodes that monitor communication, then there will be no way for other nodes to communicate. For a vertex $v \in V$ in graph $G = (V, E)$, it is defined by the following equation:

$$BC(v) = \sum_{s \in V \wedge s \neq v} \sum_{t \in V \wedge t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2}$$

$\sigma_{st}(v)$are the number of shortest paths between two vertices s and t that pass through vertex v. If s and t are not connected, $\frac{\sigma_{st}(v)}{\sigma_{st}} = 0$. It shows how influential a node is over the information flow in a network. A node having high BC potentially plays an important role in the network by controlling the flow of interactions.

4. Eigenvector Centrality (EC):

The eigenvector centrality expresses the centrality of a node as dependent on the centralities of its directly connected neighboring nodes. A node is highly ranked if and only if that node is connected to important neighbors. For an undirected graph $G = (V, E)$ and A is the adjacency matrix of network G. Then, the *eigenvector centrality* is the eigenvector C_{eiv} of the largest eigenvalue λ_{max} in absolute value, such that could be obtained from the following system of equations:

$$\lambda\overrightarrow{C_{e\imath v}} = A\overrightarrow{C_{e\imath v}} \tag{3}$$

5. K-Step Markov (KSM):

The K-Step Markov technique calculates the relative probability that the system will spend some time at any particular node, such that, it is given the start set of roots R and shall end after K steps. The author whit *et al.*[19], demonstrates that the number of steps K manage the bias distribution toward the specified roots R. In addition, as K gets larger the steady distribution coverage to PageRank results. However, for a graph $G = (V, E)$ contains vertices $u, v \in V$, and edges $(u, v) \in E$. Next, Let $P_{u,v}$ be the probability to reach v from u in one step,so this probability is the weight of the edge between u and v. Then, let $N(u)$ be the set of neighbor vertices of u. After that, the probabilities are constrained by the following equation.

$$\sum_{\substack{u \in V \\ v \in N(u)}} P_{u,v} = 1 \tag{4}$$

Furthermore, a ***random** walk* is defined as a walk that starts at a particular vertex and traverses the graph based on the$P_{u,v}$. Then, *K-Step Markov centrality* is the probability with which a random walk of length k brings a system to a particular vertex [19]. For vertex$v$, k-Step Markov (KSM) is defined by the following equation.

$$C_{ksm}(v,k) = P^0 A^k \tag{5}$$

WhereP^0 is an initial probability distribution of the vertices,G, and A is the adjacency matrix of G containing the transition probabilities. Finally, in this study we consider k to be 6.

6. *Subgraph centrality ($\boldsymbol{SC}$):*

$\boldsymbol{SC}$ ranking nodes according to the number of times a given vertex participates in different connected subgraphs of a network [15]. A *subgraph* is a closed walk that starts and ends at the same vertex. SC, weighted according to the length of the closed walks. The weight of a closed walk decreases with increase in its length. For a vertex v in undirected graph $G = (V, E)$, and A is the adjacency matrix of G. Then, the *subgraph centrality* for a node that has the length of a close walk k is computed as follows:

$$SC(v) = \sum_{k=0}^{\infty} \frac{(A^k)_{vv}}{k!} \tag{6}$$

7. *Clustering Coefficient ($\boldsymbol{CC}$):*

$\boldsymbol{CC}$ measures the degree of cohesiveness in a given graph. For a given vertex, it is defined as the ratio of actual number of adjacent edges E_i to the maximum number of possible edges. This gives the probability that two vertices with a common neighbor are connected:

$$CC(v) = \frac{2E_i}{n(n-1)} \tag{7}$$

Furthermore, CC measures the edges ratio of a certain node to the total possible number of edges, i.e., $n(n-1)/2$. CC values range in the interval of $0 \leq CC \leq 1$. Thus, as the local clustering coefficient of a certain node gets closer to 1, then it is more likely that node is highly ranked.

8. Flow Coefficient (FC):

The *clustering coefficient* of a node, as shown above, is calculated as the number of all existing connections between the node's neighbors divided by the number of all possible such connections. However, in analogy to the clustering coefficient, the ***Flow Coefficient*** [20] is defined as the number of all paths of length 2 linking neighbors of a central node that pass through the node, divided by the total numbers of all possible such paths. Furthermore, the *flow coefficient* allows estimating the capacity of a node to conduct information flow between its neighboring nodes, thus, the *flow coefficient*, is a measure of ‘‘local centrality’’. It is calculated as the number of actual paths of length 2 divided by the number of all possible paths of length 2 that traverse a central node. Hub regions that act as bridges between different communities of nodes are likely to exhibit small *clustering coefficients* and large *flow coefficients*.

9. Katz Status Index (KSI) centrality:

$\boldsymbol{KSI}$ ranks a node as highly important if a large number of vertices are connected to it. Both direct and indirect neighbors of a vertex contribute to its importance. For a graph with adjacency matrix A, KSI is defined by the following equation:

$$\overrightarrow{\boldsymbol{KSI}} = ((1 - \alpha A^T)^{-1} - I)\vec{1} \tag{8}$$

$\alpha$is a scaling factor that is less than the reciprocal of the absolute value of the largest eigenvalue of A ($0 \leq \alpha \leq 1/|\lambda_{max}|$). $A^T$the transpose of A, I is an identity vector, $\vec{1}$ is a vector of ones.

10. Node Coreness (NC):

Node coreness [21] measures the set of nodes that are highly and mutually interconnected which are known as network core. For a binary network, the k-core is the largest subgraph comprising nodes of degree at least k, and is derived by recursively removing off nodes with degree lower than k until none remain. Each node is then assigned a *core number*, which is defined as the largest k such that the node is still contained in the k-core. The k-core decomposition is applied on binary connections.

11. Within Module Z-score (WMZ):

The ***within module z-score*** [22] measures how nodes are relatively related within a module. Moreover, *modules* could be organized in different ways, such that some modules could be totally centralized with one or few nodes connected to all others. However, some modules could be totally decentralized, with all nodes having similar connectiveness. Furthermore, the nodes with similar roles are anticipated to have similar relative ***within-module connectivity***. If k_i is the number of links of node i to other nodes in its moduless_i, $\bar{k}_{s_i}$ is the average of k over all the nodes in s_i, and σ_{s_i} is the standard deviation of k in s_i Then:

$$z_i = \frac{k_i - \ \bar{k}_{s_i}}{\sigma_{s_i}} \tag{9}$$

this is the so-called *z-score*. The within-module degree z-score measures how well-connected node $\boldsymbol{i}$ is to other nodes in the module.

12. *The Burt's measure of constraint(Structural Holes (SH)):*

Structural holes[23]measures the degree to which an individual has exclusive exchange relations to otherwise disconnected partners and groups. Individuals with more structural holes are positioned for entrepreneurial action as they can control the flow between people on opposite sides of structural holes. The Burt's measure of constraint could be found using the following equation:

$$c_{ij} = (p_{ij} + \sum_{k,k\neq i,k= \neq j} p_{ik}p_{kj})^2 \tag{10}$$

Where $p_{ij}$is the adjacency matrix of G containing the transition probabilities. Given the p_{ij} matrix, the indirect constraint ($p_{ik}p_{kj}$), could be obtained with the 2-step path distance.

13. *Proximity Prestige (PP):*

Proximity prestige [24] is used to determine what is known as the node's *influence domain*. Such that the *influence domain* of a node could be defined as the set of nodes who are both directly and indirectly linked to that node. This set includes those nodes that are reachable to node i.If there exists a path then these two nodes are called reachable from one another. The *influence domain* for node i includes all nodes whose entries in the i-th column of a distance matrix are finite. The number of nodes in the *influence domain* for a node i is denoted byI_i. Furthermore, the *proximity prestige* could be measured as the ratio of the proportion of nodes which can reach $i$to the average path length of these nodes from$i$.

$$PP(n_i) = \frac{I_i/(g-1)}{\sum_{j=1}^{I_i} d(n_j\,, n_i)/I_i} \tag{11}$$

Where$I_i$is the number of nodes in the influence domain of node n_i, $g$is the total number of nodes in the graph and$d(n_j, n_i)$is the *geodesic distance* that node $j$is from node i.

14. Bary-Center Score (BCS):

BaryCenter score[19] ranks each node of the graph depending on the total shortest path of the node. It computes the shortest path distances for each node in the graph, and a score will be assigned for each node based on the lengths of the shortest paths that go through the node. Moreover, the highly central nodes in a connected component tend to have smaller overall shortest paths. On the other hand, the 'peripheral' nodes on the network tend to have larger overall shortest paths.

CHAPTER 3

LITERATURE REVIEW

The identification of gene-phenotype association has been studied in many different aspects. In literature, there are studies which utilize the protein-protein interactions (PPIs) [25] and gene expression [25] solely without the integration of prior knowledge (i.e. Gene Ontology (GO)) or integration with different data types. More recently, high-throughput technologies like microarrays and, most recently, next generation sequencing have increasingly generated massive datasets. These datasets were shown to examine false positive interactions [26, 27].Thus, several authors have proposed different techniques that some utilizing supervised machine learning and others are utilizing un-supervised machine learning for the purpose of different data type integration. Moreover, the following review several works that has published of how the aforementioned problem tackled to investigate disease-causing genes.

3.1. Gene Expression

The gene expression datasets have been used extensively for the purpose of diseases-causing gene. There are many studies utilizing gene expression in different perspectives. Such that, the genes expression profiles are used as is and fed as features to the classification problems, and some researches might filter or apply feature selection prior the classification. More specifically, this section reviews published work based on the use of gene expression profiles. The reviewed work is further categorized into: 1) Machine

Learning-Based Disease-Gene Association and 2) Co-expression Based Disease Gene-Association.

3.1.1. Machine Learning-Based Disease-Gene Association

This section reviews published work based on the machine learning techniques to predict the disease-gene association. Automated models could be built for the purpose of predicting the phenotypic genes based on the raw gene expression profiles. For instance, one of the pioneer studies published by Furey *et al.* [28]proposed support vector machines (SVMs) for the purpose of analyzing tissue samples of microarray data. The performed analysis has two folds, firstly, the classification of the tissue samples, and scanning of the data for the mislabeled tissue results. Moreover, the Furey *et al*. approach was evaluated using ovarian cancer tissues, normal ovarian tissues, and other normal tissues. For each tissue that consists of expression results for 97,802cDNAs. Furthermore, the proposed approach was able to reveal mislabeled tissue samples, after fixing these mislabeled samples and removing the outliers, the SVMs achieved perfect classification. However, the confidence level was not high.

Ramaswamy *et al.* [29] proposed to utilizes 218 tumor samples, spanning 14 common tumor types, and 90 normal tissue samples to oligonucleotide microarray gene expression analysis. The expression levels of 16,063 genes and expressed sequence tags are employed for the evaluation of accuracy of the multiclass classifier. Ramaswamy's *et al*. study implements SVMs to distinguish the tumor and normal samples. The classification model was able to achieve an accuracy rate of 78%**.** A similar study which is done by Lee *et al.* [30] proposed the use of the extended support vector machines [31], which will be used to classify gene expression data of multi-category cancer problems. The Multi-category SVM was evaluated using two gene expression data sets, first of which leukemia data that comes with three classes [32]. The other dataset is the small round blue cell tumors of childhood data [33] which is a four classes problem. The resultant of this study showed comparable classification accuracy to the other classification methods.

There are different studies, which implement different classification models for gene expression datasets. For instance, Wei *et al.* [34] proposed an automated model to classify patients and to which risk group of neuroblastoma they belong. This study utilizes gene expression profiling from cDNA microarrays that contains 42,578 clones. Moreover, an artificial neural networks classification model is adapted to build a predictor of survival for each individual patient with neuroblastoma. Additionally; principle component analysis (PCA) is applied and revealed that neuroblastoma tumors exhibited inherent prognostic specific gene expression profiles. In addition, expression levels of 37,920 good-quality clones are fed to an artificial neural network-based prognosis prediction; however, the achieved accuracy rate is 88%. Further analysis is applied using an artificial neural network-based gene minimization technique that reveals 19 genes, including 2 prognostic markers that have been reported before. They are MYCNand *CD44,* which correctly predicted outcome for 98% of these patients. Theses identified 19 genes able to classify the Children's Oncology Group-stratified high-risk patients into two subgroups according to their survival status (*P* = 0.0005).

A different study that has been published by Xia *et al.* [35], who developed an ensemble decision approach that can do gene mining tasks. Xia *et al.* has utilized two publicly available datasets for analysis purposes, which are colon data [36] and leukemia data [32]. However, the analysis has revealed 20 significant colon cancer genes, and 23 significant molecular signatures for refining the acute leukemia phenotype. The identified genes are verified using biological experiments or by alternative analysis approaches.

Dettling *et al.* [37] proposed to modify and extend the boosting classifiers for microarray expression data that comes from several tissue or cancer types. In addition, feature selection is applied and LogitBoost is used, which is integrated with several approaches for binary problems. The Dettling *et al.* algorithm was evaluated using six real and one simulated gene expression datasets. This approach does not implement sophisticated tuning or kernel selection and provides directly class membership probabilities.

Previously reviewed research studies one classification model which lacks the comparison of results with other different classification models. Guthke *et al.* [38] studied the use of supervised and unsupervised clustering techniques and machine learning algorithms for automated model relationships between gene expression data and gene functions of microorganism *Escherichia coli* [39]. Moreover, the Guthke *et al.* study utilized pre-selected subset of 265 genes (which belong to 3 functional groups) However, the functions have been predicted with an accuracy rate of 63-71% by various data mining methods. For example, some of the machine learning algorithms that were presented in this paper are: K-means clustering, Kohonen’s self-organizing maps (SOM), Eisen’s hierarchical clustering and Quinlan’s C4.5 decision tree induction algorithm. Moreover, the Guthke *et al.* study introduced the use of the fuzzy approach for gene expression data analysis. Such that, it has investigated the used of the fuzzy-C-means algorithm (FCM), the Gustafson-Kessel algorithm for unsupervised clustering, and the Adaptive Neuro-Fuzzy Inference System (ANFIS) for the purpose of the functional classification of *E. coli* genes.

Moreover, Pirooznia *et al.* [40] also studied different classification models to investigate different diseases through the gene expression datasets. The implemented classification models are SVM, Neural Networks, MLP Neural Networks, Bayesian, Decision Tree, and Random Forest. In addition, some of the unsupervised machines learning models used include: K-means, Density Based Clustering (DBC), and Expectation Maximization (EM) clustering. Furthermore, the gene expression data dimension was reduced using different feature section algorithms for comparison purposes, which include: support vector machine recursive feature elimination (SVM-RFE), Chi Squared, and Correlation based Feature Selection (CFS) [41]. More importantly, K-fold cross-validation is utilized for the evaluation purposes. For each case mentioned previously, eight different binary (two class) microarray datasets are used to investigate the following diseases: Lymphoma, Breast Cancer, Colon Cancer, Lung Cancer, Adenocarcinoma, Lymphoma, Melanoma, and Ovarian Cancer.

3.1.2. Co-expression Based Disease-Gene Association

The gene expression profiles could be used to reveal the co-expression links based on the correlation between each pair of gene expression. More interestingly then, gene expressions could be modeled as a network. In this case, the disease-causing genes could be studied in more depth, rather than studying a gene at a time. Hence, it is possible to study genes that interact with each other and might rank a certain gene to reveal its susceptibility to cause a disease of interest. More importantly, the ability to study the gene neighbors for the possibility that they might cause the same disease of interest. For instance, Lee *et al.*[42], who introduced a large-scale analysis of mRNA co-expression utilizing 60 large human data sets contains a total of 3924 microarrays. The Lee *et al.* study explored the pairs of genes that were reliably co-expressed (based on the correlation of their expression profiles) in multiple datasets. Moreover, a high-confidence network was built consisting of 8805 genes linked by 220, 649 “co-expression links” that are observed on at least three datasets. This study has revealed that the co-expression in multiple datasets is correlated with functional relatedness. Additionally, the large body of accumulated microarray data can be exploited to increase the reliability of inference about gene function.

Ruan *et al.*[43] proposed a general co-expression network-based technique that allows analysis of genes and samples obtained from microarray datasets. The Ruan *et al.* technique has a rank-based network construction method, a parameter-free module discovery algorithm, and a reference network-based metric for module evaluation. This study has utilized different datasets for evaluation purposes, i.e. yeast, Arabidopsis and human cancer microarray.

A more sophisticated technique was proposed by de Matos Simoes *et al.* [44] to infer casual gene regulatory networks from large-scale gene expression data and a proposed method called BC3NET. The proposed method is an ensemble-based on the bagging of a C3NET algorithm. Furthermore, the BC3NET algorithm matches with the Bayesian technique but with non-informative priors. The proposed algorithm was evaluated using simulated and biological gene expression data from S. cerevisiae.

Interestingly, Zhu *et al.* [45] proposed a network-based SVM with hinge loss function to classify a binary class problem through the utilization of the penalty term from F_∞-norm, which will be applied to pairwise gene neighbors for prediction and gene selection. The Zhu *et al.* algorithm was evaluated using low and high-dimensional data, additionally; two real microarrays are used for evaluation. The proposed algorithm is able to achieve a similar or higher prediction accuracy compared with the standard and the L_1 penalized SVMs.

In addition, Wu *et al.* [46] proposed the utilization of a machine learning model for the identification of prognostic biomarkers in cancer gene expression data sets, utilizing modules inferred from a highly reliable gene functional interaction network [47]. The proposed work could be divided into two folds: 1) module discovery, and 2) module validation. For the model discovery, a breast cancer tissue microarray is used [48], but for the validation module, four different independent breast cancer microarray expression datasets are used [25]. The Wu *et al.* approach algorithm build-up on the functional interaction (FI) network to find out the prognostic signatures as it is un-weighted and is not for a certain tissue of phenotype. However, the Pearson correlation coefficients (PCCs) is calculated among all functional interaction pairs in the gene expression dataset, and then the PCCs weights are assigned to the edges of the FI network. In this case, the un-weighed graph is converted to be weighed for a particular disease. Then, the Markov clustering (MCL) [49] is applied to cluster the weighted network into a series of gene interaction modules. Moreover, this approach has revealed a 31-gene signature associated with patient survival for the breast cancer application. The signature is repeated through the 5 independent gene expression studies. Furthermore, this study also has revealed a 75-gene that is signature associated to the patient survival for the application of ovarian cancer.

Most recently, Zhang *et al.*[50] proposed a network-based Cox regression model (i.e. it is called Net-Cox). The proposed model is meant to investigate the gene expression signatures (i.e. these signatures probably participate in the results of death or repetition in ovarian cancer medication). The analysis of the co-expression and the prior knowledge of gene functional relations are utilized to construct the relations between gene expressions as a gene relation network. Furthermore, the Net-Cox model is then used to study the co-expression or functional relation within the gene expression features for survival prediction outcome in ovarian cancer treatment. The Zhang *et al.* study does not solely depend on the univariate Cox regression [25].Additionally, it utilizes the co-expression and functional information to build up the gene networks. More importantly, the network-based analysis will assist in identifying sub-network signatures and help in the prognosis of survival in ovarian cancer therapy. However, Net-Cox detected a signature gene, FBN1, which was validated as a biomarker for predicting early recurrence in platinumsensitive ovarian cancer patients in a laboratory.

Table 1 Literature Overview of studies utilizing Gene Expression Profiles

Author	***Method***	***Tackled Problem***
Machine Learning-Based Disease-Gene Association		
Furey *et al. 2000*	SVM	Ovarian cancer

Author	***Method***	***Tackled Problem***
Ramaswamy*et al.* 2001	SVM	Tumor and Oligonucleotide
Guthke *et al.*2002	Supervised and unsupervised machine learning techniques	Reveal relationship between gene expression and gene functions
Lee *et al.*2003	SVM	Leukemia and Round blue cell tumor
Dettling *et al.*2003	Boosting classifier	Leukemia, Colon, Estrogen and Nodal, Lymphoma
Wei *et al.* 2004	Neural Networks/PCA	Neuroblastoma and Colon
Xia *et al.* 2004	Ensamble decision technique	Acute Leukemia
Pirooznia *et al.*2008	Supervised and unpupervised machine learning techniques/ Different features slection are used	Lymphoma, Breast Cancer, Colon Cancer, Lung Cancer, Adenocarcinoma, Lymphoma, Melanoma, and Ovarian Cancer
Co-expression Based Disease-Gene Association		
Zhu *et al.* 2009	Network-based SVM (Using biological prior knowledge to build SVM)	Parkinson's disease and Breast cancer
Ruan*et al.* 2010	Rank-based network construction	Co-expression links prediction
De Matos Simoes *et al.* 2012	BC3NET: A tool to infere gene network from gene expression	Yeast and simulated gene expression are used to validate BC3NET

Author	***Method***	***Tackled Problem***
Wu *et al.* 2012	Markov clustering scheme is used to reveal funtional group	Breast cancer biomarkers prediction
Zhang *et al.* 2013	Network-based Cox regression model	Ovarian cancer

3.2. PPIs Literature Review

Protein networks have been used for great analysis to reveal molecular evolution, and the analysis of protein functions and identification of new functions. They have been used for validating existing diseases and, more importantly, revealing new diseases. The protein network simply could be presented as graph that contains set of nodes and edges. The nodes represent the genes and the edges represent the interaction between these genes. For example, protein-protein interactions (PPIs) networks could be easily seen as undirected graphs, such that, the nodes are the proteins, and if the proteins are physically known to interact then they are connected by an undirected edge. More specifically, this section reviews published work based on the use of PPIs networks. The reviewed work is further categorized into: 1) Guilt by Association and 2) Kernel Models-based disease-gene association.

3.2.1. Guilt by Association

The guilt by association is based on the proximity concept, such that the analysis of genes those are neighbor to known disease genes (seeds). For instance, Goh *et al.* [51] assumed that if genes are neighbors in the network of a disease-causing gene, there is possibility they yield the same or a similar disease. Furthermore, the networks models have been used in different research. For instance, Navlakha *et al.* [52] proposed to use the PPIs network to investigate disease-gene association. Navlakha *et al.* has utilized several published computational methods (i.e. direct network neighbors [53], graph summarization [54], markov clustering [55], semi-supervised graph Partitioning [56], random walks [57], and network flow [58]) to rank and associate genes to a disease of interest. This study has revealed that the random-walk technique outperforms the clustering and neighborhood techniques. Furthermore, most methods propose predictions that are not made by other methods. In addition, the combination of the previously mentioned methods results in an optimal decision. Navlakha *et al.* investigated the diffuse topological distribution of disease-related proteins which negatively affects the quality of the prediction. Thus, it has the ability to reveal diseases that are adjustable to network-based prediction. However, the algorithms were evaluated using Human Protein Reference Database (HPRD) [59].

Vanunu *et al.* [60] built a gene prioritization network-based model to predict a genes and proteins complex that are related to a certain disease. The proposed algorithm is called PRINCE (PRIoritizatioN and Complex Elucidation).This approach combines the protein-protein interactions (PPIs) information with disease of similarity metrics and then the interactions will be scored to reveal how strongly those proteins associated with a certain phenotype of interest. The score methodology is done as follows: 1) the utilization of seeds of genes that have the same phenotype or similar ones, 2) the seed genes repeatedly propagates the flow to the network neighbors. The aforementioned two steps will reveal the strength-of-association (i.e. the neighbor nodes are assigned similar values), such that, each protein propagates the flow received in the prior iteration to its neighbor. Each protein has an assigned score (i.e. computed by the amount of flow it has).This score is used in the integration with a PPI network for the prediction of the protein complex that might be involved in the given disease. Furthermore, there are three diseases are studied (Prostate Cancer, Alzheimer Disease, and Non-insulin-dependent Diabetes Mellitus (Type 2)).The phenotypic genes are collected from Online Mendelian Inheritance in Man (OMIM) [61]. In addition, the used PPIs network contains 9998 proteins and 41072 interactions. These PPIs are obtained from three large scale experiments [25], additionally, Human Protein Reference Database (HPRD) [62] is used. Finally, the proposed algorithm was tested using a leave-one-out cross-validation setting.

3.2.2. Kernel Models-based disease-gene association

Kernels models based on the utilization of extracted features out of the datasets. Then, these features are applied for the purpose of disease-gene prediction. For instance, Lage *et al.* [63] developed a systematics, large-scale investigation of the human protein complexes including products that are implicated in several categories of human diseases to create the phenome-interactome network. However, this study is constructed through the integration of controlled interactions of human proteins with a computationally validated phenotype similarity score. This will allow the prediction of the previously unknown complexes that might be associated with a disease. In addition, Lage *et al.* implemented a Bayesian predictor model where a phenomic ranking of protein complexes likened to human disease are utilized for building such a model. This model is able to predict 298 out of 669 linkage intervals correctly rank the known disease-causing protein as the top candidate. In addition, the Bayesian model in 870 intervals with no identified disease-causing genes was able to reveal candidates implicated in different disorders.

Reyes *et al.* [64] proposed to tackle the problem of prediction of protein-protein interactions (PPIs) when integrating different types of biological information, such that two proteins are classified whether they do or do not interact utilizing different machine classification models. The formation of the data yields to two essential problems that can affect the results: 1) the imbalance class problem (i.e. the positive samples that really interact smaller than the negative samples), and 2) the selection of negative examples. Reyes *et al* has used the one-class classification (OCC) to predict the PPI, the advantage of OCC utilizing only one class to generate the predictive model that is independent of the type of negative examples chosen, which will have the ability to handle the imbalanced class problem.

Finally, it is obvious the concept of PPIs utilization to reveal the disease-causing-genes and the assumption of using 'seeds' of known genes that are associated to phenotype of interested has rapidly been used to validate and tackle new diseases.

Table 2 Literature Overview of studies utilizing PPIs Network

Author	***Method***	***Tackled Problem***
Kernel Models-based disease-gene association		
Lage *et al.* 2007	Bayesian predictor model	Retinitis pigmentosa, epithelial ovarian cancer, inflammatory bowel disease,

Author	***Method***	***Tackled Problem***
Kernel Models-based disease-gene association		
		amyotrophic lateral sclerosis, Alzheimer disease, type 2 diabetes and coronary heart disease
Reyes *et al.* 2007	One class classification model	Predict PPIs of a network
Guilty By Association		
Navlakha *et al.* 2010	Direct network neighbor, graph summarization, Markov clustering, semi-supervised graph Partitioning, and random walks, and network flow.	Identify family-based diseases
Vanunu *et al.* 2010	Prioritization and Complex Elucidation	Prostate Cancer, Alzheimer Disease, and Non-insulin-dependent Diabetes Mellitus (Type 2)

3.3. Genomic Data Integration Literature Review

There is a new era of the high-throughput data utilization, rather than investigating every single 'omics' data solely, are integrated to draw a bigger view of available genomics, transcriptomics, and proteomics datasets. Furthermore, the integration to such data will provide a complete description model to study and analyze a certain biological phenomena. It will also validate existing results, even more to predict new findings that are not found in previous naïve models. Furthermore, a review paper has been published by Georgia *et al.*[65] which reviews the different methodologies that have been known in the literature of how the different data types possibly could be integrated (i.e. the data integration could be done thorough Bayesian models, Kernel models and SVM, or Guilt-by-association models). In this thesis, we adopted the same categorization.

3.3.1. Guilt by Association

The guilt by association is based on the proximity concept, such that the analysis of genes those are neighbor to known disease genes (seeds). For instance, Savage *et al.* [66] proposed a method that integrates gene expression and transcription factor binding (ChIP-chip) data that allows direct inferring transcriptional modules (TMs). This method expands the hierarchical Dirichlet process mixture model to allow data fusion on a gene-by-gene basis. Moreover, this method encodes the co-expression and co-regulation that are not needed to be equivalent. Thus, the genes are not expected to group in the same way in both datasets. In addition, this method allows revealing the subset of genes that share same structure of transcriptional modules in both datasets.

Singh-Blom *et al.* [67] proposed two methods that relies on functional gene associations and phenotype-gene associations in model organisms for the purpose of disease-gene association. Two methods, the Katz measure method, and the second one, CATAPULT (Combining dATaAcross species using Positive-Unlabeled Learning Techniques) is a supervised machine learning technique that utilizes a biased support vector machine, which features are derived from walks in a heterogeneous gene-trait network. The evaluation of the Singh-Blom *et al.* algorithm is done utilizing two data sets, OMIM phenotypes and drug-target interactions. This study has revealed that the

Katz measure is better at identifying associations between traits and poorly studied genes; whereas, CATAPULT is better suited to correctly identifying gene-trait associations overall.

3.3.2. Bayesian Models

The Bayesian models could be further subdivided into: 1) Bayesian network models, 2) Bayesian classification models. The Bayesian network is constructed using Directed Acyclic Graph (DAG), such that each node (variable) shows how it conditions on other variables. More importantly, the Bayesian network could integrate datasets that are dissimilar. For instance, Huttenhower *et al.*[68] proposed a scalable Bayesian framework for predicting functional relationships from integrated microarray datasets, which is called the Microarray Experiment Functional Integration Technology (MEFIT). The MEFIT framework predicts the functional relationships within the context of specific biological processes. This study has integrated 40 Saccharomyces cerevisiae microarray datasets spanning 712 unique conditions using MEFIT. Furthermore, the test of this method based on 110 biological functions utilized from GO biological process ontology, showed a 5% or greater performance increase for 54 functions, with a 5% or more decrease in performance in only two functions.

Huttenhower *et al.* [69] developed a Sleipnir C++ library that implements multiple machine learning and data manipulation algorithms with a focus on heterogeneous data integration that could handle large biological data collections. Sleipnir has the ability to process microarray, functional ontology mining, clustering, Bayesian learning and inferences and support vector machine that could be applied to heterogeneous data on a large scale. Moreover, this library has the advantage of being integrated with other computational systems.

Nguyen *et al.* [70] proposed a toolbox for learning a globally optimal dynamic Bayesian network (DBN) structure from the gene expression data sets. Nguyen *et al* utilized the mutual information test (MIT) scoring system; such that the globally optimal learning DBN is achieved in polynomial time.

Wang *et al.* [71] proposed the identification of functional modules out of protein-protein interaction (PPI).The proposed technique is called the PPI-based Bayesian Models (PPIBM). The PPIBM approach integrates both the gene expression data analysis and clustering of protein-protein interactions. Furthermore, PPIBM utilizes the Bayesian model, which uses its base protein-protein interactions given as a part of input.

Chang *et al.* [72] proposed the use of a Bayesian network-based technique for the purpose of a single-nucleotide polymorphism (SNP) and expression microarray data integration. The integrated network models the SNP-gene interactions through the phenotype centric network. However, Chang *et al.* consists of two parts, variable selection and network learning. The learnt network elucidates how functionally dependent SNPs and how genes effects each other. In addition, the learnt network could be utilized for the purpose of predicting phenotypes. Moreover, this study has investigated the pediatric acute lymphoblastic leukemia dataset.

Yuanfang *et al.* [4] proposed an approach that overcomes the limitations of the quantitative genetics techniques through the application of functional genomics. The proposed algorithm integrates the genome-wide functional relationship network and

SVM. The functional relationship network was first constructed utilizing the laboratory mouse through the integration of multiple different data types, excluding phenotype and disease data to avoid contamination in the evaluation process. The data was integrated utilizing Gene Ontology as a gold standard and using an established Bayesian pipeline [73]. Moreover, the constructed network was fed into SVM for the prediction of related phenotypes. The proposed algorithm is evaluated through predicting genes associated with each of the 1,157 diverse phenotype ontology terms. This study has investigated bone mineral density (BMD), a phenotype related to an osteoporotic fracture, and revealed two genes (i.e. Timp2 and Abcg8) that are related to bone density defects, which are not identified in other statistical methods (i.e. genome-wide association studies / quantitative trait loci).

Wu *et al.* [47] developed a functional interaction, an (FI) network that combines both curated interactions from Reactome [74], other pathway databases, such as Panther [75] CellMap, NCI Pathway Interaction Database, and KEGG, additionally, un-curated pairwise relationships. This study utilizes collected data from physical PPIs in human and model organisms, gene co-expression data, protein domain-domain interactions, proteins interactions generated from text mining, and Gene Ontology annotations. The Wu *et al.* technique utilizes the naïve Bayes classifier (NBC) to reveal the high FIs from non-functional pairwise relationships, to neglect the false positives. The built protein functional interaction network (i.e. combines the curated pathways with non-curated source of information) covering close to 50% of the human proteome, then using this network to investigate two Glioblastoma multiforme (GBM) data sets and projecting the cancer candidate genes onto the FI network.

3.3.3. Kernel Models and SVM

Kernels models and SVM based on the utilization of extracted features out of the datasets. Then, these features are applied for the purpose of integration and disease-gene prediction. For instance, Yu *et al.*[76] investigated biological systems through the integration of both the human protein reference database [59],the protein-protein interaction (PPI) network and the gene expression omnibus [77] data to study lung cancer. Furthermore, A SVM classification model was implemented to predict the PPIs.

Most recently, Qian *et al.* [78] studied the *Multiple Sclerosis* disease using two gene expression datasets and the HPRD protein-protein interaction for the integration process. The Qian *et al.* technique was divided into four steps: 1) the inference of the gene states are done by Hidden Markov Model (HMM)/Gaussian Mixture Model (GMM) hybrid model, 2) the extraction of the bi-clusters out of the gene state sequences, 3) assigning a score for each bi-cluster, dependent on its genes' connection in the PPI network, and 4) K-nearest neighbor classification model is trained and then the SVM classification model is utilized on the collection of K-nearest neighbors.

Finally, in literature [25] has investigated that the pairwise interactions could be incomplete (false negatives) and contains noise (false positives). For that reason, many studies have been published toward integrating several data types and knowledge to provide highly reliable and efficient one single probabilistic integrated model, such that, this model is able to identify disease markers optimally [79].

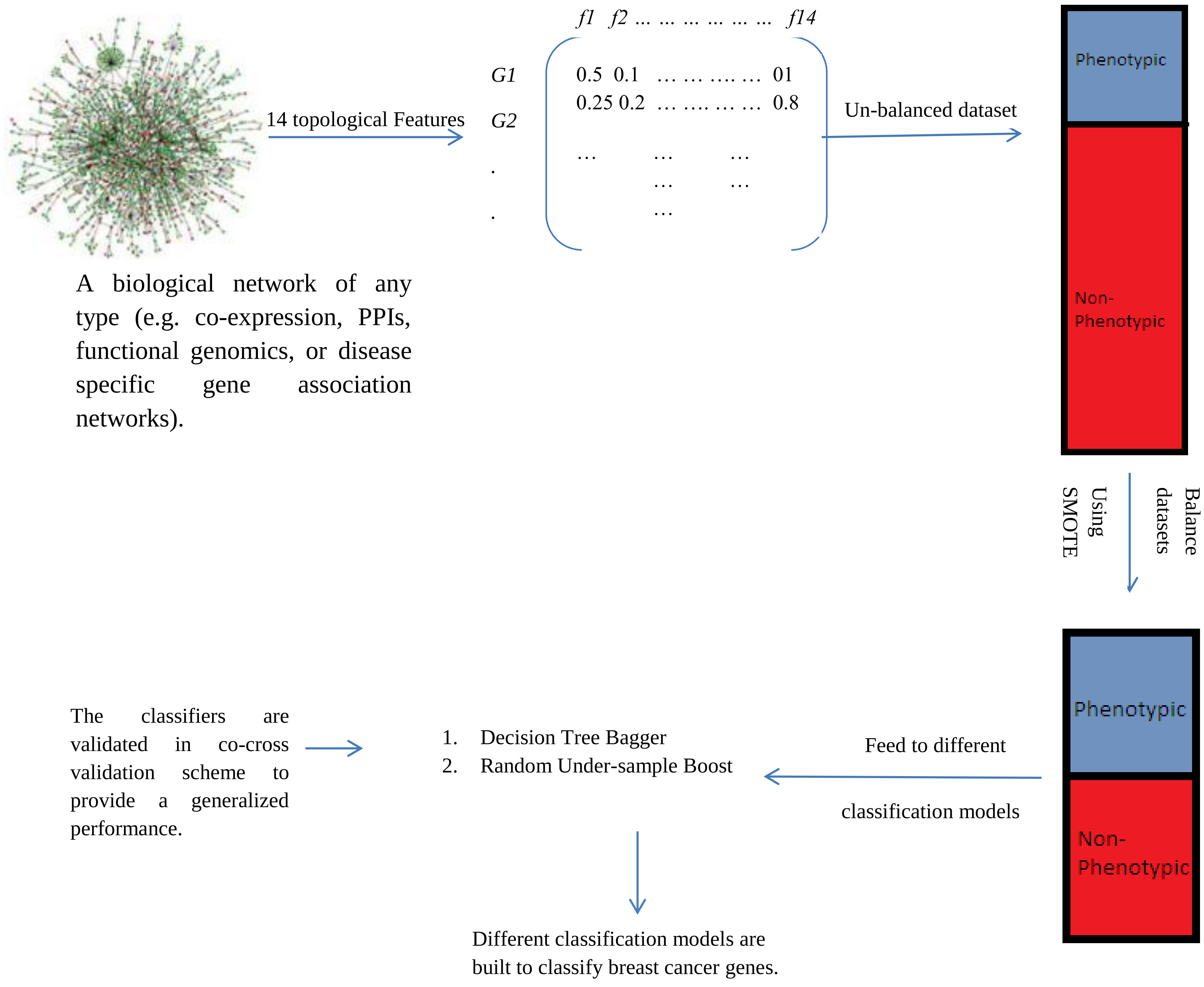
Table 3 Literature Overview of studies utilizing Functional Genomics Network

Author	***Method***	***Tackled Problem***
Guilt By Association		
Savage *et al. 2010*	Hierarchical Dirichlet process mixture	Gene expression and transcription factor integration
Sigh-Blom *et al.* 2013	Katz and CATAPULT	OMIM phenotypes and drug-target interactions prediction
Bayesian Models		
Huttenhower*et al.* 2006	Bayesian framework for predicting functional relations	Integrating various Saccharomyces cerevisiae microarray datasets
Huttenhower*et al.* 2008	Sleipnir C++ library	Allow integration of various of datasets using: 1) Bayesian classification model or 2) SVM
Yuanfang *et al.* 2010	Byesian based tehcnique to integrate different datasets/SVM for classification	Bone mineral density
Wu *et al.* 2010	Naïve Bayes classifier to integrate varous datasets	Study glioblastoma-multiforme brain tumor
Nguyen *et al.* 2011	Bayesian network construction	A tool for learning globally optimal dynamic Bayesian network structure
Chang *et al.* 2011	Bayesian network-based technique to integrate SNP and gene expression	Pediatric acute lymphoblastic leukemia

Author	***Method***	***Tackled Problem***
Wang *et al.*2012	PPI-based Bayesian integration	Integrates gene expression and PPIs network
Kernel Models and SVM		
Yu *et al.* 2013	SVM-based classification model to reveal PPIs	Lung cancer prediction
Qian *et al.* 2013	K-nearest/SVM	Multiple Sclerosis disease prediction

3.4. Our Proposed Method

In this study, we extract fourteen topological features. These features are very well-known in literature which are used to study genes either in local or global manner in the biological or social networks. More importantly, these topological features (or raking algorithms) are typically used to prioritize candidate genes based on one raking algorithm at a time. However, this study attempts to use fourteen ranking algorithm at once through the utilization of efficient and reliable classification models. Furthermore, the classification performance is enhanced using a SMOTE sampling technique to overcome the imbalance datasets problem that comes in nature in such biological networks. The classification performance is validates using a novel co-cross validation technique. Finally, the classification models power is verified utilizing three publicly datasets (functional genomics network, protein-protein interactions network, and co-expression network).



f1 f2 … … … … … … f14
G1
G2
.
.
0.5 0.1 … … …. … 01
0.25 0.2 … …. … … 0.8
… … …
… …
…
14 topological Features
Un-balanced dataset
Phenotypic
Non-Phenotypic
A biological network of any type (e.g. co-expression, PPIs, functional genomics, or disease specific gene association networks).
Balance datasets Using SMOTE
Phenotypic
Non-Phenotypic
Feed to different classification models
1. Decision Tree Bagger
2. Random Under-sample Boost
The classifiers are validated in co-cross validation scheme to provide a generalized performance.
Different classification models are built to classify breast cancer genes.

Fig. 7. A summary of the research procedure.

Fig. 7. shows the work summary that has been adopted in this research:

1. We start-off with a biological network as shown in upper left corner. Our contribution, we combine fourteen topological features instead of using a single

topological feature. These fourteen topological features are computed for every gene in the biological network.

2. We balance the datasets using a SMOTE sampling technique to enhance the classification performance.

3. We proposed a novel co-cross-validation technique that has not been used in the literature, to provide a generalized performance for the classification models.

4. Our contribution is extended in the form of a feature space that is fed to different classification models to discriminate breast cancer and non-breast cancer genes.

5. The classification models are built to predict breast cancer in the following scenarios: a) one features at a time, b) applying hill-climbing to find the most significant sub-set of topological features that increases the classification performance and c) applying Pearson's correlation coefficients features selection.

CHAPTER 4

METHODOLOGY

The complex biological networks could be represented as graphs. Such that, a node represents a gene, and the interactions between genes could be represented by links. In the literature, several ranking algorithms are developed to rank a certain node to its position in the graph. Some algorithms rank a certain gene depending on its direct neighbors. Others ranking algorithms rank a certain node globally (i.e. all nodes are considered). In this study, three different datasets (i.e. co-expression, PPIs, and functional genomic networks) are utilized to examine fourteen ranking algorithms. The computed scores for each gene are then fed into two different classification models to discriminate a disease of interest (i.e. breast cancer). Due to the high imbalance, SMOTE sampling technique is adopted.

4.1. Genomic Data Integration Methodology

This study has utilized Wu *et al.* publicly available integrated datasets of several types (i.e. gene expression, PPIs, Gene Ontology, and Domain Interaction). There is an urgent need to apply integration of multiple datasets of different types, such that knowledge from different resources is gained to construct a more reliable network. The problem of using a single dataset is that the pairwise networks do not necessarily provide fully 100% true positive interaction between two genes of proteins. For instance, the yeast two-hybrid approach might indicate that two proteins might physically interact without indicting that this interaction constitutes a part of a biological meaningful pathway in the living organism [47]. Moreover, there might be pairwise interaction datasets that may have a high false positive [26, 27]. In this section, we provide an overview of Wu *et al.* integration approach which could be summarized in the following three items.

1. Transforming pairwise interactions into probable functional interactions

The aforementioned problem of false positive could be resolve through the utilization of a pathway-based inference on the high-throughput functional data sets, such that the utilization of pathway databases will reveal a clean pairwise that are functionally related. Wu*et al.*[47] goal is to integrate both high-coverage, unreliable pairwise datasets with the low-coverage, highly reliable pathways to create a pathway informed data analysis for high-throughput analysis.

The first step to achieving this goal is that a functional interaction (FI) network is constructed that integrates a curated interaction from Reactome [74] and other pathway databases (i.e. Panther[75], CellMap[80], NCI-Nature[81], NCI-BioCarta[81], and KEGG[82]), with un-curated pairwise relationships that are collected from physical PPIs in human and model organisms, gene co-expression data, protein domain-domain interactions, protein generated from text mining, and Gene Ontology annotations. The Wu *et al.* approach makes use of the naïve Bayes classifier (NBC) to discriminate high-likelihood FIs from non-functional pairwise relationships.

2. Construction and training of a functional interaction classifier

The essential goal is to establish a network of protein functional relationships that shows functionally significant molecular events in cellular pathways. Most PPIs interaction databases are categorized as physically interactions and there is no clear proof that these interactions are involved in biomedical events. Furthermore, even other protein pairwise relationships have similar problems. An NBC [83] is implemented which scores the probability that two proteins are functionally related in a pathway event (i.e. if the probability greater or equal to 0.5).

Obviously the pathway databases are far richer than the binary PPIs and other pairwise relationships in terms of contents. Thus, Wu *et al.* has developed a relationship, called ‘functional interaction’, which defines a functional interaction as one in which two proteins are involved in the same biochemical reaction as an input, catalyst, activator, or inhibitor, or as two members of the same protein complex. The aforementioned definition is used to obtain the positive training sets from Reactome pathways for NBC. However, the NBC must be trained with both positive and negative samples. More importantly, the lack of knowledge that two proteins are not known to interact does not mean that these two protein are not interacting. Wu *et al.* adopted Zhange *et al.* [84] to create negative samples using random pairs selected from proteins in the Reactome FI set (in other words, using random pairs from the positive samples).

3. Merging the NBC with pathway data to create an extended FI network

An extended FI network is constructed with high protein and gene coverage through the combination of FIs predicted from the trained NBC with the annotated FIs extracted from the pathway databases (i.e. the pathway databases are: Reactome [74], Panther [75], CellMap [80], NCI-Nature [81], NCI-BioCarta [81], and KEGG[82]). For the purpose of increasing the coverage of the FI network, the interactions between human transcription factors and their targets from the TRED database are imported [85]. Furthermore, the result in an extended FI network consisting of 10,956 proteins and 209,988 FIs.

4.2. SMOTE: Synthetic Minority Over-sampling TEchnique

This study investigates highly imbalanced datasets (i.e. one category is far more than the other category). To enhance the classification performance the SMOTE sampling technique was used in this study, which was proposed by Chawla *et al.* [86].SMOTE has the advantage of combining in formed oversampling of the minority class with random under-sampling of the majority class. However, SMOTE is a sampling approach that could be used to overcome the problem of imbalance datasets. It over-samples the minority class by creating synthetic samples rather than over-sampling with replacements (i.e. that the positive (minority) samples are over-sampled with replacement to match the number of negative (majority) samples). However, it operates in "feature space" rather than "data space". Such that, the minority class is over-sampled by taking each minority class sample and generating synthetic samples along the line segments joining any or all of the k minority class nearest neighbors. The neighbors from the k nearest neighbors are randomly selected, and that is based on the amount of over-sampling needed.

SMOTE sampling technique computation should go as follows:

1. For each minority Sample:

    ◦ Find its k-nearest minority neighbors.

- Randomly select n of these neighbors.
- Randomly generate synthetic samples along the lines joining the minority sample and its n selected neighbors.

2. For instance:

- Select a minority sample.
- Take the difference between minority sample under consideration and its nearest neighbor.
- Multiply this difference by a random number between 0 and 1, and
- Add the resultant difference to minority sample under consideration (i.e. this will be the synthetic sample).

SMOTE was implemented in R using the ' DMwR' package [87]. It is worth mentioning that SMOTE was implemented with the following parameters: 1) the k-nearest neighbor is equal to 5 for every case scenario in this study, 2) to sample the breast cancer featuring the over sampling percentage = 400 and under sampling percentage = 125.

4.3. Biological Network Randomization

The biological network that has been utilized in this study was randomized based on the Maslov *et al.* [88] technique. The procedure rewires the connections of a given complex network, more importantly, conserving the connectiveness (keeping the degree of each vertex constant). Firstly, the algorithm uses a conventional procedure to perform the network randomization randomly choosing a pair of edges A→B and C→D. Secondly, the randomly chosen edges are rewired in the following manner, A becomes connected to D and C becomes connected to B. Furthermore, if it happens that one or both of these new links already exist in the network, then this step is canceled and a new pair of edges will be chosen. This restriction will result an out of a network that prevents the appearance of multiple edges connecting the same pair of nodes. Thirdly, a repetition of the previous rewiring procedure will provide a randomized network of the original network. Finally, the randomization of a large complex network will help to reveal significant topological patterns in a network. Such pattern might be over (or under) represented in the real network compared to the randomized network. Fig. 8. illustrate the biological network randomization methodology: 1) Select randomly pair of edges and rewire. Such that, the new re-wiring should not exist in the current network. 2) Results in a null model (randomized network) of a given complex network. Such that, the degree of direct neighbors is preserved. 3) The randomized model could be used to identify the over- under- represented pattern in the real network.

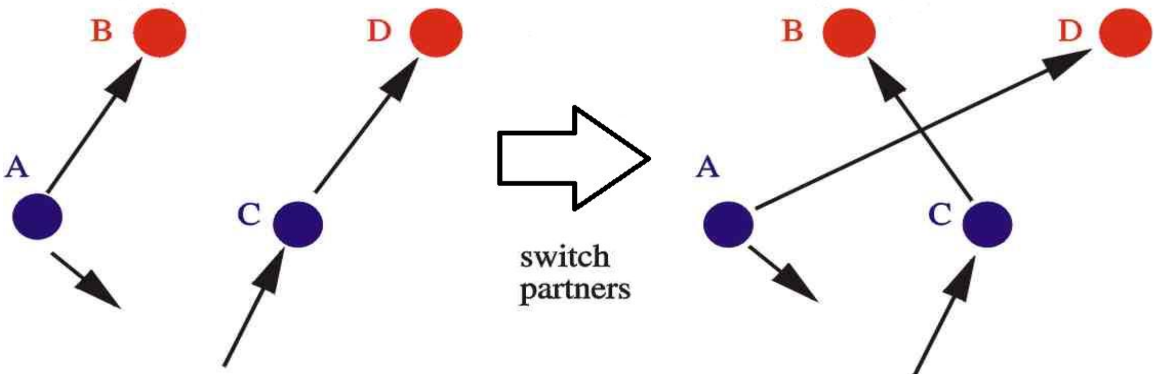


Fig. 8. Biological Network Randomization

4.4. Implemented Classifiers

In this study, there are mainly two stages to investigate whether a gene is related to a certain disease (i.e. breast cancer). Firstly, a graph theory is employed for feature extraction based on the network centrality measures and node raking as described earlier, and then SMOTE is applied to balance the minority class. Secondly, the performance of different classifiers for the prediction of different crucial diseases are analyzed (i.e. breast cancer). In particular, this study considers two classification models. First, to start off with, the decision tree bagger (DTB) [89-91] and RUSBoost [92] are both tested with a different growing number of trees. However, the used classifiers are implemented in Matlab [93]. Each model will be described briefly as follows.

4.4.1. Decision Trees Bagger

Bagging, which is also known as "bootstrap aggregation," is one of ensemble learning. The weak leaner decision tree will bag a specific dataset through generating many bootstrap replicas of the dataset, and the decision trees will be grown on these replicas. The bootstrap replica could be sampled randomly choosing from N observations out of N with replacement, where N is the dataset size. Furthermore, the average of the predictions from individual trees will give the predicted output of the trained ensemble.

The bagging technique is employed by training learners on resampled versions of the data which are often facilitated by bootstrapping observations, such that, N out of N observations with replacements are chosen for every new learner. Furthermore, the accuracy of the bagged trees could be improved, such that the tree in the ensemble could randomly choose the predictors for the decision splits.

Finally, the implemented minimal leaf sizes of the bagged trees are set to be 1. Thus, the trees grown (i.e. this study applied exhaustive growing trees starting from 50 up to 750, such that the increment pointer is 50 trees and then the best number of grown trees is chosen) with such a minimal leaf size tend to be very deep. More importantly, this setting is closer to be the optimal for the predictive power of an ensemble, but time consuming. Another important parameter is the number of predictors that are chosen randomly for every decision split. However, the parameter is determined as the square root of the number of the predictors.

4.4.2. RUSBoost

RUS is an acronym for Random Under sampling. It is meant to deal with imbalanced datasets effectively. Firstly, the algorithm is developed to take N observations to be the number of members in the class with the fewest members in the training data. These observations will be handled as the basic population for sampling. Furthermore, the majority classes are under sampled by taking only N observations of every class.

The RUSBoost ensemble is constructed with an optimal name-value pair. Called Ratio-To-Smallest, it provides a vector of K values, with each value representing the multiple of N to sample for the associated class. For instance, if the minority class contains only N = 100 members and the vector K = 2 3 4,then it could be interpreted that each weak learner has 200 members, 300 members and 400 members for the class 1, 2 and 3, respectively. Finally, if the Ratio-To-Smallest generates a value that is greater than the number of members in a specific class, then the RUSBoost samples the members with a replacement. Else and RUSBoost samples members without a replacement.

This study fixes one parameter of the RUSBoost that is the minimal leaf size, trees are set to be a default value = 1. On the other hands, the trees grown in this study are growing trees starting from 50 up to 750, such that the increment pointer is 50 trees. Then, among the grown set of trees, the one that is gives the best classification performance is selected.

CHAPTER 5

EXPERIMENTS

5.1. Experimental Data

A biologist generates massive datasets using high-throughput technologies (i.e. genomics, proteomics, or transcriptomics [65]). For instance, proteomics might provide for protein-protein interactions (PPIs) [25]through the yeast, the two-hybrid method or mass spectrometry. However, with PPIs, processing two or more genes are considered to be physically interacting. The datasets of PPIs could be viewed as a network, such that each gene in this network is represented using a node, and the interaction between genes represented using links. Using transcriptomics technologies, for instance, microarray could provide a gene expression profiling [25], such that, gene expression profiling measures the expression values of genes. Furthermore, this study utilizes three types of datasets. These datasets are BioGrid Homo sapiens PPIs [94], gene expression which made publicly available by Hedenfalk *et al.* [95], and an integrated functional protein network which made publicly available by Wu *et. al.* [96]. Prior to utilizing these networks, the common genes across these three datasets are obtained, such that, the common number of genes is 1,235.Based on this common number all three datasets are filtered to have the same size for all three networks.

5.1.1. Gene Expression Datasets

This study utilizes gene expression profiles of breast cancer for comparing and evaluating the classification and prediction purposes against the integrated genomic network of Wu *et al.* [96] and the BioGrid PPI. Firstly, the breast cancer gene expression profiles that has been utilized are obtained from Hedenfalk *et al.* [95], which consists originally of 3,226 genes and 22 gene expression profiles of patients that have breast cancer. The Hedenfalk *et al.* gene expression profiles are filtered according to the common genes (i.e. 1,235 genes).

The gene expression profiles are used to construct a co-expression network. The network construction is based on Ruan *et al.* [43]. The Ruan *et al.* approach is a rank-based network construction technique. Furthermore, the constructed co-expression network was randomized based on the Maslov *et al.* method. However, the resulted networks properties are shown in Table 4.

Table 4 Co-Expression Network Properties

Dataset	***Proteins***	***Interactions***	***Average Degree***	***Density***
Co-expression Nerwork	1,235	17,370	28.1	0.023
Randomized Co-expression Nerwork	1,235	17,370	28.1	0.023

5.1.2. Proteins-Proteins Interactions (PPIs) Networks

For further analysis, this study utilizes single PPIs of Homo sapiens to extract the topological features that have been described above and then fed into the classification models, i.e., these classification results will be reasoned against the Wu *et al.* integrated genomic network and the Hedenfalk *et al.* gene expression datasets classification results.

The datasets of PPIs are considered as a network. However, the nodes represent proteins and edges represent interactions among proteins. The PPI data of homo sapiens that were obtained from the BioGrid database has a version code of “BIOGRID-ORGANISM-Homo_sapiens-3.2.96”, which consists of 14,850 proteins and 169,166 interactions [94]. More importantly, the BioGrid PPIs is pre-processed, such that the proteins that are not found among the common genes (i.e. 1,235 genes), then are removed from the BioGrid PPIs network. In addition, the filtered PPIs network, which consists only of 1,235 genes was randomized based on the Maslov *et al.* method. However, the resulted network properties are shown in Table 5.

Table 5 PPIs Network Properties

Dataset	*Proteins*	*Interactions*	*Average Degree*	*Density*
PPIs Nerwork	1,235	12,345	19.9	0.016
Randomized PPIs Nerwork	1,235	10,471	16.9	0.014

5.1.3. Integrated Genomics Network

The third dataset of this utilized study is made publicly available by Wu *et al.*[96]. The constructed functional protein biological network has the advantages of binding both the high-coverage and unreliable pairwise datasets with the low-coverage, highly reliable pathways to make a pathway-informed data analysis system for high-throughput data analysis. However, the integration of both “high-coverage, unreliable pairwise” with “low-coverage, highly reliable pathways” datasets yield a functional interaction (FI) network. This FI network was constructed using curated interactions by experts and un-curated pairwise relationships (including protein-protein interactions in human and model organisms, genes co-expression data, protein domain-domain interactions, protein domain extracted from text mining and Gene Ontology annotations). More importantly, the FI network covers approximately 50% of the human entire set of proteins. Moreover, the proposed scoring system that is used by Wu *et al.*is a naïve Bayes classifier machine learning model. The score is an indication of the probability that a protein pairwise relationship reflects a function pathway event (refer to ch#4, section 4.1 for the integration methodology). Furthermore, Wu *et al.* FI network was filtered to contain genes that match the common genes across the other two utilized datasets (i.e. gene expression and PPIs). In addition, the filtered FI network, which consists of only 1,235 genes, was randomized based on the Maslov *et al.* method. However, the resulted networks properties are shown in Table 6.

Table 6 FI Network Properties

Dataset	***Proteins***	***Interactions***	***Average Degree***	***Density***
FI Nerwork	1,235	8,312	13.5	0.011
Randomized FI Nerwork	1,235	13,404	21.7	0.018

5.1.4. Breast Cancer Genes

Breast cancer genes and the ones that are related to these diseases are annotated and could be found in several public databases. This study has utilized several databases to extract the phenotypic genes that are related to breast cancer. Such that, the genetic association database (GAD) [2], The Mammalian Phenotype (MP) [97], The Human Phenotype Ontology [98] are made used of to extract genes that related to breast cancer.

Due to the severe unbalanced datasets, a sampling technique was utilized SMOTE to balance the data of these extracted features. This yielded consistently positive and negative samples across the three datasets (i.e. gene expression, BioGrid PPIs, and FI network). Such that, there 615 positive samples and 615 negative samples for the breast cancer.

5.2. Co-Cross Validation

The performance of each of the classifiers used in this thesis is generalized based on the use of cross-validation (CV) approach. CV is a well-known testing methodology that provides some insight on how well the system is able to classify using various combinations of testing and training data sets. The performance of phenotypic genes outcome prediction is measure by systematically excluding some data instances during the training process and testing the trained model using the excluded instances. However, we propose a novel technique based on the usual CV scheme, such that we have chosen a 5-folds CV scheme, where each time 4-folds are used to create another 5-fold to train and test certain models of the classifier, the best model is selected based on the averaged values of: F-Measure, AUC, and Geometric-Mean. The model that has been chosen is based on the averaged three values then it is tested using the 1-left-out fold from the original 5-folds. The process is repeated to cover all the dataset as a testing dataset. It is worth mentioning that we used only 5-fold co-cross validation because of the large number of samples. Hence, the division of the datasets will create sufficient number of sample for trains, and sufficient number of samples for testing.

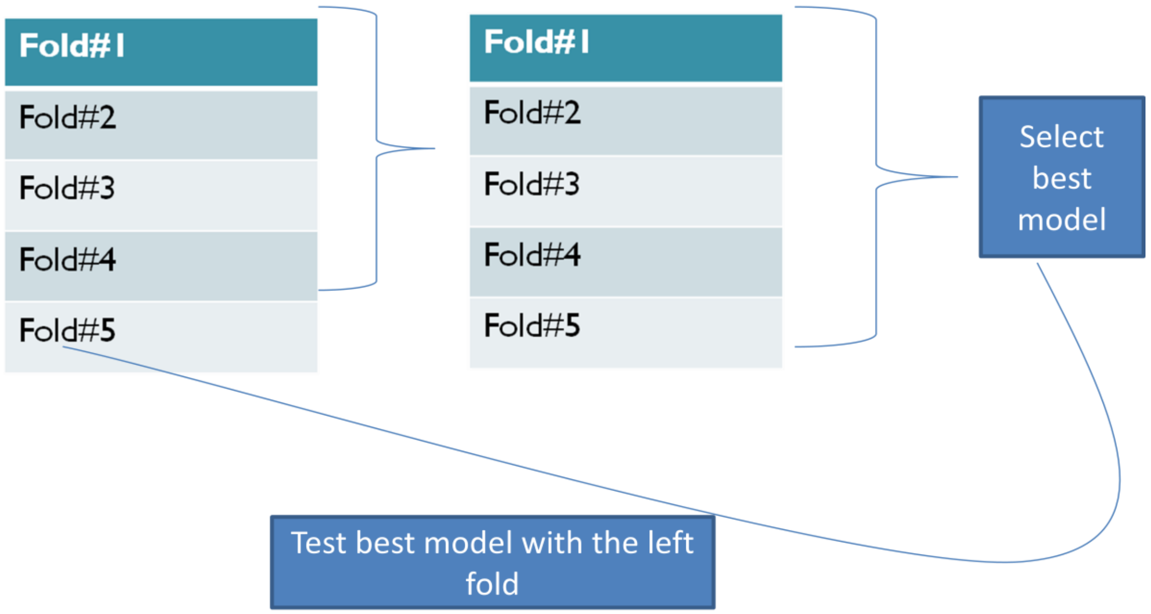


Fig. 9. Proposed co-cross validation

Fig. 9. shows the layout of our proposed co-cross validation scheme. The progress of the co-cross validation goes as follows:

1. At first stage, we create a 5-fold; out of these 5-folds we select 4-folds to create other 5-folds.
2. At the second stage, we used the created new 5-folds to train and test the system:
    a. Use 4-folds for training and the left fold for testing.
    b. Repeat this process 5-time, such that at each iteration new fold is used for testing (i.e. at first iteration fold one is used for testing and other four folds are used for training, at second iteration use the second fold for testing and other remaining folds for training, and so on).
    c. The classification performance is reported 5-time. Hence the average is taken.

d. The best model that is chosen from second phase is based on the averaged values of F-Measure, AUC, and Geometric-Mean.

3. The best model that has been chosen in the second phase will be used to test the classifier using the left fold from first stage.
4. This process is repeated 5-time at stage one (i.e. the classification performance is reported 5-time; hence the average values are taken).

5.3. Performance Evaluation

Clinical research often investigates the statistical relationship between symptoms (or test results) and the presence of a disease. When significant associations are found, it is useful to express the data in ways which are clinically relevant. Thus, to evaluate the performances of the different classifiers models, the following performance metrics are used in this thesis.

Let:

FP = false positives: Represents the number of wrongly predicted positive samples.

FN = false negatives: Represents the number of wrongly predicted negative samples.

TP = true positives: Represents the number of correctly predicted positive samples.

TN = true negatives: Represents the number of correctly predicted negative samples.

The aforementioned four counting (i.e. FP, FN, TP, and TN) could be illustrated using the confusion matrix in Table 7.

Table 7 Confusion Matrix

	Predict Number of Negative samples	Predict Number of Positive samples
Actual Number of Negative samples	TN	FP
Actual Number of Positive samples	FN	TP

The following are definitions of the metrics that have been computed to evaluate each classifier performance. They are accuracy, F-measure, area under ROC curve, and geometric mean. Additional definitions, such as sensitivity, specificity, and positive predictive value are included, because the other metrics that are used in this thesis are dependent on those.

1) *Accuracy* (ACC).Accuracy is one of the widely used performance metrics to evaluate a classifier. ACC is defined as all samples that are classified correctly over the total number of samples available (N).

$$\text{ACC} = \frac{TP + TN}{N} \tag{12}$$

2) *Sensitivity (SN).* Sensitivity refers to the proportion of cancerous genes which are correctly predicted as cancerous and total cancerous genes.

$$SN = \frac{TP}{TP + FN} \tag{13}$$

3) Specificity (SP). Specificity refers to the proportion of non-cancerous genes which are correctly eliminated and total non-cancerous gens.

$$SP = \frac{TN}{TN + FP} \tag{14}$$

4) *Positive predictive value* (PPV). Positive predictive value is the proportion of cancerous genes in the prediction which are correctly predicted as cancerous.

$$PPV = \frac{TP}{TP + FP} \tag{15}$$

5) F-measure (F). F-measure is the harmonic means of sensitivity and positive predictive value, which is defined as

$$\mathrm{F} = \frac{2 * SN * PPV}{SN + PPV} \tag{16}$$

6) *Area Under ROC plots*[99]: Performances of each of the classifiers were computed using AUC (area under receiver operating characteristic: ROC curves). The ROC plot space is a one-unit square, which the highest obtained AUC value that could be 1, meaning perfect ordering. However, in ROC graphs with a 0.5 AUC value, this represents random guessing, and values that smaller than 0.5 are not realistic as they can be negated by changing the decision criteria of the classifier. More importantly, the AUC value of a particular classifier could be viewed as the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

7) *Geometric mean (Gm).* Geometric mean has been introduced in [100] to overcome the problem that is associated with accuracy metric in unbalanced dataset learning.

$$Gm = \sqrt{SN * SP} \tag{17}$$

5.4. Hill-Climbing Feature Selection Technique

The classification performance can be enhanced through choosing a combination of important features that represent maximal separation between the classes [101]. In this thesis, we applied a hill-climbing feature selection algorithm. This algorithm initially selects the most influential features from the dataset and then repetitively adds feature to the selected feature set that positively improves the classification results or provides the least reduction in classification performances. The steps of the algorithm are as follows:

1. Initialize the select feature set $\boldsymbol{F}$ as an empty set.
2. Compute the classification performance for each of the individual features f paired with $\boldsymbol{F}$.
3. Select the feature $fbest$ for which the best classification performance is achieved and add the feature to selected feature set $\boldsymbol{F}$.
4. Repeat steps 2 and 3 until there is no remaining feature to be added to $\boldsymbol{F}$.

CHAPTER 6

RESULTS AND DISCUSSION

In this study, three datasets are utilized (i.e. co-expression network, BioGrid PPIs network, and FI network) to extract fourteen network topological features to predict breast cancer genes. These features are fed to two classifiers (i.e. Decision tree bagger (DTB) and RUSBoost) which are implemented based on growing trees of multiple of 50 trees staring from 50 trees up to 750 trees. Prior the classification and due to the severe imbalance between positive and negative samples, a SMOTE sampling technique is used, which resulted in a standard number of samples (i.e. 615 positive and negative samples). Furthermore, to validate the classification performance, we propose our co-cross validation, which in this study is implemented with a 5-fold, as described earlier. Co-cross validation technique is effective and efficient scheme to avoid the classifier over-fitting.

In this study, 1) we study the significance of the extracted 14 topological features from every individual biological networks. 2) Then, we study the most influential identified topological feature, and study its ability to reveal biological significant phenomena, in term of identifying the most central genes that are related to breast cancer.

6.1. Classification Performance Significance based on the Proposed Extracted Features

In this study, we construct a number of classification models to provide a wide view of the significant fourteen topological features and their ability to predict the phenotypic genes correctly and with high performance rates. More specifically, we find the classification performance using: 1) one feature at a time, 2) using sub-set of selected topological features using Hill-Climbing Feature Selection, and 3) using sub-set of selected topological features using Pearson correlation coefficients (PCCs) features selection.

6.1.1. Classification Performance using One Feature at a Time

In this section, we show the classification performance for each topological feature to predict the breast cancer and non-breast cancer genes using three complex biological networks, i.e., co-expression, BioGrid PPIs, and FI networks (results are shown in Table 8 - Table 10). Table 8 shows the classification results for the two classification models (i.e. DTB and RUSBoost) while adopting the extracted features from the co-expression network. We notice that within module Z-score (WMZ) topological feature provides the best classification results among the 14 topological features (i.e. DTB and RUSBoost achieve an accuracy rate of 81% (approx.), F-measure rate of 0.84, AUC rate of 0.81, and Geometric-Mean rate 0.80 while using only WMZ feature).

On the other hand, we observe that using node corness (NC) feature of co-expression network achieves the lowest classification performance among other topological features while classifying the datasets using DTB and RUSBoot to predict breast cancer (i.e. the accuracy rate of 55% (approx.), F-measure rate of 0.21, AUC rate of 0.55, and Geometric-mean of 0.34).

Table 8 Comparison of the Values of Accuracy, F-Measure, AUC, and Geometric-Mean using one feature at a time that extracted from Co-expression Network to Predict Breast cancer.

	DTB				**RUSBoost**			
Method	ACC	F	AUC	G-Mean	ACC	F	AUC	G-Mean
DC	0.7593	0.7923	0.7593	0.7403	0.7545	0.7874	0.7545	0.7361
BC	0.6008	0.6134	0.6008	0.5982	0.5756	0.5834	0.5756	0.5728
ClosC	0.6415	0.6666	0.6415	0.6363	0.5862	0.6004	0.5862	0.5824
KSM	0.6415	0.6564	0.6415	0.6396	0.6081	0.6171	0.6081	0.6068
EC	0.7650	0.7916	0.7650	0.7525	0.7602	0.7875	0.7602	0.7473
BCS	0.6423	0.6662	0.6423	0.6375	0.5854	0.5987	0.5854	0.5818
CC	0.6992	0.7274	0.6992	0.6895	0.6691	0.6926	0.6691	0.6630
FC	0.6984	0.7269	0.6984	0.6885	0.6683	0.6907	0.6683	0.6630
KSI	0.6333	0.6492	0.6333	0.6308	0.6106	0.6196	0.6106	0.6065
NC	0.5488	0.2117	0.5488	0.3419	0.5488	0.2117	0.5488	0.3419
WMZ	**0.8187**	**0.8429**	**0.8187**	**0.8040**	**0.8130**	**0.8365**	**0.8130**	**0.8001**
SC	0.6301	0.6467	0.6301	0.6276	0.6187	0.6343	0.6187	0.6165
PP	0.6423	0.6672	0.6423	0.6372	0.5862	0.6004	0.5862	0.5824
SH	0.6366	0.6549	0.6366	0.6335	0.6081	0.6245	0.6081	0.6058

To assess the power of the individual topological features, we use a different complex biological network, Namely, BioGrid PPIs network (i.e. the results are shown in Table 9). Comparison to the co-expression network features (Table 8), the BioGrid PPIs network features reveal same findings, i.e., DTB and RUSBoost classification models achieve highest classification performance using only within module z-score to predict breast cancer and non-breast cancer genes. WMZ BioGrid PPIs feature when is fed to DTB and RUSBoost is able to provide the highest performance rate, i.e., an accuracy rate of 87% (approx.), F-measure rate of 0.89, AUC rate of 0.88 and Geometric-Mean of 0.87.

On the other hand, we observe that Structural Holes (SH) of BioGrid PPIs network achieves the lowest classification performance among other topological feature for both DTB and RUSBoot classification models (i.e., the accuracy rate of 66% (approx.)). BioGrid PPIs features are able to classify better comparison to the features that are extracted from co-expression network.

Table 9 Comparison of the Values of Accuracy, F-Measure, AUC, and Geometric-Mean using one feature at a time that extracted from BioGrid PPIs Network to Predict Breast cancer.

	DTB				**RUSBoost**			
Method	ACC	F	AUC	G-Mean	ACC	F	AUC	G-Mean
DC	0.8667	0.8797	0.8667	0.8598	0.8618	0.8751	0.8618	0.8551
BC	0.7341	0.7397	0.7341	0.7336	0.7163	0.7134	0.7163	0.7154
ClosC	0.6618	0.6759	0.6618	0.6587	0.6472	0.6602	0.6472	0.6447
KSM	0.7057	0.7156	0.7057	0.7031	0.6959	0.7038	0.6959	0.6940
EC	0.7325	0.7492	0.7325	0.7283	0.7407	0.7496	0.7407	0.7390
BCS	0.7772	0.7996	0.7772	0.7683	0.7512	0.7730	0.7512	0.7436
CC	0.7740	0.7894	0.7740	0.7696	0.7675	0.7787	0.7675	0.7650
FC	0.7699	0.7852	0.7699	0.7657	0.7415	0.7579	0.7415	0.7373
KSI	0.7252	0.7349	0.7252	0.7235	0.7041	0.7109	0.7041	0.7029
NC	0.8650	0.8747	0.8650	0.8615	0.8650	0.8747	0.8650	0.8615
WMZ	**0.8748**	**0.8866**	**0.8748**	**0.8682**	**0.8772**	**0.8878**	**0.8772**	**0.8718**
SC	0.7821	0.7867	0.7821	0.7798	0.7715	0.7741	0.7715	0.7696
PP	0.6650	0.6830	0.6650	0.6608	0.6463	0.6608	0.6463	0.6436
SH	0.6602	0.6754	0.6602	0.6579	0.6106	0.6231	0.6106	0.6095

Furthermore, a third functional interaction (FI) network is utilized to extract the proposed fourteen topological features. FI fourteen topological features are fed one at a time for two different classification models (i.e. DTB and RUSBoost), the classification results to predict breast and non-breast genes are shown in Table 10.

It is consistently observed that within module z-score (WMZ) (through the three biological networks, i.e., co-expression, BioGrid PPIs, and FI networks) achieves the highest classification results when is fed to DTB and RUSBoost. WMZ topological feature of FI network when is fed to DTB and RUSBoost is able to provide an accuracy rate of 81% (approx.) for both classifiers. DTB and RUSBoost are also able to predict breast cancer genes using only WMZ with F-measure, AUC and Geometric-Mean with rates of 0.83, 0.81, and 0.79, respectively.

On the other hand, we observe that Betweenness centrality (BC) that was extracted from FI network achieves the lowest classification performance among other topological features for both DTB and RUSBoot classification models (i.e. the accuracy rate of ~64%, F-measure rate of 0.67, AUC rate of 0.64, and Geometric-mean of 0.64). Moreover, co-expression topological features remain to provide the lowest classification performance among the other complex biological networks (i.e. BioGrid PPIs and FI networks).

Table 10 Comparison of the Values of Accuracy, F-Measure, AUC, and Geometric-Mean using one feature at a time that extracted from Functional Interaction Network to Predict Breast cancer.

	DTB				RUSBoost			
Method	ACC	F	AUC	G-Mean	ACC	F	AUC	G-Mean
DC	0.7837	0.8152	0.7837	0.7647	0.7675	0.8016	0.7675	0.7471
BC	0.6463	0.6717	0.6463	0.6399	0.6431	0.6674	0.6431	0.6362
ClosC	0.6805	0.6904	0.6805	0.6790	0.6593	0.6677	0.6593	0.6573
KSM	0.6602	0.6799	0.6602	0.6571	0.6415	0.6582	0.6415	0.6388
EC	0.7772	0.8073	0.7772	0.7607	0.7593	0.7911	0.7593	0.7432
BCS	0.7984	0.8175	0.7984	0.7913	0.7683	0.7856	0.7683	0.7630
CC	0.6951	0.7350	0.6951	0.6782	0.6748	0.7112	0.6748	0.6628
FC	0.6878	0.7292	0.6878	0.6695	0.6740	0.7152	0.6740	0.6580
KSI	0.6561	0.6824	0.6561	0.6501	0.6341	0.6613	0.6341	0.6285
NC	0.7772	0.8153	0.7772	0.7494	0.7764	0.8152	0.7764	0.7473
WMZ	**0.8049**	**0.8327**	**0.8049**	**0.7874**	**0.8073**	**0.8320**	**0.8073**	**0.7937**
SC	0.6626	0.6758	0.6626	0.6607	0.6415	0.6498	0.6415	0.6403
PP	0.6789	0.6887	0.6789	0.6773	0.6569	0.6650	0.6569	0.6550
SH	0.6878	0.7122	0.6878	0.6820	0.6650	0.6866	0.6650	0.6611

In summary, '*within module z-score*' is the best feature that is identified using DTB and RUSBoost to classify breast cancer and non-breast cancer genes in all three complex biological networks (i.e. co-expression, BioGrid PPIs, and FI networks). On the other hand, DTB and RUSboost do not provide consistent view of the features that achieve the least classification performance among the extracted fourteen topological features. For instance, NC feature extracted from co-expression, SH extracted from BioGrid PPIs network, and BC extracted from FI network achieve the lowest classification results when are fed to DTB and RUSboost classification models to predict breast cancer.

6.1.2. Hill-Climbing Feature Selection Classification Performance

In this section, we show the classification performance adopting hill-climbing feature selection technique with two classification model (i.e. DTB and RUSBoost). Three complex biological networks are utilized in this study, i.e., co-expression, BioGrid PPIs, and FI networks to extract the aforementioned fourteen topological features. These fourteen features are fed to DTB or RUSBoost in hill-climbing mode (results shown in Fig. 11. - Fig. 16.) to predict breast cancer genes. DTB and RUSBoost classification models with hill-climbing feature selection will select the most significant sub-set of the topological features to increase the performance rate of breast cancer prediction. Hill-climbing approach would provide the optimal classification results, because it investigates all possible combination of the given topological features, unlike PCCs which chooses the uncorrelated-based features.

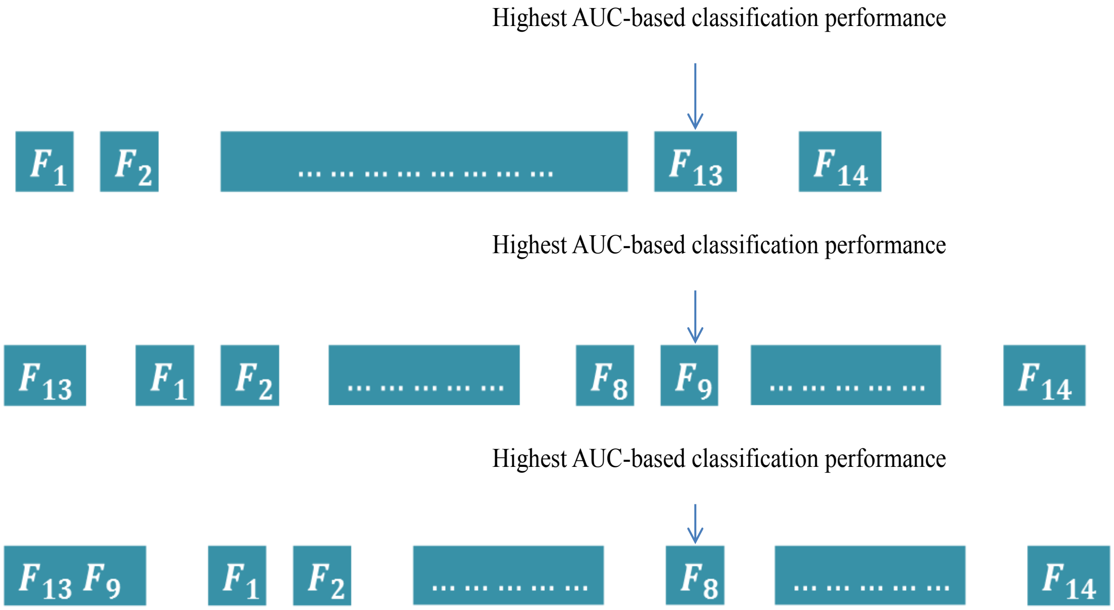

Highest AUC-based classification performance
F_1 F_2 … … … … … … … … F_{13} F_{14}
Highest AUC-based classification performance
F_{13} F_1 F_2 … … … … … F_8 F_9 … … … … … F_{14}
Highest AUC-based classification performance
F_{13} F_9 F_1 F_2 … … … … … F_8 … … … … … F_{14}

Fig. 10. Example of Hill-Climbing Approach

Fig. 10. shows an example of hill-climbing approach, i.e., at first iteration a classification model will compute AUC performance rate for every feature. In this example, we assume that f_{13}gives the highest AUC rate. Then, we fix it and try it with every other feature. Further, if f_{13} matched with f_9, we would get a better AUC performance rate. So, we fix f_{13} with f_9, and then we try to match them with every other feature. We continue to do that, until all combinations are tested and no feature is left to match with.

Table 12 shows performance rates of two classifiers to predict breast cancer while adopting co-expression network, BioGrid PPIs network, and FI network topological features individually for co-cross validation tests. The average success rate was greater than 80% for RUSBoost and 88% for DTB while using all 14 topological features that extracted from any of the individual biological network to predict breast cancer. However, when DTB and RUSBoost classifiers were trained with selected features network (i.e. selected important features are shown in Table 11), the classification performance improved while adopting the selected features (i.e. the results are shown in Table 13). Further, Fig. 11. - Fig. 16. show the trends of improvement of classification performance of the classifiers in terms of classification accuracy, F-measure, AUC, and Geometric-mean when using the hill-climbing approach.

Table 11 shows the topological features that are identified influential in discriminating breast cancer using hill climbing approach. For instance, DTB achieved an improvement by at least 90% compared to that of using all features in terms of classification accuracy (for both co-expression network and FI network individually). However, for BioGrid PPI network this performance improvement is 89% compared to that of using all the 14 topological features. It is noticeable that the huge improvement in classification accuracy was achieved using 9 topological features for co-expression network and 10 topological features for FI network respectively. Fig. 11. , Fig. 13. , and Fig. 15. show the trends of performance improvement visually for individual networks (i.e. co-expression, BioGrid PPIs, and FI, respectively) while using DTB. It is interesting to note that after using few selected features the performance starts declining for the both classifiers. This evidently guides us to use a subset of the 14 features in place of using all features to accurately classify the dataset.

To further assessing the performance of important topological features, we computed F-measure. As shown in Table 13, DTB using 9 topological features of co-expression to achieve F-measure value of 0.90, 7 topological features of BioGrid PPIs network to achieve F-measure value of 0.90, and 10 topological features of FI network to achieve F-measure value of 0.90 (i.e. Fig. 11. , Fig. 13. , and Fig. 15. show the trends improvement for co-expression, BioGrid PPIs, and FI networks topological features, respectively). On the other hand, RUSBoost classifier uses 8 features of co-expression network to achieve maximum F-measure value of 0.82, only 1 topological feature of BioGrid PPIs network to attain maximum F-measure of 0.89, and 7 topological features of FI network to achieve F-measure of 0.84 (i.e. the visual trends improvement are shown in Fig. 12. , Fig. 14. , and Fig. 16. , respectively).

A similar classification performance improvement trend is achieved when the performance of the classifiers is evaluated using area under the curve (AUC). For DTB, achieved mean AUC of 0.90 while using 9 topological features of co-expression network (Fig. 11.) , AUC of 0.89 while using 7 topological features of BioGrid PPIs network (Fig. 13.), and AUC of 0.90 while using 10 topological features of FI network (Fig. 15.). On the other hand, RUSBoost acquired a mean AUC of 0.82 using 8 features of co-expression network (Fig. 12.), the maximum classification performance (i.e., AUC value: 0.88) was achieved while using only WMZ of BioGrid PPIs network (Fig. 14.), for FI network 7 features is used to achieve AUC of 0.83 (Fig. 16.).

For further assessing the performance of important topological features, we compute one more statistical indicator which is Geometric-mean including the ones mentioned above, which are accuracy, F-measure and AUC. However, DTB using 9 topological of co-expression network to achieve the highest Geometric mean value of 0.90 (Fig. 11.), 7 topological features of BioGrid PPIs network to achieve maximum Geometric-mean value of 0.89 (Fig. 13.), For FI network the selected topological features are 10 to achieve a Geometric-mean value of 0.90 (Fig. 15.). Furthermore, RUSBoost acquired a Geometric-mean of 0.81 using 8 features of co-expression network (Fig. 12.), the maximum classification performance (i.e., AUC value: 0.87) was achieved while using only WMZ of BioGrid PPIs network (Fig. 14.), for FI network 7 features is used to achieve Geometric-mean of 0.83 (Fig. 16.).

A comparison between Table 12 and Table 13 reveals that utilizing a combination of selected features an enhanced classification performance was achieved for both DTB and RUSBoost classifiers while adopting any of the individual biological networks (i.e. co-expression network, BioGrid PPIs network, or FI network). DTB classifier showed at least 2% improvements and RUSBoost displayed at least 1% improvements while adopting selected set of topological features.

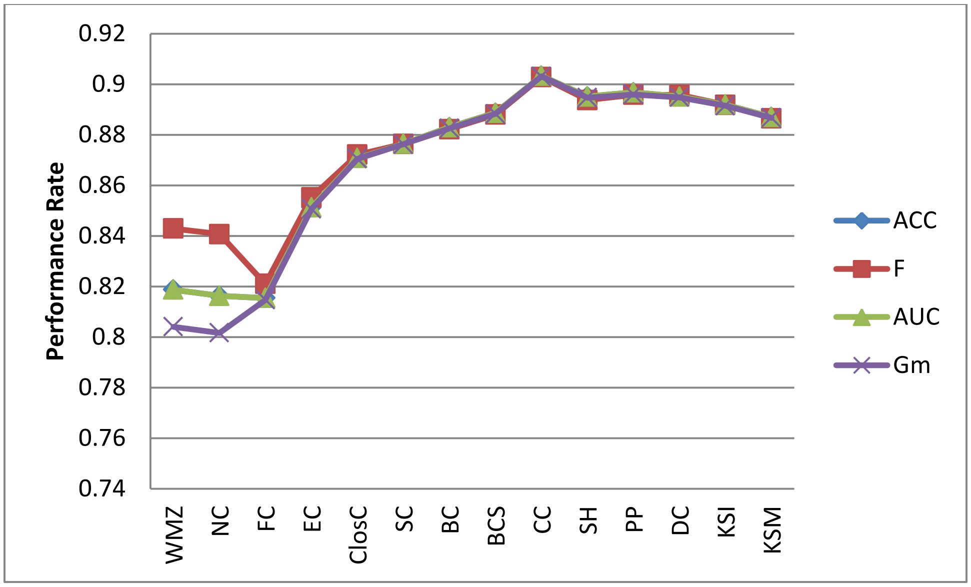


Fig. 11. Comparison of Values of Accuracy, F-Measure, AUC, and Geometric-Mean using Selected Features Based on Hill-Climbing to Predict Breast Cancer using DTB/Co-expression Network Topological Features

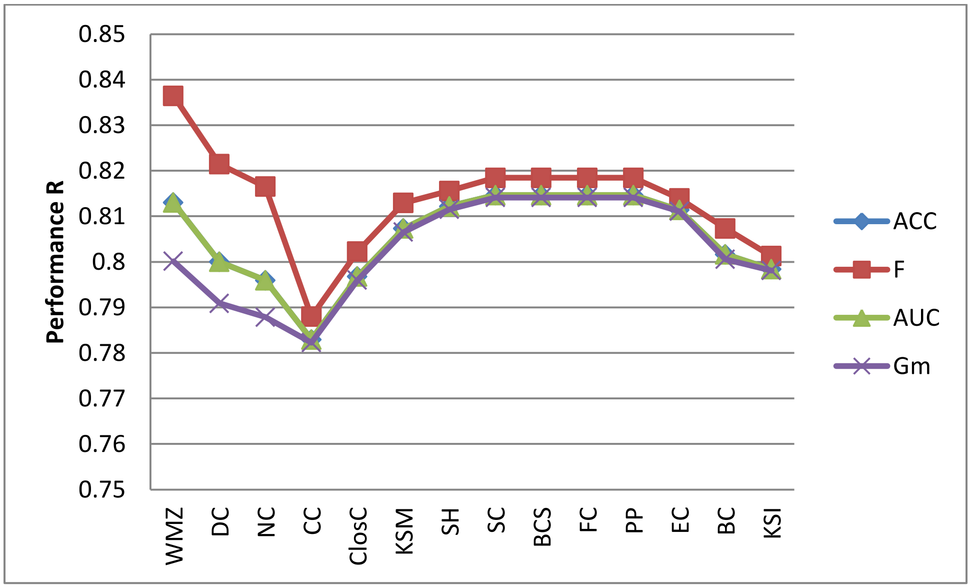


Fig. 12. Comparison of Values of Accuracy, F-Measure, AUC, and Geometric-Mean using Selected Features Based on Hill-Climbing to Predict Breast Cancer using RUSBoost/Co-expression Network Topological Features

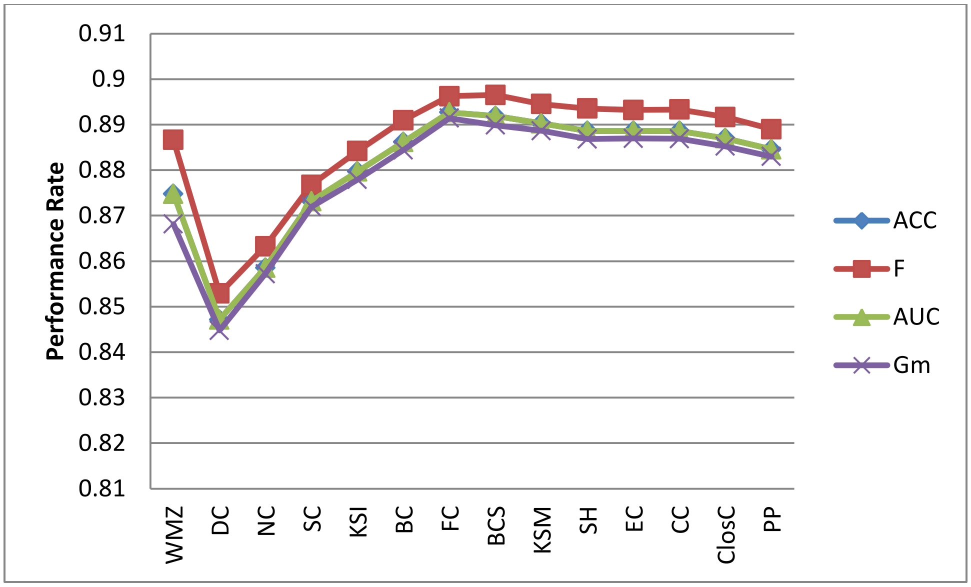


Fig. 13. Comparison of Values of Accuracy, F-Measure, AUC, and Geometric-Mean using Selected Features Based on Hill-Climbing to Predict Breast Cancer using DTB/BioGrid PPIs Network Topological Features

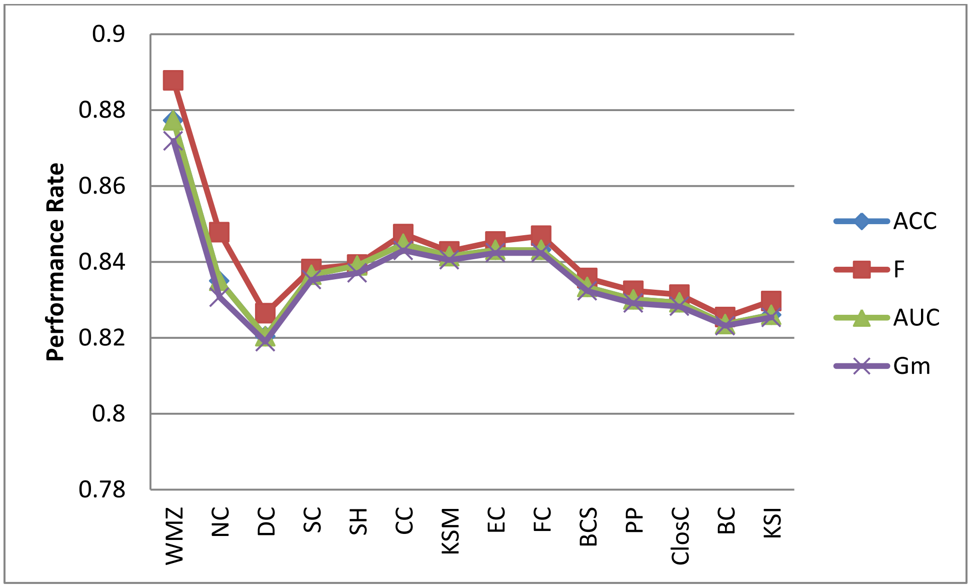


Fig. 14. Comparison of Values of Accuracy, F-Measure, AUC, and Geometric-Mean using Selected Features Based on Hill-Climbing to Predict Breast Cancer using RUSBoost/BioGrid PPIs Network Topological Features

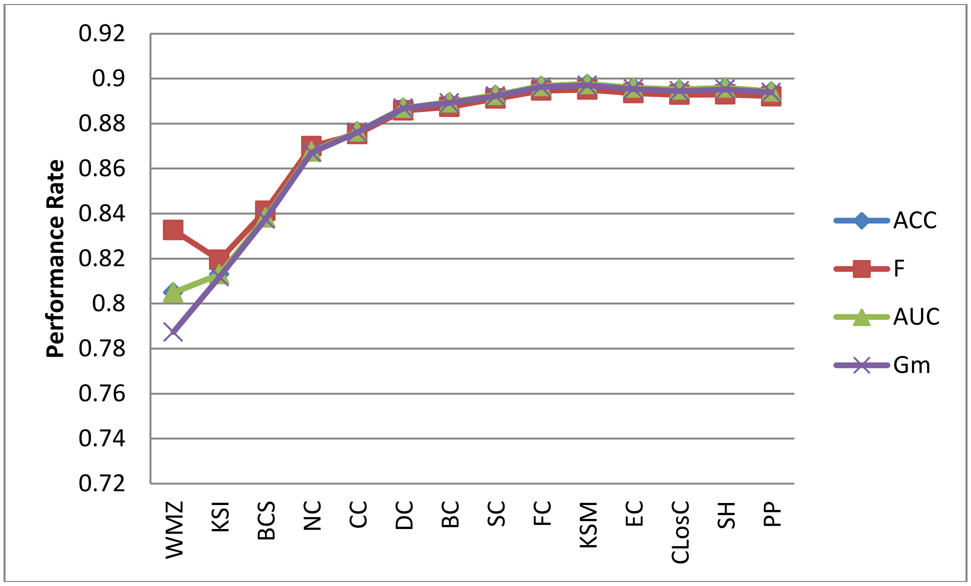


Fig. 15. Comparison of Values of Accuracy, F-Measure, AUC, and Geometric-Mean using Selected Features Based on Hill-Climbing to Predict Breast Cancer using DTB/FI Network Topological Features

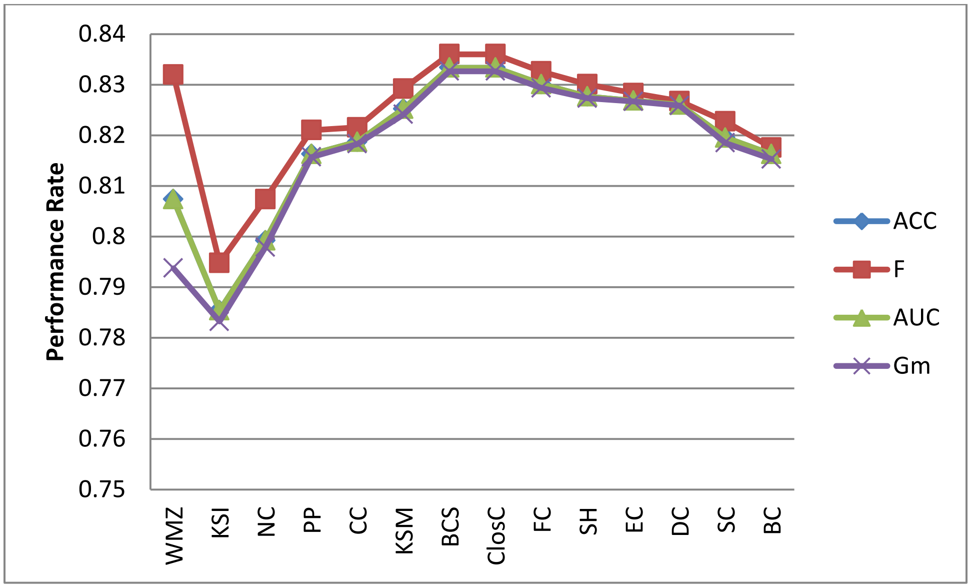


Fig. 16. Comparison of Values of Accuracy, F-Measure, AUC, and Geometric-Mean using Selected Features Based on Hill-Climbing to Predict Breast Cancer using RUSBoost/FI Network Topological Features

Table 11 Selected Important Features Based on DTB and RUSBoost Classification Models with Hill-Climbing Approach to Predict Breast Cancer From three Different Datasets

DTB	Co-Expression Network	WMZ	NC	FC	EC	ClosC	SC	BC	BCS	CC	
	BioGrid PPIs Network	WMZ	DC	NC	SC	KSI	BC	FC			
	FI Network	WMZ	KSI	BCS	NC	CC	DC	BC	SC	FC	KSM
RUSBoost	Co-Expression Network	WMZ	DC	NC	CC	ClosC	KSM	SH	SC		
	BioGrid PPIs Network	WMZ									
	FI Network	WMZ	KSI	NC	PP	CC	KSM	BCS			

Table 12 Classification Performance, Accuracy, F-Measure, AUC And Gm Of 5-Fold Co-Cross Validation Tests Using All 14 Topological Features

	DTB				**RUSBoost**			
Method	ACC	F	AUC	G-Mean	ACC	F	AUC	G-Mean
Co-Expression Network	0.887	0.887	0.887	0.887	0.798	0.801	0.798	0.798
BioGrid PPIs Network	0.885	0.889	0.885	0.883	0.826	0.830	0.826	0.825
FI Network	0.894	0.892	0.894	0.894	0.816	0.818	0.816	0.815

Table 13 The Maximum Classification Performance, Accuracy, F-Measure, AUC And Geometric-Mean Of 5-Fold Co-Cross Validation Tests Using Selected Topological Features

	DTB				RUSBoost			
Method	ACC	F	AUC	G-Mean	ACC	F	AUC	G-Mean
Co-Expression Network	0.903	0.903	0.903	0.903	0.815	0.818	0.815	0.814
BioGrid PPIs Network	0.893	0.896	0.893	0.891	0.877	0.888	0.877	0.872
FI Network	0.898	0.895	0.898	0.897	0.833	0.836	0.833	0.833

6.1.3. Pearson Correlation Coefficients (PCCs) Feature Selection

To provide some insight about the features, we analyze the correlation between each pair of the features. In this regard, we used Pearson correlation coefficients (PCCs) which is defined as $\rho\,(a, b) = \frac{E\,(ab)}{\sigma^2{}_a\sigma^2{}_b}$, where a and b are two zero-mean real valued random variables, E (ab) is the cross-correlation between a and b, $\sigma^2{}_a = E(a^2)$ and $\sigma^2{}_b = E(b^2)$ [102]. More importantly, we set a threshold of 0.6 to select the highly correlated features.

Fig. 17. shows the level of correlation between each pair of the features where the features were extracted from co-expression dataset. An analysis of the figure reveals that the feature ‘Within Module Z-core’ is highly correlated with ‘Degree Centrality’, ‘k-step markov centrality’, ‘eigenvector centrality’, ‘katz status index’, ‘Subgraph centrality’ and ‘Structural holes’. Further, the feature ‘Betweenness Centrality’ is found to be highly correlated with the features ‘Closeness Centrality’, ‘Bary-center score’ and ‘proximity prestige’. Similarly, ‘Clustering coefficient’ is highly correlated with ‘flow coefficient’. In summary, the features: ‘Within module Z-score’, ‘Betweennness centrality’, ‘Node Coreness’, and ‘Clustering coefficient’ are considered to be uncorrelated in the context of co-expression network topological features and could be considered for building the classification models.

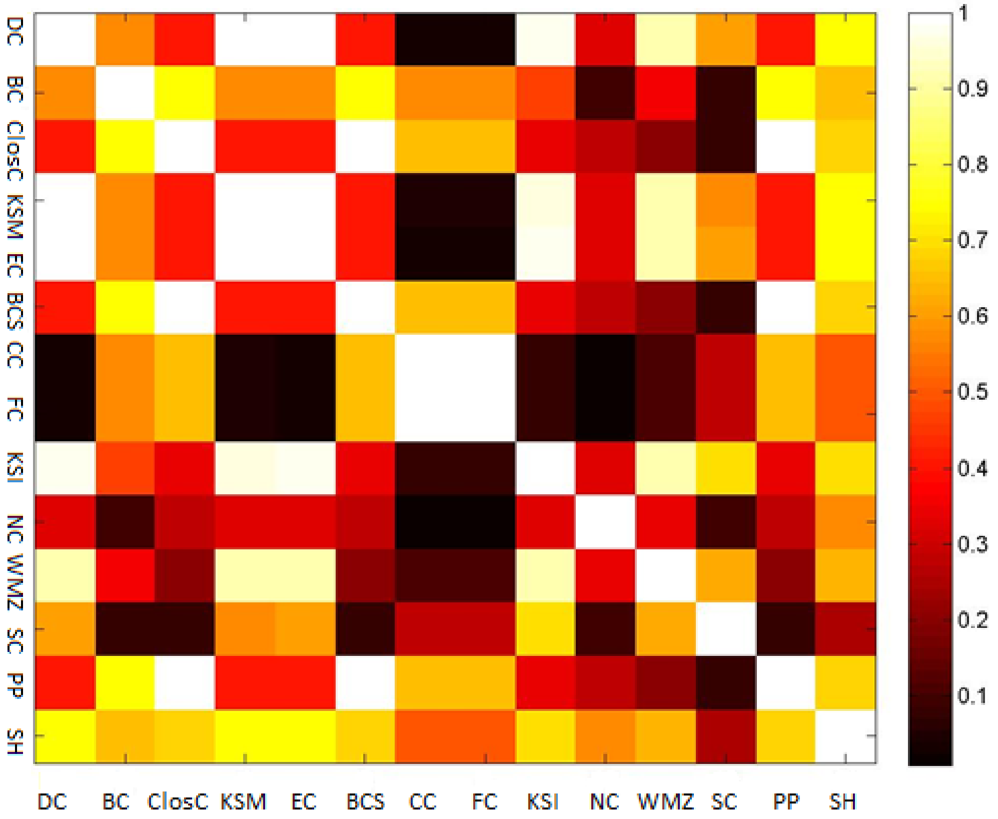


Fig. 17. Pearson correlation Heat-map of Extracted Features from Co-expression network.

Fig. 18. shows the level of correlation between each pair of the features where the features were extracted from BioGrid PPIs dataset. An analysis of the figure reveals that ‘Within Module Z-score’ is highly correlated with ‘Degree centrality’, ‘Betweenness Centrality’, ‘K-step markov centrality’, ‘Eigenvector Centrality’, ‘Katz Status Index’, ‘Node Coreness’, and ‘Subgraph Centrality’. Further, ‘Closeness Centrality’ is highly correlated with ‘Proximity Prestige’. Hence, ‘Within Module Z-score’, ‘Closeness centrality’, ‘Bary-center score centrality’, ‘Clustering coefficient’, ‘Flow coefficient’ and ‘Structural Holes’ are considered to be uncorrelated and could be used for building classification models.

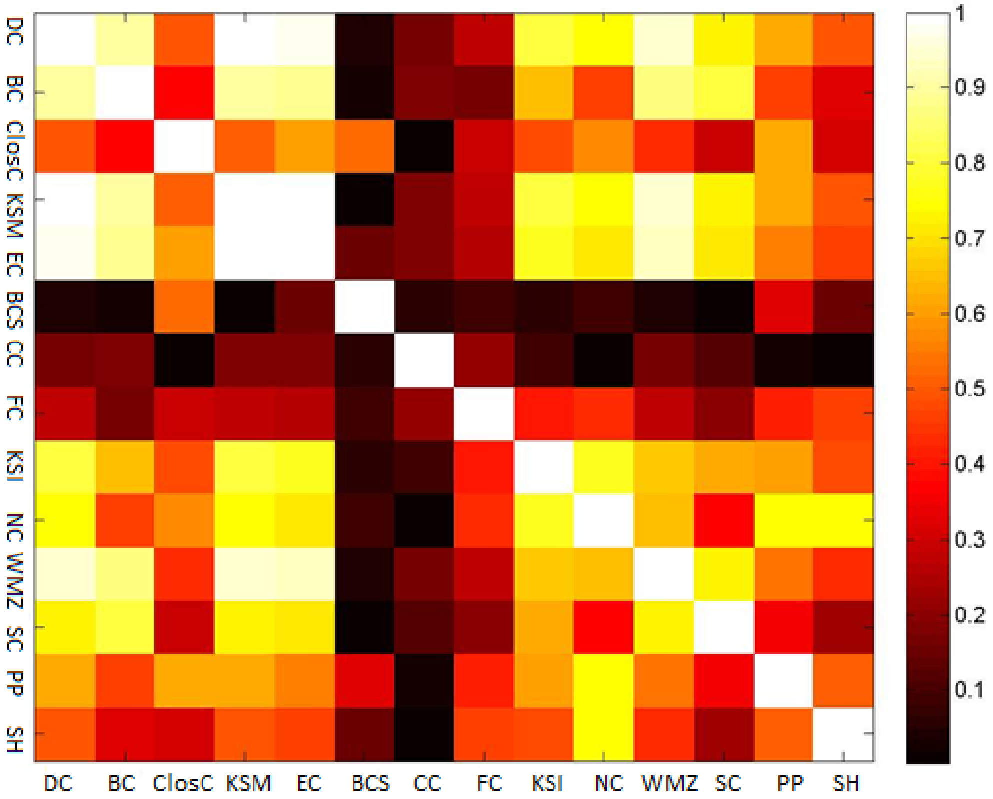


Fig. 18. Pearson correlation Heat-map of Extracted Features from BioGrid PPIs network.

Fig. 19. shows the level of correlation between each pair of the features where the features were extracted from functional interaction network. More specifically, ‘Within Module Z-score’ is highly correlated with ‘Degree centrality’, ‘K-step markov centrality’ and ‘Eigenvector centrality’. Further, ‘Betwenness Centrality’ is highly correlated with ‘Node Coreness’. In addition, ‘Closeness Centrality’ is highly correlated with ‘Bary-

Center Score'. Furthermore, 'Proximity Prestige' is highly correlated with 'Structural Holes'. Hence, 'Within Module Z-score', 'Betweennness centrality', 'Closeness centrality', 'Clustering coefficient', 'Flow coefficient', 'Katz Status Index', 'Subgraph centrality', and 'Proximity Prestige', are considered to be uncorrelated and could be used for building the classification models.

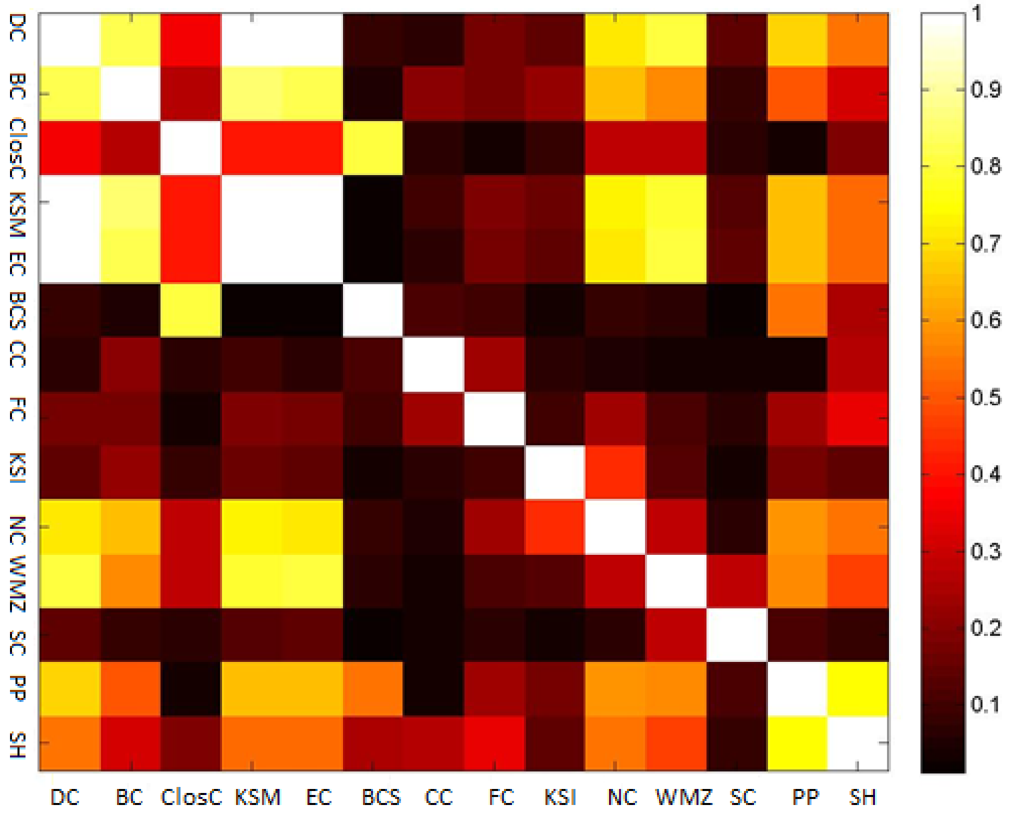


Fig. 19. Pearson correlation Heat-map of Extracted Features from Functional Interaction network.

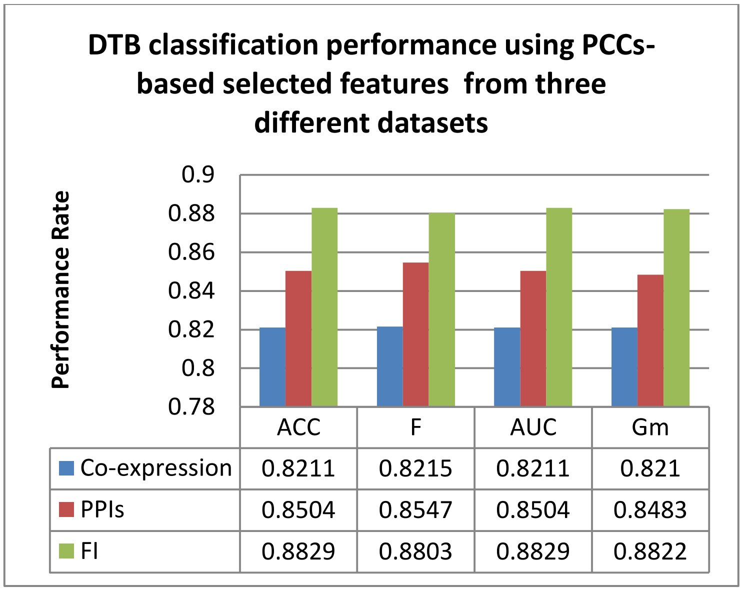
Table 14 Selected Uncorrelated Features Based on Pearson Correlation Coefficients to be Considered for Classification From Three Different Networks

Co-Expression Network	WMZ	BC	NC	CC				
BioGrid PPIs Network	WMZ	ClosC	BCS	CC	FC	SH		
FI Network	WMZ	BC	ClosC	CC	FC	KSI	SC	PP

Table 14 shows the potential selected uncorrelated features based on PCCs that could be adopted to build different classification models to predict breast cancer genes. Such that, four topological features, six topological features, and eight topological features are selected from co-expression network, BioGrid PPIs network, and FI network, respectively. These selected topological features from every individual biological network will be used to build two classification models (i.e. DTB and RUSBoost) to predict breast cancer and non-breast cancer genes.

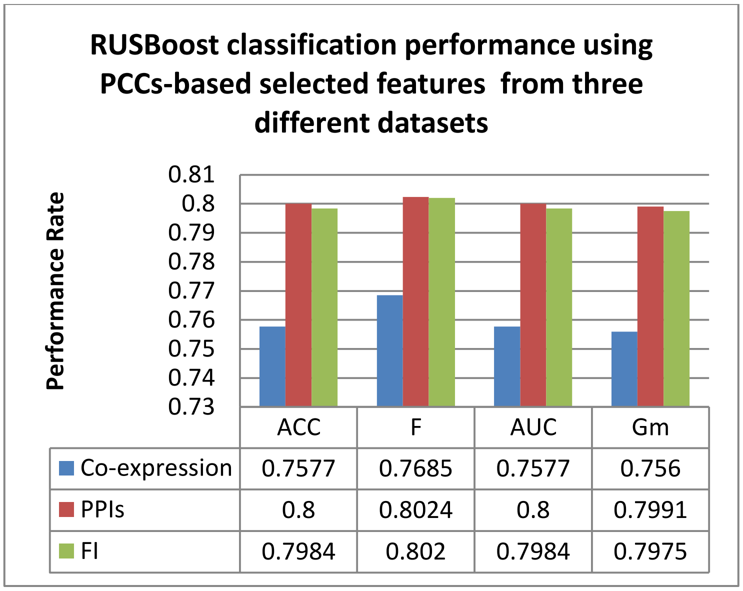
Fig. 20. shows the classification results of DTB, when is combined with FI network topological features we get the best classification performance, i.e., the accuracy rate of 88% (approx.), F-measure rate of 0.88, AUC rate of 0.88, and Geometric-mean of 0.88.

RUSBoost while adopting selected topological features based on PCCs either for BioGrid PPIs or FI networks, is able to achieve comparable results (Fig. 21.) . For instance, RUSBoost combined either with BioGrid PPIs network or FI network, is able to predict the breast cancer and non-breast cancer genes with an accuracy rate of 80% (approx.). Further, RUSBoost combined with BioGrid PPIs or FI network, is able to predict the breast cancer genes with F-measure, AUC, and Geometric-mean with rates of 0.80, 0.80, 0.80, respectively.



	ACC	F	AUC	Gm
🟦 Co-expression	0.8211	0.8215	0.8211	0.821
🟥 PPIs	0.8504	0.8547	0.8504	0.8483
🟩 FI	0.8829	0.8803	0.8829	0.8822

Fig. 20. Comparison of Values of Accuracy, F-Measure, AUC, and Geometric-Mean using Selected Features Based on PPCs to Predict Breast Cancer using DTB



	ACC	F	AUC	Gm
🟦 Co-expression	0.7577	0.7685	0.7577	0.756
🟥 PPIs	0.8	0.8024	0.8	0.7991
🟩 FI	0.7984	0.802	0.7984	0.7975

Fig. 21. Comparison of Values of Accuracy, F-Measure, AUC, and Geometric-Mean using Selected Features Based on PPCs to Predict Breast Cancer using RUSBoost

In summary, this study shows that the performance rates achieved for DTB combined with selected sub-set topological features from Functional Interaction network for breast cancer detection is more than satisfactory and, in addition, its automated nature also makes it suitable to be used in real clinical conditions. Further, the proposed approach has the feasibility of a real-time utilization, after offline training and testing. It is a beneficial to develop a system as a result of this study that may provide a robust feedback to the experts for the classification of breast cancer quickly and accurately based on the identified significant sub-set topological features based on PCCs or Hill-Climbing features selection.

6.2. Biological Significance of Selected Genes based on Within Module Z-score

The complex biological networks are powerful tools to define the interactions. In addition, these tools provide insight to predict the function and the evolution of the components (i.e. genes) in the biological networks. Obviously, the large size of the complex biological networks makes it difficult, if not possible, to extract the insightful knowledge from such biological networks. In this regard, we have used in this study 14 topological features with different classifiers that systematically allow identifying patterns and predicting phenotypic genes in complex biological networks. These 14 topological features could identify a crucial role for the genes in the biological networks, i.e., the most central within the network structure. Furthermore, we have analyzed every feature and we were able to identify that ‘Within Module Z-score’ is the best feature to predict the breast cancer genes. Other topological features are identified that might increase the prediction performance rate in hill-climbing feature selection approach. However, we focus our attention on WMZ and its ability to predict the breast cancer genes. WMZ is based on finding modules in the complex biological network, and then ranking the genes in that module based on the z-score. Every module could be arranged in a different manner, i.e., some modules could be centralized with one or few nodes connected to all the others. Other modules could be decentralized with all nodes having similar connectivies. More importantly, nodes with similar roles are predicted to have similar within module connectivity.

Due to the large size of complex biological networks, it is complicated to collect insightful knowledge out of hundreds or thousands of vertices and links. Though different centralities are computed but every centrality measure lack the ability to identify phenotypic genes with high accuracies. However, ‘within module z-score’ was proven to yield remarkable prediction results with two different classifiers. It is able to extract the central nodes based on a scale-specific context through the identification of the crucial modules (clusters). Hence, the identified modules summarize the connections and information of network structure, and that contain components (i.e. genes) of high similarities. Moreover, it is likely that the set of connected nodes in a certain module reveal similar roles or functions (i.e. related to breast cancer).

6.2.1. Parameter to Select Hubs Genes based on WMZ

We define hub genes based on 'within module z-score', i.e., the nodes with a $\text{within module z} - \text{score} \geq 2.5$ are catecorized as module hubs genes, else we classify nodes with a $\text{within module z} - \text{scores} < 2.5$ as non-hubs genes. Hence, the non-hubs genes are discarded and only the hubs genes are further studied because they tend to be important and they might influence other genes. Moreover, Fig. 22. shows the selected hubs genes out of co-expression network and their corresponding z-score, i.e., 36 genes are selected. Among the selected genes from co-expression network based on within module z-score, there are 18 genes that already known to be related to breast cancer (i.e., POLB, RAB25, FOXM1, TNFRSF1B, AKT1, IRF1, BCL2A1, LSP1, PCM1, RGS16, TRIM24, SP2, VDR, CKS2, DDX17, RAB11FIP1, APLP2, MAD2L1).

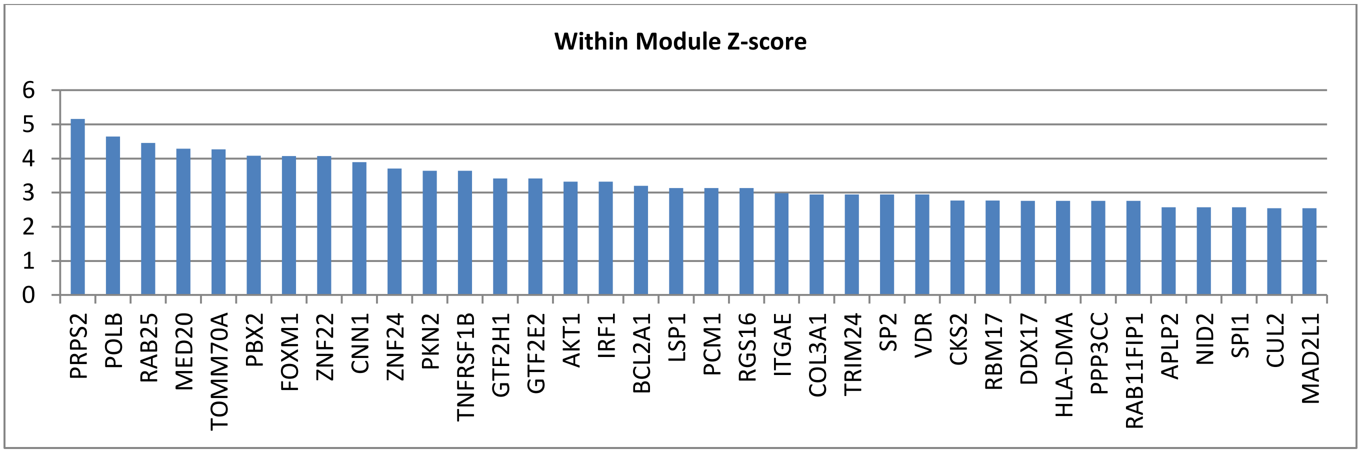


Fig. 22. Within module Z-score of Selected Genes from Co-expression Network.

Fig. 23. shows the selected hubs-genes out of BioGrid PPIs network and their corresponding z-score, i.e., 24 hubs-genes are selected and among the selected genes from BioGrid PPIs based on within module z-score, there are 23 genes that already known to be cancerous except for one genes which is CUL2.

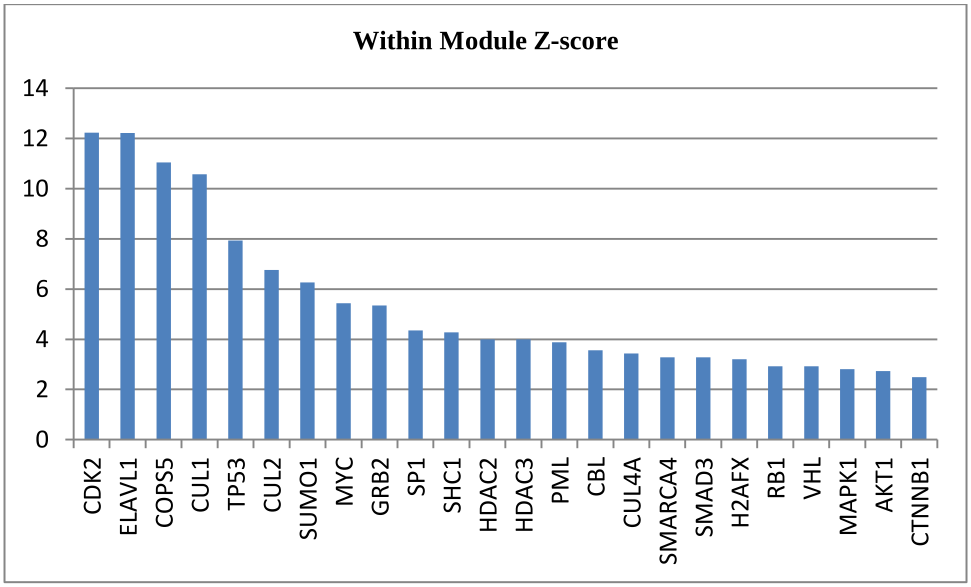


Fig. 23. Within module Z-score of Selected Genes from BioGrid PPIs Network.

Furthermore, Fig. 24. shows the selected genes out of FI network, i.e., 42 hubs-genes are selected and among the selected genes from FI network based on within module z-score, there are 36 genes that already known to be linked to breast cancer except for 6 genes which are CPSF1, CSNK1A1, DHX38, GTF2F2, GTF2B, NCBP2.

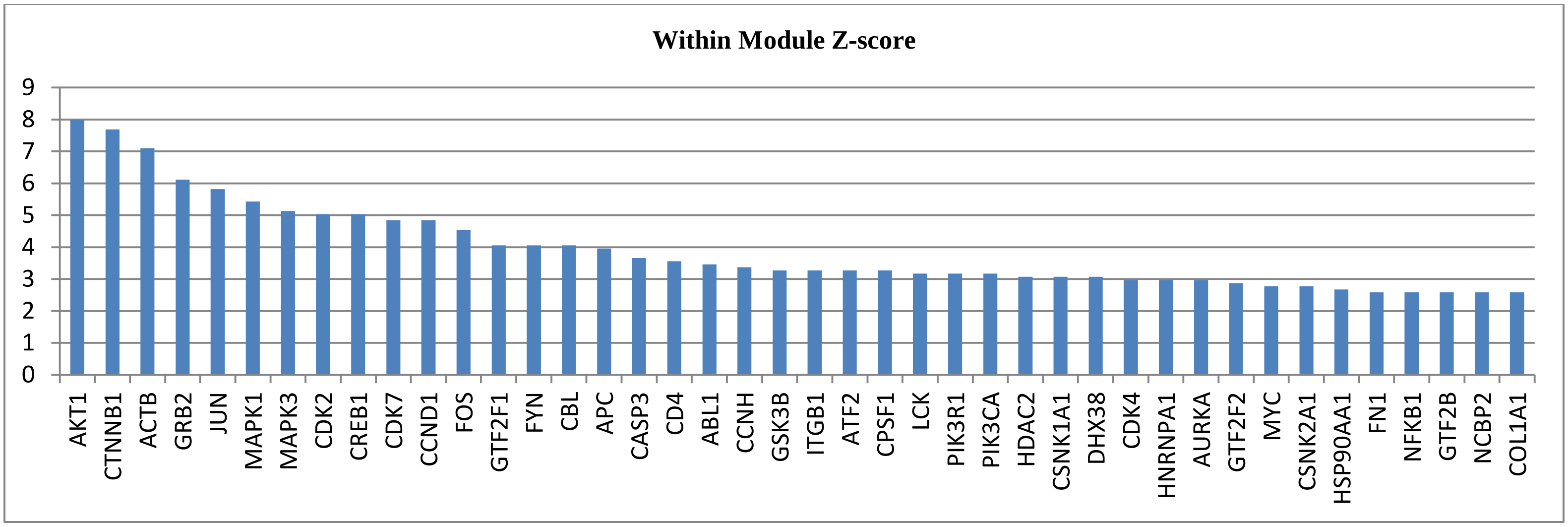


Fig. 24. Within module Z-score of Selected Genes from FI Network.

Finally, the 'Within module z-score'-based selected genes across the three individual biological networks (i.e. co-expression network, BioGrid PPIs network, FI network) shows remarkable results. Interestingly, majority of the identified hubs-genes are related to breast genes. Nonetheless, the selected hubs-genes, i.e., they are unknown whether they are related to breast cancer or not could be further studied and identify the functionally related similarities.

6.2.2. David's Tools Analysis

To provide some insight about the selected genes, we analyze the selected genes using Database for Annotation, Visualization and Integrated Discovery (DAVID) [103] to reveal which hubs-genes are similar to which hub-genes. Specifically, we use gene similarity search tool, this tool incorporates association of set of annotation terms for any given gene. Genes tend to participate in similar biological processes, if they are share set of those terms. The gene similarity search algorithm, adopts kappa statistics to quantitatively measure the degree of the agreement how genes share the similar annotation terms (i.e.

kappa values ranges from 0 to 1). Intuitively, the higher the kappa value, the stronger the relationship among the genes. More importantly, we set a threshold of 0.6 to select the highly similar genes.

Table 15 shows the selected genes from co-expression network and their corresponding functionally related genes with kappa statistics to detect any functional relationship. For instance, ‘ZNF22’ gene shows a high functional relationship with ‘SP2’ gene (i.e. the kappa value is 0.64). This indicates that the two genes ‘ZNF22’ and ‘SP2’ in considerable agreement and highly participate in the same biological processes, more so than by random event. Hence, ‘ZNF22’ is highly involved to breast cancer.

Table 15 Selected Genes from Co-expression Network based on Within Module Z-score and their Corresponding Functionally Related Genes.

Symbol	Entrez Gene ID	Breast Cancerous	Functionally Related Genes (Kappa rate)
PRPS2	5634	Unknown	None
POLB	5423	Known	None
RAB25	57111	Known	None
MED20	9477	Unknown	PBX2 (0.40), GTF2E2 (0.39), FOXM1(0.37), SP2(0.36),
TOMM 70A	9868	Unknown	None
PBX2	5089	Unknown	SPI1(0.47), IRF1(0.45),GTF2H1(0.42), MED20(0.40)
FOXM1	2305	Known	SP2(0.40), IRF1(0.38), MED20 (0.37), SPI1 (0.37), GTF2E2 (0.37)
ZNF22	7570	Unknown	SP2 (0.64), ZNF24 (0.62)
CNN1	1264	Unknown	None
ZNF24	7572	Unknown	SP2(0.67), ZNF22(0.62), SPI1(0.41), TRIM24(0.36)
PKN2	5586	Unknown	None
TNFRS F1B	7133	Known	None
GTF2H 1	2965	Unknown	GTF2E2(0.45), PBX2(0.42)

GTF2E2	2961	Unknown	GTF2H1 (0.45), MED20(0.39), FOXM1(0.37)
AKT1	207	Known	None
IRF1	3659	Known	PBX2(0.45), SPI1(0.44), FOXM1(0.38)
BCL2A1	597	Known	None
LSP1	4046	Known	None
PCM1	5108	Known	None
RGS16	6004	Known	None
ITGAE	3682	Unknown	None
COL3A1	1281	Unknown	None
TRIM24	8805	Known	VDR(0.38), ZNF24(0.36), SPI1(0.36)
SP2	6668	Known	ZNF24(0.67), ZNF22(0.64), FOXM1(0.40), MED20(0.36)
VDR	7421	Known	TRIM24(0.38)
CKS2	1164	Known	None
RBM17	84991	Unknown	None
DDX17	10521	Known	None
HLA-DMA	3108	Unknown	None
PPP3CC	5533	Unknown	None
RAB11FIP1	80223	Known	None
APLP2	334	Known	None
NID2	22795	Unknown	None
SPI1	6688	Unknown	PBX2(0.47), IRF1(0.44),ZNF24(0.41), FOXM1(0.37), TRIM24(0.36)
CUL2	8453	Unknown	None
MAD2L1	4085	Known	None

Table 16 shows the functionally related genes that are selected based on ‘within module z-score’ from BioGrid PPIs network. Interestingly, all selected genes are known to be involved in breast cancer except for ‘CUL2’ genes. ‘CUL2’ is highly functionally related to ‘CUL4A’ and ‘CUL1’ genes with kappa values of 0.85 and 0.66, respectively. These achieved kappa statistics values indicate strongly that ‘CUL2’ gene is involved in breast cancer.

Table 16 Selected Genes from BioGrid PPIs Network based on Within Module Z-score and their Corresponding Functionally Related Genes.

Symbol	Entrez Gene ID	Breast Cancerous	Functionally Related Genes (Kappa Rate)
CDK2	1017	Known	None
ELAVL1	1994	Known	None
COPS5	10987	Known	None
CUL1	8454	Known	CUL2(0.66), CUL4A(0.66)
TP53	7157	Known	PML(0.45), SMAD3(0.35)
CUL2	8453	Unknown	CUL4A(0.85), CUL1(0.66)
SUMO1	7341	Known	None
MYC	4609	Known	None
GRB2	2885	Known	SHC1(0.52)
SP1	6667	Known	None
SHC1	6464	Known	GRB2(0.52)
HDAC2	3066	Known	HDAC3(0.49), SMARCA4(0.40), RB1(0.37)
HDAC3	8841	Known	HDAC2(0.49)
PML	5371	Known	TP53(0.45)
CBL	867	Known	None
CUL4A	8451	Known	CUL2(0.85), CUL1(0.66)
SMARCA4	6597	Known	HDAC2(0.40)
SMAD3	4088	Known	TP53(0.35)
H2AFX	3014	Known	None
RB1	5925	Known	HDAC2(0.37)
VHL	7428	Known	None
MAPK1	5594	Known	AKT1(0.37)
AKT1	207	Known	MAPK1(0.37)
CTNNB1	1499	Known	None

Table 17 shows the functionally related genes that are selected based on ‘within module z-score’ from Functional Interaction network. There are 6 unknown genes whether they are involved in the breast cancer. However, among these genes is ‘GTF2F2’ gene which shows a high similarities with ‘GTF2F1’ (i.e. kappa value is 0.69). This is a strong indication that ‘GTF2F2’ share major biological processes with ‘GTF2F1’, hence, ‘GTF2F2’ could be strongly involved in breast cancer disease.

Table 17 Selected Genes from FI Network based on Within Module Z-score and their Corresponding Functionally Related Genes.

Symbol	Entrez Gene ID	Breast Cancerous	Functionally Related Genes (Kappa rate)
AKT1	207	Known	MAPK1(0.37)
CTNNB1	1499	Known	None
ACTB	60	Known	None
GRB2	2885	Known	MAPK3(0.41), PIK3R1(0.36)
JUN	3725	Known	FOS(0.53)
MAPK1	5594	Known	MAPK3(0.65), AKT1(0.37), GSK3B(0.36)
MAPK3	5595	Known	MAPK1(0.65), CSNK2A1(0.43), GRB2(0.41), CDK4(0.39), PIK3CA(0.39), CDK2(0.36)
CDK2	1017	Known	CDK4(0.66), CDK7(0.49), CSNK2A1(0.46), CSNK1A1(0.41), MAPK3(0.36)
CREB1	1385	Known	FOS(0.36)
CDK7	1022	Known	CCNH(0.61), GTF2F2(0.52), CDK2(0.49), CDK4(0.47), GTF2F1(0.42), CSNK2A1(0.39), CSNK1A1(0.37)
CCND1	595	Known	None
FOS	2353	Known	JUN(0.53), CREB1(0.36)
GTF2F1	2962	Known	GTF2F2 (0.69), GTF2B(0.52), CCNH(0.49), CDK7(0.42)
FYN	2534	Known	LCK(0.41), ABL1(0.38),
CBL	867	Known	None
APC	324	Known	None
CASP3	836	Known	None
CD4	920	Known	None
ABL1	25	Known	FYN(0.38), LCK(0.38)
CCNH	902	Known	CDK7(0.61), GTF2F2(0.52), GTF2F1(0.49), GTF2B(0.42)
GSK3B	2932	Known	CSNK1A1(0.39), MAPK1(0.63)
ITGB1	3688	Known	None
ATF2	1386	Known	None
CPSF1	29894	Unknown	None

LCK	3932	Known	FYN(0.41), ABL1(0.38)
PIK3R1	5295	Known	PIK3CA(0.47), GRB2(0.36)
PIK3CA	5290	Known	PIK3R1(0.47), MAPK3(0.39)
HDAC2	3066	Known	None
CSNK1A1	1452	Unknown	CSNK2A1(0.55), CDK4(0.45), CDK2(0.41), GSK3B(0.39), CDK7(0.37)
DHX38	9785	Unknown	GTF2F2(0.43)
CDK4	1019	Known	CDK2(0.66), CDK7(0.47), CSNK2A1(0.46), CSNK1A1(0.45), MAPK3(0.39)
HNRNPA1	3178	Known	NCBP2(0.50)
AURKA	6790	Known	None
GTF2F2	2963	Unknown	GTF2F1(0.69), CDK7(0.52), CCNH (0.52), GTF2B(0.44), DHX38(0.43)
MYC	4609	Known	None
CSNK2A1	1457	Known	CSNK1A1(0.55), CDK4(0.46), CDK2(0.46), MAPK3(0.43), CDK7(0.39)
HSP90AA1	3320	Known	None
FN1	2335	Known	None
NFKB1	4790	Known	None
GTF2B	2959	Unknown	GTF2F1(0.52), GTF2F2(0.44), CCNH(0.42)
NCBP2	22916	Unknown	HNRNPA1(0.50)
COL1A1	1277	Known	None

Finally, we shed the light specifically on three genes which are ‘ZNF22’ from co-expression network, ‘CUL2’ from BioGrid PPIs network, and ‘GTF2F2’ from FI network. These three genes showed high similarities in term of kappa values. However, the other selected unknown genes are highly potential candidates to be involved in breast cancer disease.

6.3. Impact of our Work

In our work, we are able to identify significant sub-set of topological features that could help biologists and researchers to adopt those biological features to study the structure and functions of biological networks. Specifically, we found that 'within module z-score' and 'node coreess' topological features have the ability to accurately identifying the central genes (i.e. hub-genes that are highly connected to other genes) when are fed to two classifiers. Hub-genes are potential candidates that must be studied because they are essential to influence other neighboring genes.

Furthermore, 'within module z-score' is utilized to perform in-depth analysis. Particularly, we set a threshold that any gene has a within module z-score equal or greater 2.5, then it is selected to be a hub-gene. Interestingly, we observed that the majority of selected hubs-genes are involved in breast cancer disease. Obviously, the selected hubs-genes that are unknown whether they are involved in breast cancer disease, they are potential candidates to be related to breast cancer. However, a gene similarity analysis revealed that ‘ZNF22’ from co-expression network, ‘CUL2’ from BioGrid PPIs network, and ‘GTF2F2’ from FI network provide high statistical significance that are participating in biological processes that typically breast cancerous genes are involve in. Hence, our analysis suggests that ‘ZNF22’, ‘CUL2’, and ‘GTF2F2’ are involved in breast cancer diseases.

CHAPTER 7

CONCLUSION AND FUTURE WORK

The bioinformatics techniques that are studied for disease-gene association are evaluated using three publicly available datasets (i.e. co-expression, BioGrid PPIs, and FI networks). These complex biological networks were utilized to apply analysis for breast cancer prediction. The work of this study could be divided into three folds. Firstly, the graph theory discipline which has been a significant part to studying complex biological networks. It has been applied to different complex biological networks to extract hidden knowledge (i.e. we have used fourteen ranking algorithms for every network). These topological properties are then fed into different classification models as features. Secondly, the phenotype that has been studied (i.e. breast cancer) is associated with a few number of samples. Thus, the negative classes are far more in number than in the positive classes. For this reason, a SMOTE sampling mechanism is adopted to overcome the problem of unbalanced datasets. Thirdly, at this stage, different classification models, such as: decision tree bagger and RUSBoost, are implemented to discriminate between the positive classes and negative classes. Moreover, the power of network topology analysis is limited, as we are dealing with the biological networks as a statics perspective, which in fact they are highly dynamic systems. Thus, the used biological networks have been randomized and the classification results are generated once again in the attempt to reveal some hidden patterns that are uncovered from the real networks. However, it remains the study of dynamic biological network over-times a bottleneck which will be investigated more in our future research.

7.1. Limitations

In this research, we were able to build a systematic approach that could association genes to diseases. However, there were some limitations of this study because of its nature. First, only one balanced sampled datasets are generated from each network topological features using SOMTE. We could get more insights about the classification performance significance by generating more number of sampled datasets using SMOTE or even using different sampling techniques. Second, the classification performance could be improved by extracting more robust topological features. Third, different classifiers need to be tested that might provide a better classification performance, and different weak learners could be ensemble.

Appendices

Appendix A

In this appendix, we include a number of results that were omitted from the main thesis write-up. In section A, we show the classification performance using the fourteen extracted topological features from the randomized co-expression, BioGrid PPIs, and FI networks to predict breast cancer genes.

A Randomized Co-expression, BioGrid PPIs, and FI networks Features Classification Performance

In this section, we show the classification performance using all fourteen topological features at once to predict the breast cancer and non-breast cancer genes. More importantly, these fourteen topological features are extracted from the randomized models of the three complex biological networks (i.e. co-expression network, BioGrid PPIs network, and FI network). The adopted randomization scheme is illustrated in chapter 4, section 4.3. Then, these fourteen topological features that have been extracted from the randomized biological networks are fed to two different classification models i.e., DTB and RUSBoost (results shown in Table 18) to predict breast cancer genes.

A.1 Classification Performance using Fourteen Topological Features

Cancer phenotype participates in dysregulation of several pathways controlling essential cell processes [104] (e.g. breast cancer phenotype). In this case, the complex biological networks are affected by having altered pathways (in other words, the interactions are altered), which will cause a genetic mutation dependent on the environmental context. Moreover, the potential of topological properties analysis is not fully utilized since we are dealing solely with statics networks. Rather the complex biological networks tend to be highly dynamics, as new interactions and genes are involved in the biological processes. Hence, the utilization of the randomized network provides the advantages of covering the hidden patterns that might be under (or over) represented in the real network. Therefore, a randomized version of the three biological networks (i.e. co-expression network, BioGrid PPIs, and FI) is generated based on the Maslov *et al.* technique. The fourteen topological properties are then applied to extract the features. Prior classification, the raw data are balanced using SMOTE (i.e. the positive and negative samples are equal to 615). Next the classification models are applied to predict the breast cancer and non-breast cancer genes using the extracted features from the randomized network. The results are shown in Table 18.

The randomized BioGrid PPIs network topological features when are adopted for any classification model (i.e. DTB or RUSBoost); tend to provide highest classification performance. More specifically, DTB combined with randomized BioGrid PPIs network topological features provide best classification results. Further, DTB/ randomized BioGrid PPIs topological features is able to predict breast cancer and non-breast cancer genes with an accuracy rate of 87% (approx.), and could predict breast cancer genes correctly with F-measure rate of 0.88, AUC rate of 0.87, and geometric-mean rate of 0.87 (Table 18) Secondly, the RUSBoost achieved the second best classification performance using the randomized PPIs network topological features, where the accuracy, F-measure, AUC, and geometric-means are 83% (approx.), 0.84, 0.83, and 0.83, respectively.

Moreover, the randomized co-expression network topological features, when used for training and testing the DTB and RUSBoost, the classification trends are slightly better than using the randomized FI network (Table 18).

Table 18 Comparison of the Values of Accuracy, F-Measure, AUC, and Geometric-Mean using All Fourteen Features to Predict Breast cancer.

	DTB				**RUSBOOST**			
METHOD	ACC	F	AUC	G-MEAN	ACC	F	AUC	G-MEAN
Co-Expression Network	0.871	0.87	0.871	0.871	0.775	0.777	0.775	0.774
BioGrid PPIs Network	0.872	0.878	0.872	0.87	0.832	0.835	0.832	0.832
FI Network	0.867	0.869	0.867	0.867	0.77	0.773	0.77	0.769

Appendix B

In this appendix, we include a number of results that were omitted from the main thesis write-up. In Section B, we provide classification results following the same procedure applied in the thesis on four disease specific gene association networks.

B Disease Specific Gene Association Networks

In our thesis work, we have extracted fourteen features from publicly gene expression profiles, protein-protein interaction networks, and functional interaction network. Then, we used a sampling technique SMOTE to overcome the problem of imbalanced datasets that are inherited in nature in such types of datasets. After that, the imbalanced datasets are used to train the classification models (i.e. decision tree bagger and RUSBoost) to investigate the phenotype-gene association in breast cancer. More importantly, in this section we validate our approach that has been proposed in this thesis using more four disease specific gene association networks to reveal breast cancer genes patterns.

In analogy with our previous procedure, we have built also two classification models that could be used for discriminating breast cancer genes in more four datasets which are disease specific gene association networks. The classification results are shown in Table 19.

Table 19 Disease Specific Networks Classification Performance To Predict Breast Cancer Genes

CLASSIFIER		BASAL NETWORK	HER2 NETWORK	LUMA NETWORK	LUMB NETWORK
DTB	ACC	0.931 ± 0.005	0.927 ± 0.006	0.923 ± 0.008	0.929 ± 0.013
	F	0.935 ± 0.005	0.931 ± 0.006	0.927 ± 0.008	0.929 ± 0.013
	AUC	0.931 ± 0.005	0.927 ± 0.006	0.923 ± 0.008	0.929 ± 0.013
	G-Mean	0.929 ± 0.006	0.925 ± 0.006	0.921 ± 0.008	0.929 ± 0.013
RUSBOOST	ACC	0.902 ± 0.005	0.892 ± 0.004	0.881 ± 0.008	0.875 ± 0.01
	F	0.903 ± 0.005	0.894 ± 0.003	0.883 ± 0.008	0.876 ± 0.01
	AUC	0.902 ± 0.005	0.892 ± 0.004	0.881 ± 0.008	0.875 ± 0.01
	G-Mean	0.902 ± 0.005	0.891 ± 0.004	0.881 ± 0.008	0.875 ± 0.01

The proposed methodology in this thesis showed great potential when it was tested using three publicly available datasets: co-expression network, protein-protein interactions network, and function interaction network. Furthermore, we validate the strength of our proposed fourteen topological features along with the selected classification models performance in predicting the phenotype-gene association using features extracted from different four disease specific networks. Table 19 reports the performance measures which are defined earlier, and more importantly the table show a comparison among three classification models (i.e. decision tree bagger and RUSBoost). Obviously, decision bagger provides best classification results which adopt SMOTE sampling technique and using the topological features extracted from disease specific gene association networks. That is an indication of the strength of the generated rules of the decision tree bagger during the training phase. Moreover, that is also shows the efficiency of the used co-crossed validation that is proposed in this study to obtain high and efficient classification performance rates during the testing phase.

The classifiers are evaluated with accuracy rate, along with other performance measures (i.e. F-measure, AUC, geometric mean) to provide a better indication of the classification performance. However, the combination of Basal gene specific network and decision tree bagger achieved the highest classification performance in term of accuracy, F-measure, AUC, and geometric mean. Such that, the accuracy rate is 0.931, this shows how the decision tree bagger capable to correctly classify cancerous and non-cancerous breast cancer genes. On the other hand, geometric mean which is an efficient metrics for unbalanced datasets showed a proportion of 0.929 which is an indication of capability of decision tree bagger to classify breast cancerous genes correctly.

References

1. McKusick, V.A., *Mendelian Inheritance in Man and its online version, OMIM.* American journal of human genetics, 2007. **80**(4): p. 588.
2. Becker, K.G., et al., *The genetic association database.* Nature genetics, 2004. **36**(5): p. 431-432.
3. Özgür, A., et al., *Identifying gene-disease associations using centrality on a literature mined gene-interaction network.* Bioinformatics, 2008. **24**(13): p. i277-i285.
4. Guan, Y., et al., *Functional genomics complements quantitative genetics in identifying disease-gene associations.* PLoS computational biology, 2010. **6**(11): p. e1000991.
5. Ackert-Bicknell, C.L., et al., *Mouse BMD quantitative trait loci show improved concordance with human genome-wide association loci when recalculated on a new, common mouse genetic map.* Journal of Bone and Mineral Research, 2010. **25**(8): p. 1808-1820.
6. Freeman, L.C., *Centrality in social networks conceptual clarification.* Social networks, 1979. **1**(3): p. 215-239.
7. Junker, B.H. and F. Schreiber, *Analysis of biological networks*. Vol. 1. 2008: Wiley Online Library.
8. Cormen, T.H., et al., *Introduction to algorithms*. 2001: MIT press.
9. Huber, W., et al., *Graphs in molecular biology.* BMC bioinformatics, 2007. **8**(Suppl 6): p. S8.

10. Dijkstra, E.W., *A note on two problems in connexion with graphs.* Numerische mathematik, 1959. **1**(1): p. 269-271.

11. Floyd, R.W., *Algorithm 97: shortest path.* Communications of the ACM, 1962. **5**(6): p. 345.

12. Pavlopoulos, G.A., et al., *Using graph theory to analyze biological networks.* BioData mining, 2011. **4**(1): p. 10.

13. Freeman, L.C., *A set of measures of centrality based on betweenness.* Sociometry, 1977: p. 35-41.

14. Wuchty, S. and P.F. Stadler, *Centers of complex networks.* Journal of Theoretical Biology, 2003. **223**(1): p. 45-53.

15. Estrada, E. and J.A. Rodriguez-Velazquez, *Subgraph centrality in complex networks.* Physical Review E, 2005. **71**(5): p. 056103.

16. Bonacich, P., *Power and centrality: A family of measures.* American journal of sociology, 1987: p. 1170-1182.

17. Levy, S.F. and M.L. Siegal, *Network hubs buffer environmental variation in Saccharomyces cerevisiae.* PLoS biology, 2008. **6**(11): p. e264.

18. Zotenko, E., et al., *Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality.* PLoS computational biology, 2008. **4**(8): p. e1000140.

19. White, S. and P. Smyth. *Algorithms for estimating relative importance in networks*. in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003. ACM.

20. Honey, C.J., et al., *Network structure of cerebral cortex shapes functional connectivity on multiple time scales.* Proceedings of the National Academy of Sciences, 2007. **104**(24): p. 10240-10245.

21. Hagmann, P., et al., *Mapping the structural core of human cerebral cortex.* PLoS biology, 2008. **6**(7): p. e159.

22. Guimera, R. and L.A.N. Amaral, *Functional cartography of complex metabolic networks.* Nature, 2005. **433**(7028): p. 895-900.

23. Cho, H., et al., *Social networks, communication styles, and learning performance in a CSCL community.* Computers & Education, 2007. **49**(2): p. 309-329.

24. Wilson, G. and W. Banzhaf. *Discovery of email communication networks from the Enron corpus with a genetic algorithm using social network analysis*. in *Evolutionary Computation, 2009. CEC'09. IEEE Congress on*. 2009. IEEE.

25. !!! INVALID CITATION !!!

26. Von Mering, C., et al., *Comparative assessment of large-scale data sets of protein–protein interactions.* Nature, 2002. **417**(6887): p. 399-403.

27. Formstecher, E., et al., *Protein interaction mapping: a Drosophila case study.* Genome research, 2005. **15**(3): p. 376-384.

28. Furey, T.S., et al., *Support vector machine classification and validation of cancer tissue samples using microarray expression data.* Bioinformatics, 2000. **16**(10): p. 906-914.

29. Ramaswamy, S., et al., *Multiclass cancer diagnosis using tumor gene expression signatures.* Proceedings of the National Academy of Sciences, 2001. **98**(26): p. 15149-15154.

30. Lee, Y. and C.-K. Lee, *Classification of multiple cancer types by multicategory support vector machines using gene expression data.* Bioinformatics, 2003. **19**(9): p. 1132-1139.

31. Lee, Y. and Y. Lin. *Grace wahba. Multicategory support vector machines*. in *Proceedings of the 33rd Symposium on the Interface*. 2001.

32. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* science, 1999. **286**(5439): p. 531-537.

33. Khan, J., et al., *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.* Nature medicine, 2001. **7**(6): p. 673-679.

34. Wei, J.S., et al., *Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma.* Cancer research, 2004. **64**(19): p. 6883-6891.

35. Li, X., et al., *Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling.* Nucleic Acids Research, 2004. **32**(9): p. 2685-2694.

36. Alon, U., et al., *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, 1999. **96**(12): p. 6745-6750.

37. Dettling, M. and P. Bühlmann, *Boosting for tumor classification with gene expression data.* Bioinformatics, 2003. **19**(9): p. 1061-1069.

38. Guthke, R., et al., *Gene expression data mining for functional genomics using fuzzy technology*, in *Advances in Computational Intelligence and Learning*. 2002, Springer. p. 475-487.

39. Tao, H., et al., *Functional genomics: expression analysis ofEscherichia coli growing on minimal and rich media.* Journal of Bacteriology, 1999. **181**(20): p. 6425-6440.

40. Pirooznia, M., et al., *A comparative study of different machine learning methods on microarray gene expression data.* BMC genomics, 2008. **9**(Suppl 1): p. S13.

41. Hall, M.A., *Correlation-based feature selection for machine learning*, 1999, The University of Waikato.

42. Lee, H.K., et al., *Coexpression analysis of human genes across many microarray data sets.* Genome research, 2004. **14**(6): p. 1085-1094.

43. Ruan, J., A.K. Dean, and W. Zhang, *A general co-expression network-based approach to gene expression analysis: comparison and applications.* BMC systems biology, 2010. **4**(1): p. 8.

44. de Matos Simoes, R. and F. Emmert-Streib, *Bagging statistical network inference from large-scale gene expression data.* PLoS One, 2012. **7**(3): p. e33624.

45. Zhu, Y., X. Shen, and W. Pan, *Network-based support vector machine for classification of microarray samples.* BMC bioinformatics, 2009. **10**(Suppl 1): p. S21.

46. Wu, G. and L. Stein, *A network module-based method for identifying cancer prognostic signatures.* Genome biology, 2012. **13**(12): p. R112.

47. Wu, G., X. Feng, and L. Stein, *Research A human functional protein interaction network and its application to cancer data analysis.* 2010.

48. Van De Vijver, M.J., et al., *A gene-expression signature as a predictor of survival in breast cancer.* New England Journal of Medicine, 2002. **347**(25): p. 1999-2009.

49. van Dongen, S.M., *Graph clustering by flow simulation.* 2000.

50. Zhang, W., et al., *Network-based Survival Analysis Reveals Subnetwork Signatures for Predicting Outcomes of Ovarian Cancer Treatment.* PLoS computational biology, 2013. **9**(3): p. e1002975.

51. Goh, K.-I., et al., *The human disease network.* Proceedings of the National Academy of Sciences, 2007. **104**(21): p. 8685-8690.

52. Navlakha, S. and C. Kingsford, *The power of protein interaction networks for associating genes with diseases.* Bioinformatics, 2010. **26**(8): p. 1057-1063.

53. Oti, M., et al., *Predicting disease genes using protein–protein interactions.* Journal of medical genetics, 2006. **43**(8): p. 691-698.

54. Navlakha, S., R. Rastogi, and N. Shrivastava. *Graph summarization with bounded error*. in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 2008. ACM.

55. Van Dongen, S., *Graph clustering via a discrete uncoupling process.* SIAM Journal on Matrix Analysis and Applications, 2008. **30**(1): p. 121-141.

56. Navlakha, S., et al. *Finding biologically accurate clusterings in hierarchical tree decompositions using the variation of information*. in *Research in Computational Molecular Biology*. 2009. Springer.

57. Köhler, S., et al., *Walking the interactome for prioritization of candidate disease genes.* The American Journal of Human Genetics, 2008. **82**(4): p. 949-958.

58. Vanunu, O. and R. Sharan. *A Propagation-based Algorithm for Inferring Gene-Disease Assocations*. in *German Conference on Bioinformatics*. 2008. Citeseer.

59. Prasad, T.K., et al., *Human protein reference database—2009 update.* Nucleic acids research, 2009. **37**(suppl 1): p. D767-D772.

60. Vanunu, O., et al., *Associating genes and protein complexes with disease via network propagation.* PLoS computational biology, 2010. **6**(1): p. e1000641.

61. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic acids research, 2005. **33**(suppl 1): p. D514-D517.

62. Peri, S., et al., *Human protein reference database as a discovery resource for proteomics.* Nucleic acids research, 2004. **32**(suppl 1): p. D497-D501.

63. Lage, K., et al., *A human phenome-interactome network of protein complexes implicated in genetic disorders.* Nature biotechnology, 2007. **25**(3): p. 309-316.

64. Reyes, J.A. and D. Gilbert, *Prediction of protein-protein interactions using one-class classification methods and integrating diverse data.* 2007.

65. Tsiliki, G. and S. Kossida, *Fusion methodologies for biomedical data.* Journal of proteomics, 2011. **74**(12): p. 2774-2785.

66. Savage, R.S., et al., *Discovering transcriptional modules by Bayesian data integration.* Bioinformatics, 2010. **26**(12): p. i158-i167.

67. Singh-Blom, U.M., et al., *Prediction and validation of gene-disease associations using methods inspired by social network analyses.* PloS one, 2013. **8**(5): p. e58977.

68. Huttenhower, C., et al., *A scalable method for integration and functional analysis of multiple microarray datasets.* Bioinformatics, 2006. **22**(23): p. 2890-2897.

69. Huttenhower, C., et al., *The Sleipnir library for computational functional genomics.* Bioinformatics, 2008. **24**(13): p. 1559-1561.

70. Vinh, N.X., et al., *GlobalMIT: learning globally optimal dynamic bayesian network with the mutual information test criterion.* Bioinformatics, 2011. **27**(19): p. 2765-2766.

71. Wang, D., et al. *Combining Gene Expression Profiles and Protein-Protein Interactions for Identifying Functional Modules*. in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*. 2012. IEEE.

72. Chang, H.-H. and M. McGeachie. *Phenotype prediction by integrative network analysis of SNP and gene expression microarrays*. in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. 2011. IEEE.

73. Guan, Y., et al., *A genomewide functional network for the laboratory mouse.* PLoS computational biology, 2008. **4**(9): p. e1000165.

74. Vastrik, I., et al., *Reactome: a knowledge base of biologic pathways and processes.* Genome biology, 2007. **8**(3): p. R39.

75. Mi, H., et al., *PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways.* Nucleic acids research, 2007. **35**(suppl 1): p. D247-D252.

76. Yu, W., et al., *Dynamic protein-protein interaction subnetworks of lung cancer in cases with smoking history.* Chinese journal of cancer, 2013. **32**(2): p. 84-90.

77. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic acids research, 2002. **30**(1): p. 207-210.

78. Qian, L., et al., *Classification of Time Series Gene Expression in Clinical Studies via Integration of Biological Network.* PLOS ONE, 2013. **8**(3): p. e58383.

79. Lu, L.J., et al., *Assessing the limits of genomic data integration for predicting protein networks.* Genome research, 2005. **15**(7): p. 945-953.

80. CellMap. *http://cancer.cellmap.org*.

81. Database, N.P.I., *http://pid.nci.nih.gov*.

82. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucleic acids research, 2004. **32**(suppl 1): p. D277-D280.

83. Witten, I.H. and E. Frank, *Data Mining: Practical machine learning tools and techniques*. 2005: Morgan Kaufmann.

84. Zhang, L.V., et al., *Predicting co-complexed protein pairs using genomic and proteomic data integration.* BMC bioinformatics, 2004. **5**(1): p. 38.

85. Jiang, C., et al., *TRED: a transcriptional regulatory element database, new entries and other development.* Nucleic acids research, 2007. **35**(suppl 1): p. D137-D140.

86. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* arXiv preprint arXiv:1106.1813, 2011.

87. Torgo, L., *Data mining with R: learning with case studies*. 2010: Chapman & Hall/CRC.

88. Maslov, S. and K. Sneppen, *Specificity and stability in topology of protein networks.* Science Signaling, 2002. **296**(5569): p. 910.

89. Breiman, L., *Bagging predictors.* Machine learning, 1996. **24**(2): p. 123-140.

90. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

91. Breiman, L.; Available from: http://www.stat.berkeley.edu/~breiman/RandomForests/.

92. Seiffert, C., et al. *RUSBoost: Improving classification performance when training data is skewed*. in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. 2008. IEEE.

93. Guide, M.U.s., *The mathworks.* Inc., Natick, MA, 2012. **5**.

94. Breitkreutz, B.-J., et al., *The BioGRID interaction database: 2008 update.* Nucleic acids research, 2008. **36**(suppl 1): p. D637-D640.

95. Hedenfalk, I., et al., *Gene-expression profiles in hereditary breast cancer.* New England Journal of Medicine, 2001. **344**(8): p. 539-548.

96. Wu, G., X. Feng, and L. Stein, *A human functional protein interaction network and its application to cancer data analysis.* Genome biology, 2010. **11**(5): p. R53.

97. Smith, C.L., C.-A.W. Goldsmith, and J.T. Eppig, *The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information.* Genome biology, 2004. **6**(1): p. R7.

98. Robinson, P.N., et al., *The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease.* The American Journal of Human Genetics, 2008. **83**(5): p. 610-615.

99. Fawcett, T., *An introduction to ROC analysis.* Pattern recognition letters, 2006. **27**(8): p. 861-874.

100. Kubat, M. and S. Matwin. *Addressing the curse of imbalanced training sets: one-sided selection*. in *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*. 1997. MORGAN KAUFMANN PUBLISHERS, INC.

101. Begg, R. and J. Kamruzzaman, *A machine learning approach for automated recognition of movement patterns using basic, kinetic and kinematic gait data.* Journal of biomechanics, 2005. **38**(3): p. 401-408.

102. Benesty, J., et al., *Pearson correlation coefficient*, in *Noise reduction in speech processing*. 2009, Springer. p. 1-4.

103. Dennis Jr, G., et al., *DAVID: database for annotation, visualization, and integrated discovery.* Genome biol, 2003. **4**(5): p. P3.

104. Kreeger, P.K. and D.A. Lauffenburger, *Cancer systems biology: a network modeling perspective.* Carcinogenesis, 2010. **31**(1): p. 2-8.

Vitae

Name	:Sadiq Abdulmohsin Makki Al-Insaif
Nationality	:Saudi
Date of Birth	:9/9/1988
Email	:sadiq.alinsaif@gmail.com
Address	:P.O.Box 3758, Ras-Tanura 31911, Saudi Arabia
Last Degree	:Bachelor of Science in Computer Science, King Fahd University of Petroleum and Minerals (2012)

Publications

1. D. Colak, O. AlHarazi, ***S. Al-Insaif***, N. Kaya, S. Bin Amer, "Identifying Subnetwork markers associated with Breast Cancer in Middle Eastern Women", Proceedings: Qatar Foundation Annual Research Conference, Doha, Qatar, Nov. 2013.

2. Mazin Saeed, Faisal Saleh**,** ***Sadiq Al-Insaif***, Mohamed El-Attar. "Evaluating the Cognitive Effectiveness of the Visual Syntax of Feature Models."**,** proceedings: The first Asia Pacific Requirements Engineering Symposium (APRES 2014), Auckland, New Zealand, March 2014.

3. *Mohamed Deriche*, ***Sadiq Al-Insaif***, *Samer Arafat, Mohammed Abdul-Azeem Siddiqui*. ” Comparative Study of Different Classifiers for Epileptic Seizures EEG-Based Classification*.” Information Technology in Biomedicine, IEEE Transitions on*, (2013). Submitted.
4. *Emad Ramadan, Sadiq Al-Insaif, Md Rafiul Hassan and Divd Tuck. "Network Topology Measures for Identifying Disease-Gene Association in Breast Cancer", (2014)*. Soon to be submitted
5. *Sadiq Al-Insaif and Md Rafiul Hassan. “Automated Classification of IVF Using Machine Learning Tools”, (2014)*. Soon to be submitted.