MACHINE LEARNING APPLICATION ON GENETIC ANALYSIS FOR DISEASES USING HUMAN INTERACTOME

by

Pakeeza Akram

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science

Summer 2018

MACHINE LEARNING APPLICATION ON GENETIC ANALYSIS FOR DISEASES USING HUMAN INTERACTOME

by

Pakeeza Akram

Approved: _____
Kathleen Mccoy, Ph.D.
Chair of the Department of Computer and Information Science

Approved: _____
Babatunde A. Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved: _____
Ann Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Li Liao, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Cathy Wu, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Keith Decker, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Fei Wang, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

I express my sincere gratitude to my advisor, Dr. Li Liao, for his support and encouragement throughout my study. I would also like to thank my committee members for their cooperation. I like to thank the University of Delaware for giving me this opportunity to work with the finest people in the field. I also extend my gratitude to Fulbright Scholarship program which gave me the opportunity to work and collaborate with the international community.

I would also like to thank my mom Tehmina Akram whose support was always there to break the traditional norms and letting me get a higher education. I would also like to thank my husband Dr. Muhammad Asif Warsi and my kids Muhammad Ibrahim and Emaan Fatima and my siblings for their love and support. I would also like to thank all my family members and friends especially Debarati Roychowdhary for her countless support throughout my research in the study and personal matters. This thank you note will be incomplete without mentioning the person who envisioned me getting this degree, my father, Muhammad Akram Khan, sitting in stars now to see my success. Ph.D. along with family helped me in managing work in the given timeline and an organized manner.

Thank you all for your love and support, you made this journey smooth and easy for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

As a huge amount of omics data and clinical data is being gathered, analyzing the data and identifying genetic causes of diseases has become a one central task in bioinformatics. In this dissertation, we worked on computational analysis of comorbid diseases at three different levels. Comorbidity is the phenomenon of having two or more diseases co-occurring not by random chance. These diseases present enormous challenges to accurate diagnosis and treatment. The primary goal of this research is to incorporate as much information as we can and develop and apply the state-of-art algorithms for solving several questions related to comorbid disease prediction. The dissertation is also aimed to provide new information useful to explore further the human genome and its behavior.

First, at the sequence level, we studied the effect of non-synonymous single nucleotide polymorphisms (SNPs) on diseases (cancers to be specific). Specifically, we investigated how connecting SNPs in the context of haplotype and interacting sites of proteins encoded by affected genes can improve the prediction performance. We trained classifiers on both sequential and structural features extracted from the affected genes and assessed the predictions made by the trained classifiers using cross-validation. We found that accuracy was consistently enhanced by combining sequential and structural features, with the increase ranging from a few percentages points up to more than 20 percentage points. The results of putting SNPs in the context of interacting sites were less consistent compared to individual SNPs prediction, whereas the SNPs that appear together in haplotype showed a stronger correlation with one another and with the

phenotype, and therefore led to significant improvement in prediction performance, with ROC score increased from 0.81 to 0.95. We found similar prediction performance in context of residue prediction at interacting site and non-interacting site, where ROC score increased from 0.66 to 0.86.

Second, at gene cluster level, we worked on the identification of common genes associated with comorbid diseases. This task can be critical in understanding the pathobiological mechanisms of disease comorbidity. We developed a novel method to predict missing common genes related to a comorbid disease pair. Specifically, searching for missing common genes is formulated as an optimization problem to minimize network-based module separation from two subgraphs produced by mapping genes associated with disease onto the interactome. Using cross-validation on more than 600 disease pairs, our method achieves significantly higher average receiver operating characteristic ROC score of 0.95 compared to a baseline ROC score 0.60 using randomized data. Missing common genes prediction is aimed at completing the gene set associated with comorbid disease, to provide a better understanding of biological intervention such as gene-targeted therapeutics related to comorbid diseases. We also provided a few case studies to showcase the pathobiology of genes and their correlation to metabolic pathways.

Third, at the disease level, as an effort toward better understanding the genetic causes of comorbidity, we developed a method to predict how likely two given diseases are comorbid. Intuitively, two diseases that share more common genes shall have increased chance of being comorbid. Previous work shows that after mapping the associated genes onto the human interactome, the distance between the two disease modules (subgraphs) is correlated with comorbidity, and hence can used for

comorbidity prediction. In order to fully incorporate structural characteristics of interactome as features for more accurate prediction of comorbidity, we developed a new method that embeds the human interactome into a high dimensional geometric space and uses the projection onto different dimension to "fingerprint" disease modules. A supervised machine learning classifier is then trained to discriminate comorbid diseases from non-comorbid diseases. In cross-validation using a dataset of more than 10,000 disease pairs, we reported that our model achieved a remarkable performance of ROC score=0.90 for comorbidity at relative risk RR=0 and ROC score =0.76 for comorbidity at relative risk RR=1, which significantly outperformed the previous method. This validated our hypothesis that embedding the interactome to a high dimension space aids the extraction of informative features for effective learning and opens the possibility of further incorporating domain specific information such as weighting known disease related pathways.

Chapter 1

INTRODUCTION

In this chapter, we introduce a brief background on the biological and computational concepts relevant to this dissertation. Then we discuss the related work already done towards solving problems addressed. We also list some publicly accessible data sources that have been utilized in our research. In the end, we will briefly discuss the drop-down of the dissertation with an emphasis on the specific contribution made during the study.

1.1 Background

1.1.1 Genes and Proteins

Genes are essential functional macromolecules of life. A gene is a segment of DNA that encodes functional RNA (Ribonucleic acid) and protein products. DNA, an acronym for deoxyribonucleic acid is made up of four nucleotide bases: Adenine (A), Guanine (G), Thymine (T) and Cytosine(C). These nucleotide bases are considered as a building block of the life that carries the genetic instructions used in the growth, development, functioning of all known living organisms. RNA is like DNA with few differences. Chemically, both are nucleic acids of nitrogen-containing bases joined by the sugar-phosphate backbone. Structurally, RNA is a single-stranded whereas DNA is double-stranded; and DNA has Thymine, whereas RNA has Uracil. Moreover, RNA nucleotides contain ribose sugar, rather than the Deoxyribose sugar by DNA. Functionally, DNA maintains the genetic information, but RNA uses genetic

information to synthesize the protein. The chain of processes consists of transcription, transport, and translation. During these processes genes are transcribed to RNA, then preprocessing and removal of the non-coding region occur, and then RNA is transported out of the nucleus. Later the process of translation takes place, i.e., of conversion of RNA into protein. In a nutshell, all these processes are called the central dogma of life.

Proteins are large macromolecules, consisting of one or more long chains of amino acid residues proteins are involved in performing a range of functions in living organisms. Some of these functions are catalyzing metabolic reactions, regulating the biological process, transporting materials, building immunization system

During the central dogma processes, an error can occur in the DNA code usually at the time of replication. Usually, this error gets removed during error-prone repair process like microhomology-mediated end joining, but sometimes the errors remain in individual genome leading to dysfunction of DNA, RNA or protein. This process is known as mutation where permanent alteration of the nucleotide sequence may occur. Gene mutation can lead to different types of changes in the DNA/RNA sequences. These changes may result in altering the product of the gene, preventing the gene from functioning properly or may have no effect.

Mutations can be divided into several types by effect on structure, impact on function and impact on fitness. Mutation types are further categorized to substitution, insertion, and deletion, where each category name suggests that either the nucleotide is exchanged to another base, or few bases have been added or deleted respectively. Substitution mutation is also known as a point mutation or single-nucleotide polymorphism (SNP). It is a variation at a specific genome position in a single nucleotide where each change is present to some appreciable degree within a population

(e.g.,> 1%). To illustrate, at the specific position the base G appear in most individuals, but for minority base, T appears. This position has SNP with two possible variations G or T. SNPs may lead to malfunction of protein causing diseases. There is a wide range of known SNPs associated with several conditions. For example, a single base mutation in BRCA1 gene may lead to breast cancer.

1.1.2 Protein-Protein Interaction (PPI) and PPI Networks

Proteins perform their function in the conjugation of another protein. This cooperation is known as protein-protein interaction (PPI). During PPIs, there is physical contact between two or more proteins because of biochemical events geared by electrostatic forces like ionic bonding, hydrogen bonding, covalent bonding or hydrophobic interactions. Therefore, to have a better understanding of biochemical processes, intracellular signaling pathways and modeling of complex protein structures several efficient and accurate PPI detection methods have already been developed and getting improved day by day. Traditionally, high-throughput experimental methods like yeast two-hybrid, co-immunoprecipitation and many more were used. These methods are prone to have high false-negative rates along with high cost. Therefore, several computational methods have been developed to predict protein interaction with each other by calculating their similarity using sequence homology, gene co-expression, or phylogenetic profile. With this development PPIs, discovery and validation have been expedited resulting in a reasonable size of PPI Networks to be formed. Formally PPI network can be represented as a graph G= (V, E) with V nodes (proteins) and E edges (interactions). G is defined by the adjacency matrix A with V x V dimensions.

$$A(i.j) = \begin{Bmatrix} 1, if(i,j) \in E \\ 0, if(i,j) \notin E \end{Bmatrix} \qquad 1.1$$

Where in Eq (1.1), i and j are two nodes in the nodes set V, and (i, j) represents an edge between i and j, (i, j) ∈ E. Consequently, researchers have invested more time to study PPI networks to get more and more information for better understanding of individual protein and its biological function in isolation as well as in protein complex.

1.2 Related Work

We plan to have a systematic study of the comorbid diseases and their genetic study from various aspects given PPI network and gene set associated with each comorbid disease pair. Here we will discuss previous work that focus on building prediction of mutations and their impact on diseases, gene-disease relation, and disease-disease association.

1.2.1 SNPs Effect on Protein-Protein Interaction

It has been widely accepted that genetic variations can be associated with diseases. Missense non-synonymous single nucleotide polymorphism (nsSNP) is considered as one of the most common type of variation [1]. Missense nsSNP is a variation in which an amino acid in the protein sequence is changed due to a single point mutation. Because of the association between genetic variations and diseases, there has been active research to identify SNPs and to determine their phenotypic effects, with some reported success in finding the variants as causes to diagnose, treat and prevent complex diseases [2]. Understanding how these nsSNPs affect protein function remains a critical task. Protein-Protein interaction sites have been considered as a hotspot for nsSNP associated with diseases [3]. To unveil genetic variations and functional effect on a protein, multiple methods have been developed, such as enzyme activity prediction [1, 4], and detection of disease potential of a SNP [5]. Recently, the computational

alanine scanning method has been developed to study SNPs effect on protein-protein interaction, essentially by replacing every single residue with alanine to see the effect on protein by estimating free energy change between the wild and the mutated one [6-10]. Another recent work has been done for disease associated nsSNPs on protein-protein interactions by investigating the change in binding energy using force field and electrostatic calculation [11].Most methods have primarily focused on either using sequence based properties, such as conservation score like SIFT [12], or using only structure based properties, such as PoPMuSiC [13]. However, most recently there have been attempts at hybrid approaches for SNP prediction, such as Polyphen2, which have shown promising prediction results as compared to using sole properties of structure or sequence [7]. It has also been reported that individual SNPs and haplotypes have different effect on the protein function [14]. In certain cases, with the presence of two SNPs, the disease-causing SNP becomes recessive and does not exert effect on protein function [14]. Despite the progress made using specific protein features, accurate prediction of effect of nsSNP on PPI that lead to specific diseases remains a major challenge, especially, on the rare focused area of effect of SNP at interacting site of protein or while acting as haplotype. An effort can also be made towards associating SNPs with comorbid diseases to determine the underlying cause of respective disease pair. It is well established that one gene can perform a set of functions and these genes can intertwine or crosstalk between pathways. Genetic variation in one gene may lead to more than one disease in the living organism.

1.2.2 Gene-Disease Association

Biological cause of diseases is examined through several methods. These methods include genes analysis, pathway dysfunction, age factor analysis and

environmental factor analysis. Despite having a wide range of knowledge, the interconnectivity of these factors is not fully understood. System biology is considered to play an empirical role in uniting this information on one platform. These studies lead to the development of human diseasome which combines set of all disease and their implicated genetic changes.

Majorly there are three methods to measure similarity score between diseases: annotation-based, function-based and topology-based. These measures use the different biological information to predict disease associations. Given a pair of disease, annotation-based measure calculates similarity score based on the overlap of their annotated genes using Jaccard index. Function-based measure generates similarity score based on the overlap of disease pair associated biological functions obtained from GO annotations. The topology-based measure use topology information from the underlying PPI network for given disease pair, and estimates disease similarity scores based on the topological similarity of their annotated genes [15]. There is a limitation of using annotation based or functionally based measure since they do not incorporate any protein-protein interaction (PPI) information. PPI provides the specific information regarding gene and their interaction with other genes. This information provide links necessary to evaluate biological pathway affected and resulting in comorbid diseases. In the past, few years research interest is inclined towards topology-based similarity measure. Several successful studies have been conducted using a topology-based measure such as disease catalog through a network-based analysis of associations among genes, proteins, metabolites, intermediate phenotype and environmental factors that influence patho-phenotype was generated [16]. There is a construction of a “viral disease network” of disease associations to analyze the role of viruses and disease

phenotypes [17]. This study represented several diseases that have not previously been associated with infection by the corresponding viruses. Quite an identical method was used in another study to map disease relationships through a network derived from metabolic data instead of viruses. They represented an association that known metabolic coupling between enzyme-associated diseases reveal comorbidity patterns between diseases in patients [18]. Another group studied the disease genes position within the human interactome to predict new cancer-related genes [19]. There is also a gene-centric approach to disease association discovery was examined, this group took 110 diseases for which a set of disease genes are known, and compared gene sets and their positions within the gene network to infer associations of related disorders [20]. Another group developed data fusion approach to mine human disease-disease associations. They revised the links between diseases using related systems-level data, including protein-protein and genetic interactions, gene co-expression, metabolic data, drug-target relations. By fusing these data, they identified several disease-disease associations that were not present in Disease Ontology and validate their existence from literature and significant comorbidity effects in associated diseases [21]. All this effort is towards associating several biological factors to diseases. It is beneficial to have an inference how disease-associated gene location may affect when overlay as a network. In this aspect, recent work of disease-disease association has been formulated using disease module. Disease module is a subgraph where genes associated with various diseases appear as distinct modules in the human interactome. This group predicted that co-occurring disease pair has the overlapping module. This study also states that genes associated with such a disease pair are closely located on the human interactome [22]. There is still a need to accelerate research for associating diseases with causal effects

including related genes, pathways and metabolic data to get a meaningful understanding, especially for the comorbid disorders to diagnose, treat or prevent them at their early stage, since it is contributing towards the cause of high mortality rate.

1.2.3 Disease-Disease Association

Human health is considered as the most critical task in health informatics. Malfunction of the gene and its products can lead to a disease which can interfere with the normal functioning of human body. It is well studied that one gene can play multiple functions resulting in causing not only single but a pair or more diseases to a person simultaneously [23, 24]. The phenomenon of having two or more disorders in one person at a time is not by chance and is known as disease comorbidity [25-27]. Disease comorbidity has the adverse prognosis and profound consequences, like frequent visits and more extended stays at hospitals and high mortality rate [28, 29]. For instance, it is studied that sleep apnea is the secondary cause of hypertension [30]. It was depicted using a small dataset that 56% of people having sleep apnea are suffering from hypertension at the same time. Another study presented that the people with both cardiovascular diseases (CVD) and chronic kidney disease (CKD) were 35 percent more likely to have recurrent cardiovascular events or die than those with CVD alone [5]. Drug toxicity and intolerance is also a significant problem while treating such patients as multiple drugs are incorporated to treat several disorders, where these drugs might have possible negative interaction with one another [31].

Researchers are engaged in providing an ultimate solution towards finding the comorbid behavior using various properties and attributes to each protein. They focus on specific property associated with disease to distinguish. The Human Disease Network (HDN) suggest common mutant genes is the cause of disease comorbidity [19].

It is also seen that disease comorbidity is possible due to enzymes catalyzation during metabolic reactions in the metabolic network [18, 32]. There are several conclusions drawn about disease comorbidity using shared protein-protein interactions (PPIs) resulting in disease network with a disease associated rewired PPI [33-35]. There are few computational approaches that have been proposed to predict disease comorbidity. In a study PPI networks were used to locate PPIs associated with co-occurrences of conditions [36], it was found that protein localization attributes to identify comorbidity in genetic diseases [37]. Another study provided the association of phenotypically similar diseases might have a connection through evolutionary associated genes [38]. Recently, comoR, a useful tool has been developed to predict disease comorbidity by incorporating several existing tools into one package [25]. This package is a valuable tool with a limitation that each process works independently. For instance, one method is, Comorbidity Path which predicts disease comorbidity based on disease-associated pathways only and the other tool comorbidity OMIM consider disease gene associated with OMIM database under certain threshold only.

Another study considered each disease as a disease module [22]. Disease module is the subgraph of all the genes associated with a disease on the human interactome. They modeled an algorithm to find comorbid diseases called module separation. Module separation is the average of all pair shortest distance of genes within the $disease_A$ and $disease_B$. The algorithm suggests that lower the module separation higher will be the comorbid relation. They find that only 7% of the disease pairs follow this criterion. Most recently, PCID an algorithm has been developed for comorbidity prediction based on the integration of multi-scale data. They use the heterogeneous information to describe disease association. The information includes genes, protein interactions,

pathways, and phenotypes but instead of focusing on all disease their focus is to predict only those diseases which co-occur with some primary illness, where the primary infection should be a well-studied and tended to be comorbid. This criterion resulted in only 73 disease pairs in their dataset [39].

1.3 Databases

In this section, we list some publicly accessible PPI databases and web-services that related to our research.

- 3DID: 3DID is a database of three-dimensional interacting domains (3did). It has a collection of high-resolution three-dimensional structural templates for domain-domain interactions. It contains templates for interactions between two globular domains as well as novel domain-peptide interactions. http://3did.irbbarcelona.org/
- BioGRID: BioGRID is an interaction repository with data compiled through comprehensive curation efforts. http://thebiogrid.org/
- GenBank: GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. http://www.ncbi.nlm.nih.gov/genbank
- Gene Ontology (GO) Database: The GO Consortium provides an ontology of defined terms representing gene product properties, including cellular component, molecular function, biological process. http://www.geneontology.org
- PubMed: PubMed is a search engine accessing primarily the MEDLINE database of references and abstracts of papers on life sciences and biomedical topics. http://www.ncbi.nlm.nih.gov/pubmed
- Reactome: It is a free, open-source, curated and peer reviewed pathway database. Reactome provides tools for the visualization, interpretation and

analysis of pathway knowledge to genome analysis and systems biology. http://www.reactome.org/

- STRING: STRING is a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations. http://string-db.org
- UCSC: The University of California Santa Cruz (UCSC) Genome Browser is a web-based resource for the scientific research with convenient access to a database of high-quality genome sequence and annotations. https://genome.ucsc.edu/
- Uniprot: It is a free database of protein sequence and functional information. Several entries are being derived from genome sequencing projects. It also contains huge amount of information about the biological function of proteins obtained from the research literature as well. http://www.uniprot.org/

1.4 Contributions

As we have argued in the previous section, many of the algorithms currently used in biological prediction problems have limitations. In this thesis, we propose several methods focusing on three different levels to contribute towards system biology utilizing the power of the state-of-art machine learning algorithms. These three levels are as follows:

1. SNP Prediction for several cancer types, specifically at concentrated positions of SNP's as well as on comorbid disease
2. Gene and Interaction Prediction, specifically identification of missing common genes and missing common edges associated with comorbid diseases
3. Disease-Disease association, especially emphasizing on the protein-protein interactions and genes associated with a disease as the source.

In Chapter 2, we study prediction of non-synonymous SNPs on several cancers, particularly in the context of haplotype and interaction sites. We formalize the prediction of SNP effects on diseases as a classification problem and then apply machine learning techniques. This work was presented as a poster in International Joint Conference on Artificial Intelligence International Workshop on Biomedical Informatics with Optimization and Machine Learning (IJCAI BOOM 2016), and at The 11th International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2017) and ultimately published in Journal of Biomedical Science and Engineering volume 10 pages 28-44 as "Cancer Specific Non-Synonymous Single Nucleotide Polymorphism Prediction in the Context of Haplotype and Protein Interacting Sites" by Pakeeza Akram and Li Liao [40].

In Chapter 3, we present a novel method to predict missing common genes associated with comorbid diseases using human interactome. There are three fundamental goals achieved in this study. One, that we were able to formulate an optimization task for predicting the missing commonalities between two diseases. Second, this prediction helps in connecting the missing dots on incomplete human interactome. Third, we showcase the significance and implication of genes and their behavior by mapping on metabolic pathways. This study was accepted for conference 6th IEEE International Conference on Computational Advances in Bio and Medical Sciences ICCABS 2016 [41] and was also published in BMC Genomics as "Prediction of missing common genes for disease pairs using network-based module separation on incomplete human interactome" by Pakeeza Akram and Li Liao [42].

In Chapter 4 we will present a new model to predict comorbid diseases for large dataset. Our model utilizes the information of PPI network in geometric space and

incorporates protein interaction and pathway association. This work is accepted for conference 14th International Symposium on Bioinformatics Research and Applications, and it is also accepted for journal publication.

Chapter 2

DISEASE SPECIFIC NON-SYNONYMOUS SINGLE NUCLEOTIDE POLYMORPHISM PREDICTION

2.1 Introduction

An integral part of a complicated biological rewiring is the network of protein-protein interactions (PPIs). Due to protein interaction, a single gene mutation is not restricted to the actions of its gene products but can result in disruption of the whole system, also affecting other gene products which, previously, might be performing their functions correctly.

In this chapter, we study the prediction effect of non-synonymous single nucleotide polymorphism (SNPs) on several cancers, i.e., acute myeloid leukemia, breast cancer, colorectal cancer, and esophageal cancer, particularly in the context of haplotype and interaction sites. We formalize the prediction of SNP effects on diseases as a classification problem and then apply machine learning techniques, including support vector machines (SVM) and random forest (RF), to learn from training examples and to classify unseen SNPs. Our comprehensive comparative analysis of different classifiers using a set of evaluation metrics explores not only the utility of various machine learning methods for SNP prediction, but it also shed light on whether and how the prediction of SNP's effect is affected for genetic variations by their presence at interacting sites and non-interacting sites of the protein, or for individual SNPs versus SNPs as haplotype associated with a specific disease. We hypothesize that prediction of SNP effect using both sequence and structure-based protein properties

together will be better that using sequence-based features only or structure-based feature only respectively. We also hypothesize that residues at interacting site have higher prediction performance that residues at non-interacting sites. Finally, it was also hypothesized that SNPs as haplotype pair provide better prediction performance as compared to individual SNP.

2.2 Methods

As mentioned above, we formalize the prediction of SNP's effects on proteins associated with specific diseases as a classification problem and adopt supervised learning strategy. Specifically, two powerful classifiers, random forest [43] and support vector machines (SVM) [44], are selected for this study (explained in next section).

2.2.1 Classification

Features, both sequential and structural properties, of proteins encoded by genes with SNPs that are believed to be relevant for the phenotypic properties, are collected and quantified for use as input vector x to the classifier. Specifically, for this study, we are interested in two types of phenotypic properties: detrimental (a SNP that cause a disease) or polymorphic (a SNP that does not cause a disease), corresponding to the output y of the binary classifier, namely, y = 1 for detrimental and 0 for polymorphic. The classifier is to learn the actual mapping from input to output: y ← x, with a hypothesis function F (x, ɵ), where ɵ collectively represents the parameters of the classifier, for example, the degree d of a polynomial kernel for SVM. The classifier is trained to minimize the empirical error as shown in equation (2.1)

$$min\left\{\sum_{i=1\ to\ n} |(y_i) - F(x_i, \theta)|\right\} \qquad 2.1$$

or a set of n training examples x_i, i = 1 to n, whose phenotypic property y_i is known. Once the classifier is trained, it is used to make prediction / classification on unseen data, i.e., SNPs whose phenotypic property is not known *a priori*.

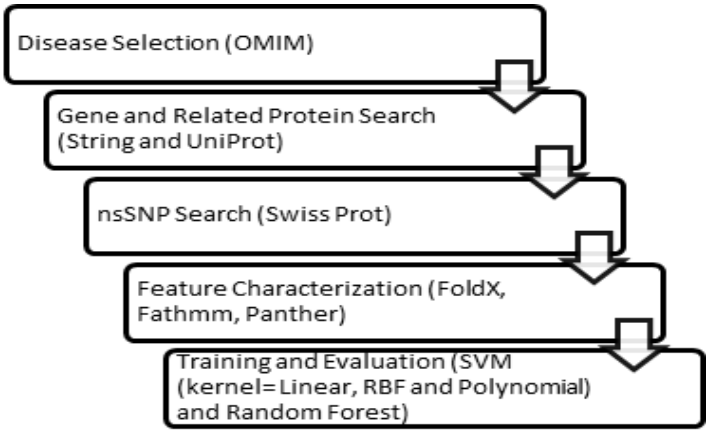


Figure 2.1: Pipeline constructed for nsSNP prediction starting from gene collection to classification estimation.

2.2.2 SVM and its Kernels

SVM is a powerful statistical learning method, and it was proposed by Vapnik (1998). This method has recently been applied with remarkable success in bioinformatics problems.

The basic idea of SVMs is to find a hyperplane that separates two classes represented as points in a vector space, with the maximum margin to the separating hyperplane.

$F(x_i) = \text{sign}(w \bullet x_i + b)$

Where the norm vector w and intercept b are trained with the constraints

$\mathbf{w} \cdot \mathbf{x}_i + b > 0$ if $y_i = +1$

$\mathbf{w} \cdot \mathbf{x}_i + b < 0$ if $y_i = -1$

This maximum margin ensures good generalization, that is, unseen data are then correctly classified according to their location with respect to the hyperplane. The constrained optimization problem can be solved using the Lagrangian technique.

$\max_{\alpha} L(\alpha)$

Where $L(\alpha) = \sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$, and α_i are the Lagrangian multipliers, under the constraints: $\alpha_i \geq 0$, and $\sum \alpha_i y_i = 0$.

Data representation plays a partial role to strengthen SVM where an entity like gene can be represented by set of associated attributes. But the contribution made by the set of attributes to distinguish positive instances from negative one as shown in Figure 2.1 is complicated process [78]. The algorithm for SVM finds a non-linear mapping $\phi()$ that transform the data from the input space (original space) to a feature space (which is often a higher-dimensional space), where the data can be linearly separable.

The mapping in SVM can be quite complex and the dimension can be very high in order for the mapped data to be linearly separable. Therefore, a kernel function is

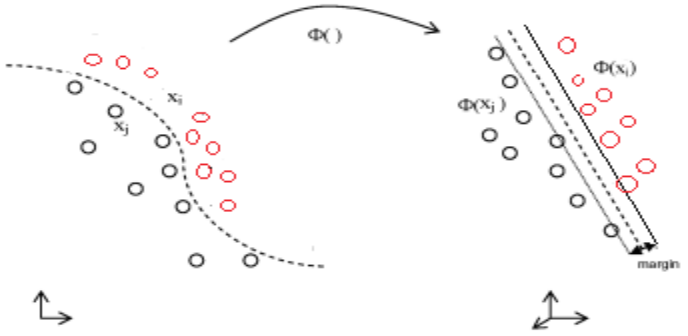


Figure 2.2: Schematic view of non-linear mapping of data from input space to feature space for SVM.

used in place of an explicit mapping, that is, the kernel function $K(x_i, x_j)$ defines how the dot product between two vectors points $\phi(x_i) \bullet \phi(x_j)$ is computed in the feature space. The use of kernel functions avoids explicit mapping to high-dimensional feature space;

high dimensionality often poses difficult problems for learning such as over-fitting, thus termed the curse of dimensionality. Two commonly used generic kernels are Gaussian Radial Basis Function (RBF) and polynomial functions.

For SVM, three different kernels were adopted and assessed: Linear, Radial Basis Function Radial Basis Function

$$KG(x, x') = \exp\left(-\frac{\gamma(\|x - x'\|)^2}{C}\right)$$

where the values for C=3.46 and $\gamma = 1.07$ and Polynomial

$$KP(x, x') = (\langle x, x' \rangle + 1)^d$$

with degree d = 2 was applied. These values of C and degree of polynomial d were optimized by using Opunity 1.1.1, a python package.

2.2.3 Random Forest

Random forests are a combination of tree predictors. Where each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges as the number of trees in the forest becomes large. The strength of the individual trees in the forest and the correlation between them determines the generalization error of a forest of tree [43]. Following are key features due to which random forests are popular. It gives better accuracy than other algorithms. They can handle large datasets efficiently.

It considers all the predictor variables (predictors are selected randomly for each of the trees). It tells about the important variables in the dataset (Variable importance). We can visualize the forest error rate as the forest grows and then decide on the size of the forest accordingly

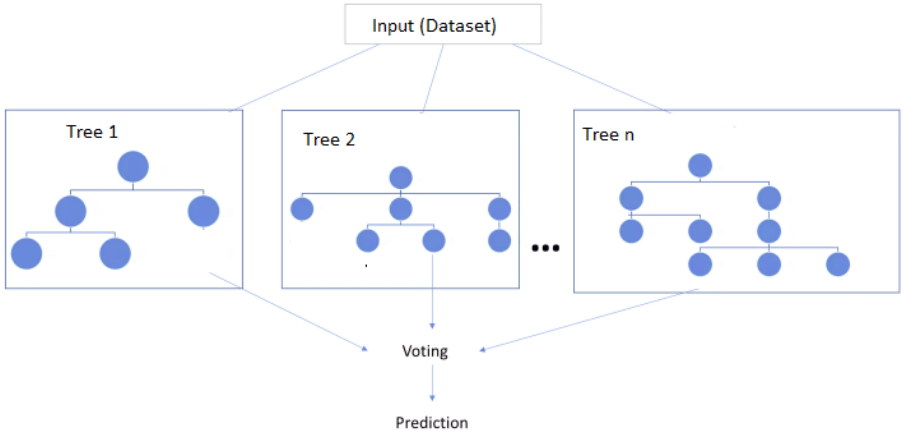


Figure 2.3: Schematic view of Random Forest algorithm.

Feature selection plays a critical role in ensuring adequate learning and reliable prediction. It has been known that mutations that occur at the interface between interacting proteins are more likely to cause detrimental effect as compared to present on other sites. Also, previous studies suggest that haplotype may influence whether a SNP may or may not manifest its phenotype. Therefore, in this study, we are particularly interested in predicting the effect of non-synonymous SNPs on four types of common cancers in the context of SNPs being on protein interaction sites or within a haplotype.

The pipeline developed for this study consists steps for data collection, feature characterization/quantification, classifier training, testing and evaluation, as shown in Figure 2.1. Detail for each step is given in the following subsections.

2.2.4 Data and Feature Characterization

SNPs and phenotypic effect for the four different types of cancers, i.e., acute myeloid leukemia (MIM # 601626), breast cancer (MIM#114480), colorectal cancer (MIM#114500) and esophageal cancer (MIM#114480) – are collected from OMIM.

OMIM is one of the most extensive databases which provides detailed information about phenotype-genotype relation [45].

To determine whether SNPs occur at protein-protein interaction sites, we used STRING database to identify the interaction sites for the affected proteins (i.e., the gene products) [46]. For Acute Myeloid Leukemia, 16 genes are involved, which result in 171 proteins that have specific interactions with each other. Due to unsolved 3D protein structures, the set is reduced to 111 proteins. There are several databases which provide SNP data, including UniProt and dbSNP. For this study, we used SNP from UniProt database [47] because of its extensive collection as compared to other databases. The queries to UniProt identified 1399 nsSNP for these 111 proteins. The same data collection protocol is used for the other three cancers as well. After filtering with required protein structural as well as sequence properties, the final data set consists of 4056 SNP's in total, as listed in Table 2.1.

Table 2.1:Data distribution for cancer type representing polymorphic and detrimental SNP's.

Cancer	Polymorphic	Detrimental	Total
Acute Myeloid Leukemia	1131	268	1399
Breast Cancer	1087	145	1232
Colorectal Cancer	983	131	1114
Esophageal Cancer	961	94	1055
Total	3473	583	4056

We constructed the feature vector by incorporating nsSNP and their several properties related to both sequence and their respective structure. FoldX software was used to calculate parameters which are essential for protein stability. This tool provides

the difference between the energy derived from original protein (wild type) and the substituting the residue with SNP (mutant type) identified for different bond and sites of the protein chain. All energy values are calculated in kcal/mol unit [48]. It provides several important features along with the calculation of the total energy for the mutant and the wild-type protein. Panther software calculates the Substitution Position-Specific Evolutionary Conservation (subPSEC) Scores, and it is based on hidden Markov model (HMM) [49]. It was used to collect subPSEC score. Fathmm was applied to calculate HMM cancer-specific pathogenicity weights [50]. In total 20 features were obtained and all these features are shown in Table 2.2.

We also collected haplotype data for genes associated with the Acute Myeloid Leukemia. A haplotype is considered as the set of polymorphic, which are inherited together. It is referred to a combination of alleles or a set of SNPs that are found on the same chromosome [14]. To collect haplotype information two databases were used in this study. One is HapMap Project, and the other is UCSC genome browser [51, 52]. HapMap Project has a wide range of SNPs, which are collected from dbSNP. Since our dataset consists of SNPs obtained from SwissProt, to capture as many as haplotype data, we incorporate UCSC genome browser, which provides gene-based common allele variants taken from 1000 genome project [53].

Table 2.2:Name and description of each feature.

	Feature Name	Description
1	Fatthm Score	Fatthm score determining the cancerous nature of SNP calculated from fatthm tool
2	SubPSEC	Substitution Position-Specific Evolutionary Conservation (subPSEC) Score
3	Pdeleterious	Probability that a given variant will cause a deleterious effect on protein function calculated by Panther tool
4	Total Energy	Total energy difference of wild and mutant type
5	BackBone HBond	The contribution of backbone Hbonds
6	Sidechain Hbond	The contribution of sidechain-sidechain and sidechain-backbone Hbonds
7	Van der Waals	Contribution of the Van der Waals
8	Electrostatics	Electrostatic interactions
9	Solvation Polar	Penalization for burying polar groups
10	Solvation Hydrophobic	Contribution of hydrophobic groups
11	Vander Waals clashes	Energy penalization due to VanderWaals' clashes (inter-residue)
12	Entropy sidechain	Entropy cost of fixing the side chain
13	Entropy main chain	Entropy cost of fixing the main chain
14	Torsional clash	VanderWaals' torsional clashes (intraresidue)
15	Backbone clash	Backbone-backbone VanderWaals.
16	Helix dipole	Electrostatic contribution of the helix dipole
17	Disulfide	Contribution of disulfide bonds
18	Electrostatic kon	Electrostatic interaction between molecules in the pre-complex
19	Partial covalent bonds	Interactions with bound metals
20	Energy Ionization	Contribution of ionization energy

2.2.5 Cross-Validation and Evaluation

To assess the prediction performance, we adopt the widely accepted cross-validation scheme. Specifically, we used 10-fold cross-validation. The data is randomly split into ten equal-sized subsets, and one set is reserved for testing, and the remaining nine subsets are combined into a training set to train the classifier. This process is repeated ten times, with each subset being used as test set once and the average performance from 10 runs is reported. We used some commonly used measurements to

report the performance, which includes accuracy, precision, recall, F1 score, defined as follows.

$$Recall = \frac{TP}{TP + FN}$$
$$Precision = \frac{TP}{TP + FP}$$
$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$
$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where TP stands for true positive when a SNP is correctly predicted as detrimental, TN for true negative when a SNP is accurately predicted as polymorphism, where TP stands for true positive when a SNP is correctly predicted as detrimental, TN for true negative when a SNP is accurately predicted as polymorphism, FP for false positive when a SNP is incorrectly predicted as detrimental; and FN for false negative when a SNP is incorrectly predicted as polymorphism.

We also evaluate the performance using receiver operating characteristic (ROC) curve and score. ROC is a graphical representation that illustrates the performance of a binary classifier system. The plot is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity or recall while false-positive rate is also known (1 − specificity) [44].

2.2.6 Results and Discussions

In this study, we carried out comprehensive comparative analysis of predicting SNPs effects on four types of cancers. Specifically, we examined the following four different scenarios:

1. Comparison using structural properties only, or sequence properties only or combine effect of both properties using different classifiers (Protein Property based)
2. Specific cancer SNP's prediction or collection of cancers SNP's prediction (Cancer Collection based)
3. SNP's prediction for residues at interacting sites or non-interacting sites (Interaction Site based)
4. SNP's prediction for SNPs within haplotype or individual SNP's (Haplotype Based)

Note that, due to data collection issues, the last two types of analysis were only performed for Acute Myeloid Leukemia.

2.2.6.1 Comparison based on Protein Properties

For the 4056 SNP's listed in Table 2.1, three different datasets were generated. All three datasets have the same number of instances but different dimensionality of the feature vector. First dataset had 3 (sequential) features in it, second dataset had 18 (structural) features and the last dataset had all 21 features in it. Receiver operating characteristic (ROC) score was calculated for 10-Fold cross validation and the mean of those scores is represented in Table 2.3 and Figure 2.4 respectively.

The results show that using structural and sequence-based features together for SNP Prediction provides better results as compared to individual protein properties. It also suggests that hybrid features offer better results for any combination of sequence and structure-based features used. It also shows that random forest performs better among other classifiers used in this task.

Table 2.3: Mean ROC score for SNP prediction using different classifiers for specific protein-based properties.

Classifier	Sequence Based Features	Structure Based Features	Hybrid Features
SVM Linear	0.63	0.5	0.74
SVM RBF	0.76	0.7	0.81
SVM Polynomial	0.58	0.6	0.67
Random Forest	0.9	0.82	0.92

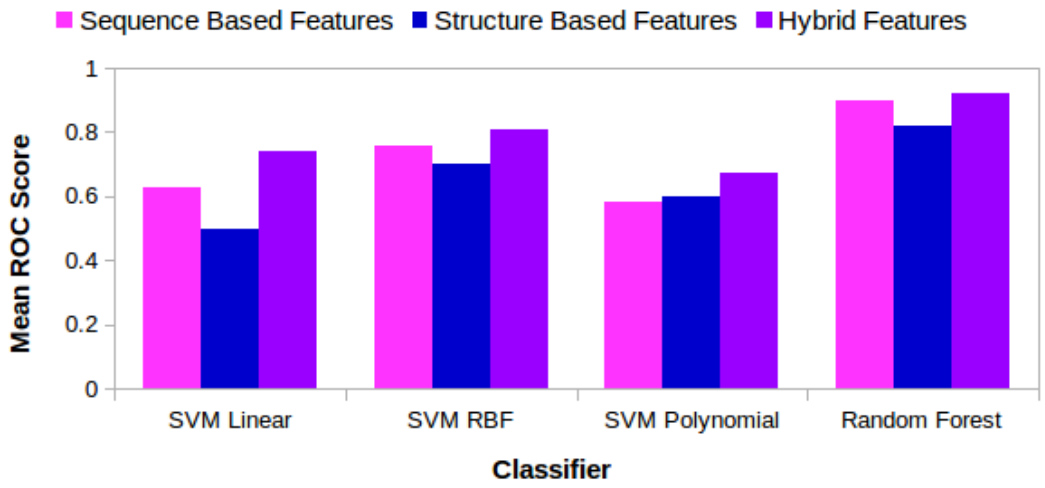


Figure 2.4: Classifier performance using ROC Score for sequence based, structure based and hybrid protein properties.

2.2.6.2 Comparison based on Cancer Collection

For this task, data was collected for four different cancers that are breast cancer, colorectal cancer, esophageal cancer and acute myeloid leukemia, see Table 2.1. It was observed that very few genes, such as TP53, were common for all types of cancers collected for this study and generally in all kinds of cancers. It can be seen from Table 2.1 that the number of detrimental SNPs is low as compared to the polymorphic SNPs.

The difference is almost three times between two types of SNPs. Prediction performance for every classifier for each disease was studied. Table 2.4 lists the performance of each classifier, i.e., SVM Linear, SVM RBF, SVM Polynomial and Random Forest on both detrimental as well as polymorphic SNP.

Table 2.4: Evaluation metric score for each cancer using four different classifiers.

Classifier	SNP type	Precision	Recall	F1-Score	Accuracy
		Acute Myeloid Leukemia			
SVM	Polymorphic	0.87	0.91	0.89	
Linear	Detrimental	0.53	0.41	0.46	0.82
SVM	Polymorphic	0.84	0.94	0.89	
RBF	Detrimental	0.5	0.26	0.34	0.81
SVM	Polymorphic	0.83	0.96	0.89	
Polynomial	Detrimental	0.51	0.17	0.26	0.81
Random	Polymorphic	0.86	0.92	0.89	
Forest	Detrimental	0.51	0.35	0.41	0.82
		Breast Cancer			
SVM	Polymorphic	0.88	0.9	0.89	
Linear	Detrimental	0.13	0.1	0.11	0.81
SVM	Polymorphic	0.88	0.91	0.90	
RBF	Detrimental	0.11	0.08	0.09	0.81
SVM	Polymorphic	0.88	0.9	0.89	
Polynomial	Detrimental	0.13	0.11	0.12	0.81
Random	Polymorphic	0.89	0.88	0.88	
Forest	Detrimental	0.14	0.15	0.15	0.81
		Colorectal cancer			
SVM	Polymorphic	0.88	0.96	0.91	
Linear	Detrimental	0.00	0.00	0.00	0.84
SVM	Polymorphic	0.88	0.96	0.92	
RBF	Detrimental	0.07	0.02	0.03	0.85
SVM	Polymorphic	0.88	0.95	0.91	
Polynomial	Detrimental	0.00	0.00	0.00	0.84
Random	Polymorphic	0.89	0.94	0.91	
Forest	Detrimental	0.26	0.17	0.20	0.84
		Esophageal Cancer			
SVM	Polymorphic	0.91	0.99	0.95	
Linear	Detrimental	0.00	0.00	0.00	0.90
SVM	Polymorphic	0.92	0.99	0.95	
RBF	Detrimental	0.44	0.07	0.13	0.91
SVM	Polymorphic	0.91	0.99	0.95	
Polynomial	Detrimental	0.08	0.01	0.02	0.90
Random	Polymorphic	0.91	0.98	0.94	
Forest	Detrimental	0.14	0.03	0.05	0.90

The above Table 2.4 represents that SVM RBF performs better for esophageal and colorectal cancer and SVM linear executed better for acute myeloid leukemia, while all classifiers performed about equally well on breast cancer. It also shows that for polymorphic SNP prediction precision and recall is much better as compared to the detrimental SNPs. This difference is may be attributed to the skewed data distribution. It is also noticeable that regarding accuracy there is only 1% difference while using different classifiers.

Further, all the cancer types were lumped together to analyze their performance (shown in Table 2.6). It showed that random forest once again performed better. To further evaluate predictive power without using a fixed threshold to determine positive versus negative, receiver operating characteristic (ROC) score was calculated for all classifiers using 10-fold cross-validation. The mean ROC score is represented in Figure 2.5 and Table 2.5. Results from mean ROC score shows that except for acute myeloid leukemia for each disease random forest provides better score. And in general, most of the ROC scores are above 0.70.

Table 2.5: Mean ROC score for each classifier for SNP's classification prediction.

Mean ROC Score	**Combine Cancers**	**Acute Myeloid Leukemia**	**Breast Cancer**	**Colorectal Cancer**	**Esophageal Cancer**	**Uncommon Genes**	**Single Instance**
SVM Linear	0.61	0.81	0.51	0.55	0.58	0.45	0.69
SVM RBF	0.71	0.81	0.68	0.70	0.69	0.38	0.73
SVM Polynomial	0.70	0.78	0.57	0.72	0.66	0.45	0.72
Random Forest	0.90	0.80	0.69	0.73	0.72	0.50	0.77

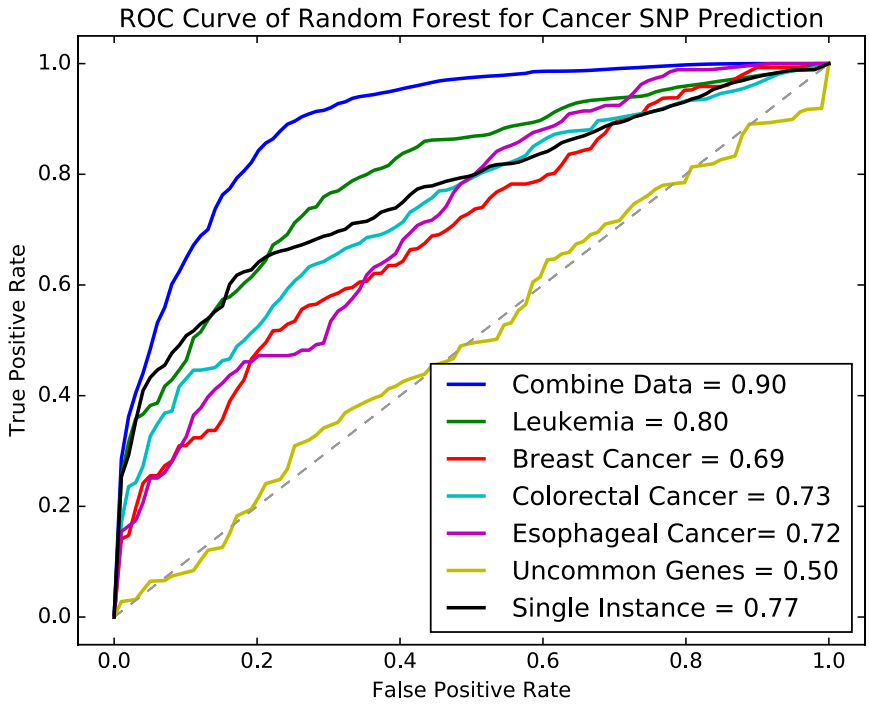


Figure 2.5: Mean ROC score plot for each cancer type using random forest (best classifier for study).

Initially, it was hypothesized that SNP classification for an individual disease would be better than that of combine diseases, but results reflect the opposite. To further investigate a couple of tasks were performed. It was noticed that there were six genes which are common and associated with cancer types selected for this study. These common genes were removed entirely from the data set, and classification was performed. Results showed that mean ROC score for all the cases was less than 0.6 (shown in Figure 2.5). It provides a clue that if there is no common gene among diseases than SNP prediction for individual cancer type will be better but in general, almost all the cancers have specific common genes like TP53 etc.

Another task was performed to examine how training be affected if the combination of all disease SNP without redundancy, i.e., the only single instance of SNP occurs in the final dataset when this gene is shared by more than one of the cancer type. In this case, ROC score was like every individual cancer type SNP classification.

It was noticed and mentioned earlier that like all real time data, detrimental SNP are much less in number than the polymorphic SNPs. It produces an unbalanced dataset. To see what impact data would make if the number of detrimental SNP is equal to polymorphic SNPs. The number of SNPs for the polymorphic class was reduced, and then classification task was performed. It does not show any change in ROC score for the best classifier, but the F1-score for detrimental SNPs was rapidly increased from 0.45 to 0.87 as shown in Table 2.6. This change in detrimental SNP evaluation can be seen from as well as from the Figure 2.6. Lastly, the mean ROC score was calculated using 10-fold cross validation for each classifier as shown in Table 2.6 and found that random forest provides better results as compared to any other classifier. Note that there is no change in the mean ROC score for the best classifier, but SVM with its different kernels is performing better.

Table 2.6:Evaluation metric score for combined cancer SNP using four different classifiers.

Classifier	SNP type	Precision	Recall	F1-Score	Accuracy	ROC Score
			Combine Cancers			
SVM Linear	Polymorphic	0.86	0.98	0.91		
	Detrimental	0.15	0.02	0.04	0.84	0.61
SVM RBF	Polymorphic	0.87	0.97	0.92		
	Detrimental	0.49	0.16	0.25	0.86	0.71
SVM Polynomial	Polymorphic	0.86	0.98	0.92		
	Detrimental	0.45	0.08	0.13	0.85	0.7
Random Forest	Polymorphic	0.9	0.96	0.93		
	Detrimental	0.6	0.35	0.45	0.87	0.9
			Single Instance			
SVM Linear	Polymorphic	0.77	0.84	0.8		
	Detrimental	0.24	0.18	0.21	0.69	0.69
SVM RBF	Polymorphic	0.8	0.85	0.82		
	Detrimental	0.34	0.27	0.3	0.72	0.73
SVM Polynomial	Polymorphic	0.79	0.93	0.85		
	Detrimental	0.39	0.14	0.21	0.75	0.72
Random Forest	Polymorphic	0.82	0.82	0.82		
	Detrimental	0.39	0.4	0.4	0.73	0.76
			Balanced Data			
SVM Linear	Polymorphic	0.83	0.99	0.9		
	Detrimental	0.93	0.79	0.88	0.89	0.88
SVM RBF	Polymorphic	0.82	0.95	0.88		
	Detrimental	0.94	0.79	0.86	0.87	0.88
SVM Polynomial	Polymorphic	0.63	0.89	0.74		
	Detrimental	0.81	0.47	0.6	0.68	0.88
Random Forest	Polymorphic	0.83	0.92	0.87		
	Polymorphic	0.83	0.92	0.87	0.87	0.9

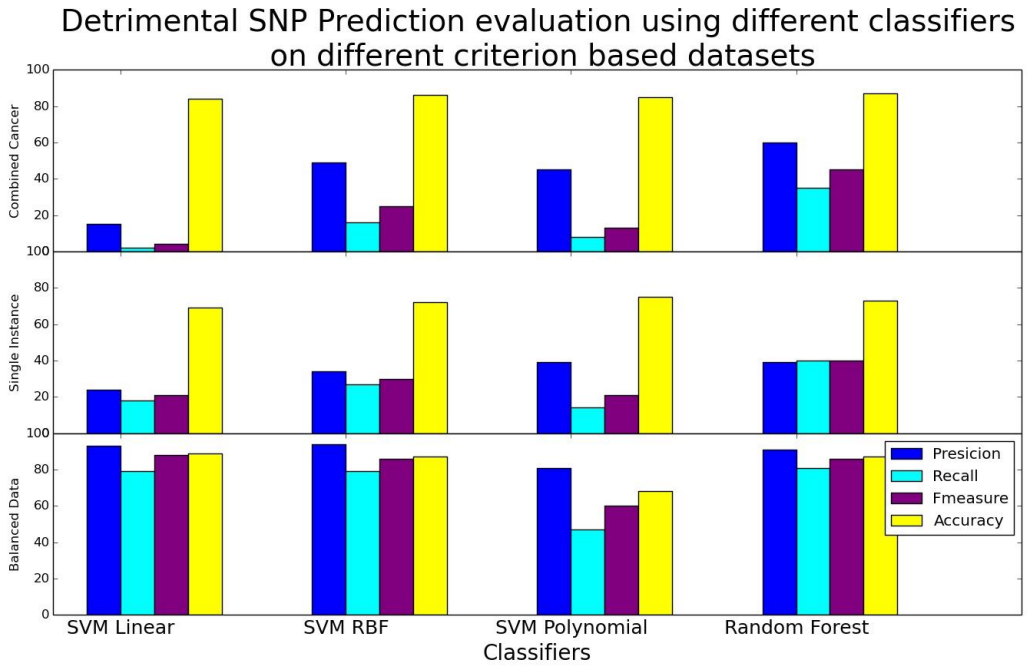


Figure 2.6: Detrimental SNP evaluation for Combined data, single instance data (non-redundant) and Balanced data.

We performed a t-test to assess the statistical significance for the difference between that set of combine cancers, and the set of Acute Myeloid Leukemia on the ROC score of both datasets using the random forest, and the p-value is 0.007458. This analysis concludes that random forest performs better than other classifiers when SNP's prediction is made for specific cancer. This task also proves that specific genes (not including the common genes associated with the multiple cancer types) associated with specific cancer provide better prediction performance.

2.2.6.3 SNP Prediction for Residues at Interacting Site or Non-Interacting Site

The nsSNP prediction was done at the interacting site as well as the non-interacting site. 3DID database (release: June 2015) was used to observe the presence of a residue at the interacting site. It was found that among 40 proteins associated with

acute myeloid leukemia having solved 3D structure and nsSNP there are only 18 proteins which had information for their interacting and non-interacting residues recorded in the database. Two subsets were created for this problem: one having SNPs at interacting residues and the other with SNPs at non-interacting residues. Each of these subsets had 20 features mentioned in Table 2.2 Data distribution is shown in Figure 2.7 for the datasets.

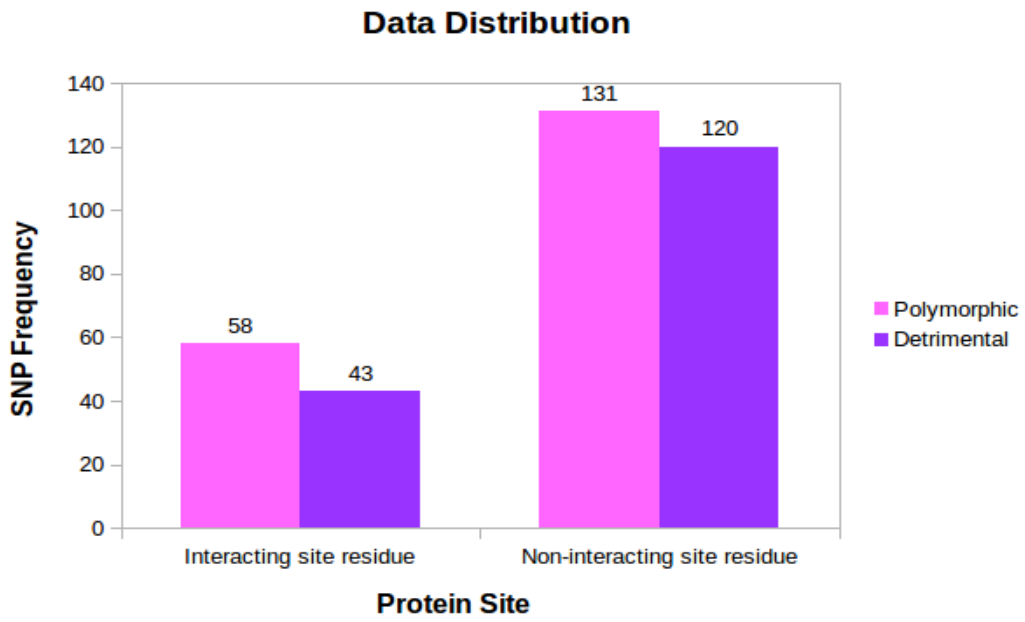


Figure 2.7: SNP data distribution for acute myeloid leukemia at interacting and non-interacting site of protein.

Classification prediction was performed using same classifiers. Their performance concerning precision, recall, F1-Score, and accuracy is given below in Table 2.7. Data distribution is balanced for both subsets, and thus it provides improved results for both datasets when compared to task one datasets regarding polymorphic and detrimental prediction.

Table 2.7: Evaluation metric score for SNPs at interacting and non-interacting sites using four different classifiers.

Classifier	SNP type	Precision	Recall	F-measure	Accuracy	ROC
Interacting Site Residues						
SVM Linear	Polynomial	0.68	0.71	0.7	0.65	0.83
	Detrimental	0.6	0.58	0.59		
SVM RBF	Polynomial	0.68	0.66	0.67	0.62	0.86
	Detrimental	0.56	0.58	0.57		
SVM Polynomial	Polynomial	0.58	0.84	0.69	0.57	0.86
	Detrimental	0.44	0.16	0.24		
Random Forest	Polynomial	0.68	0.74	0.71	0.67	0.72
	Detrimental	0.61	0.53	0.57		
Non-interacting Site residues						
SVM Linear	Polynomial	0.54	0.56	0.55	0.53	0.61
	Detrimental	0.5	0.49	0.5		
SVM RBF	Polynomial	0.54	0.58	0.56	0.53	0.64
	Detrimental	0.5	0.47	0.48		
SVM Polynomial	Polynomial	0.59	0.89	0.71	0.61	0.66
	Detrimental	0.72	0.32	0.44		
Random Forest	Polynomial	0.56	0.56	0.56	0.55	0.53
	Detrimental	0.52	0.53	0.52		

While the overall performance has been dropped, there is an improvement in performance for prediction of detrimental SNP's. Further, ROC score was determined for all classifiers for both datasets as shown in Figure 2.8. The upper panel is for all the classifier trained and tested for SNPs at interacting sites, and the lower group is for non-interacting site SNP's. Mean ROC score for SVM RBF and SVM polynomial were same, i.e., 0.86 for both datasets but in case of non-interacting site residues, SVM polynomial is performing better with 0.66 scores. It concludes that when overall performance of two datasets is considered, SVM polynomial has better performance than any other classifier for this task and it also validates our hypothesis of better prediction of residues at interacting site. Lastly, to verify the statistical significance of

the performance difference, a t-test was performed on the 10-fold cross-validation of SVM polynomial ROC score on interacting site and non-interacting site sets, and it was found that p-value is 0.020197, confirming the statistical significance of the difference.

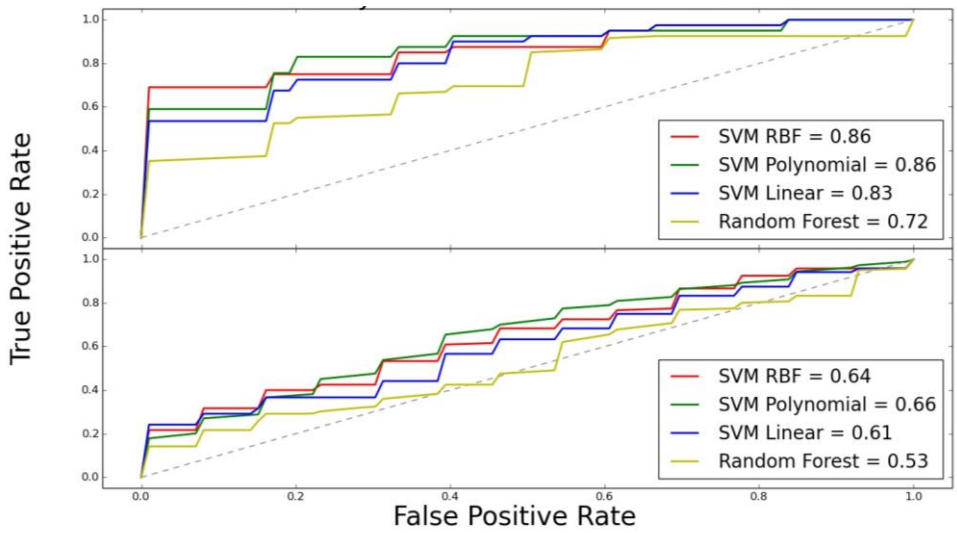


Figure 2.8:Mean ROC score plot for several classifiers at interacting site (upper plot) and at noninteracting site (lower plot) of protein.

2.2.6.4 SNP Prediction Individual SNPS vs SNPS within Haplotype

In this analysis, we examine predicting SNPs effect in the context of the haplotype, i.e., the prediction of individual SNPs versus SNPs within a known haplotype. The search against the database from HapMap Project and the other is UCSC genome browser only identified haplotypes from 14 genes from the gene pool associated with acute myeloid leukemia. Haplotypes were considered in pair only that means every single SNP in haplotype was compared to every haplotype allelic change within the same gene including self-replication. In this task, two subsets were generated: one set consists of haplotypes pairs and the other set consists of all individual SNPs associated

with genes involved in acute myeloid leukemia. Data distribution for these two subsets is given in Figure 2.9.

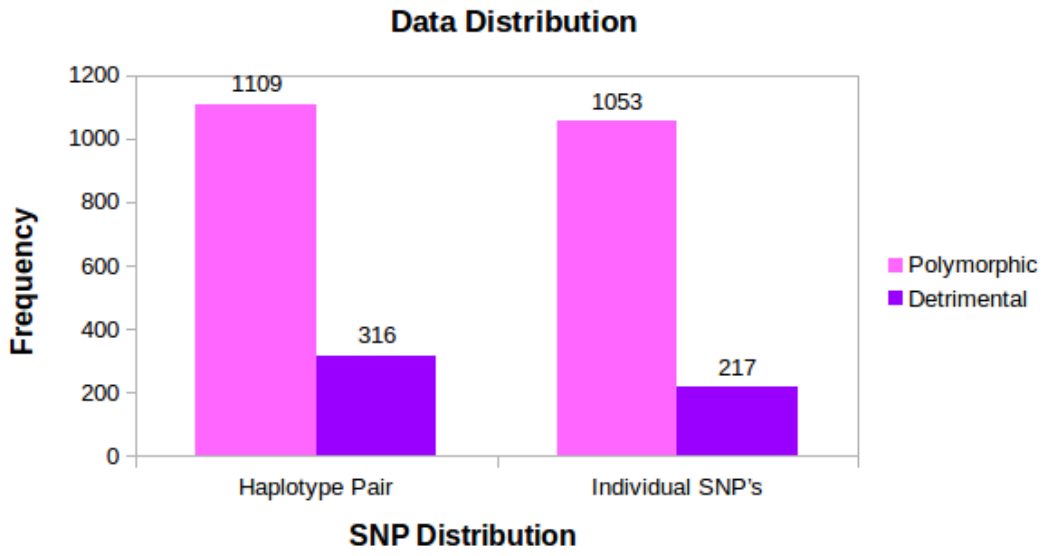


Figure 2.9:SNP data distribution for acute myeloid leukemia as Haplotype and Individual SNP's

For training 10-fold cross-validation was applied to both datasets using SVM with three kernels and random forest. The results for this classification problem are shown in Table 2.8

.

Table 2.8: Evaluation metric score for SNPs in haplotype pair or individual SNP using four different classifiers.

Classifier	SNP type	Precision	Recall	F1-score	Accuracy	ROC
			Haplotype Pair			
SVM Linear	Polymorphic	0.85	0.81	0.83		
	Detrimental	0.43	0.5	0.46	0.74	0.7
SVM RBF	Polymorphic	0.88	0.87	0.88		
	Detrimental	0.57	0.6	0.58	0.81	0.71
SVM Polynomial	Polymorphic	0.88	0.88	0.88		
	Detrimental	0.59	0.59	0.59	0.82	0.67
Random Forest	Polymorphic	0.96	0.92	0.94		
	Detrimental	0.75	0.88	0.81	0.91	0.95
			Individual SNP			
SVM Linear	Polymorphic	0.87	0.91	0.89		
	Detrimental	0.44	0.35	0.39	0.81	0.79
SVM RBF	Polymorphic	0.85	0.92	0.88		
	Detrimental	0.37	0.24	0.3	0.80	0.81
SVM Polynomial	Polymorphic	0.86	0.93	0.9		
	Detrimental	0.45	0.27	0.34	0.82	0.77
Random Forest	Polymorphic	0.85	0.87	0.87		
	Detrimental	0.32	0.3	0.32	0.79	0.79

In Table 2.8 we can see that the best accuracy in predicting haplotype pair is 0.91, a significant increase over 0.82, the best accuracy in predicting individual SNPs. Also, we notice a clear advantage of Random forest for predicting haplotype pairs across the board on all four metrics, whereas SVM Polynomial performs slightly better for predicting of individual SNPs. It is worth noting that the F1-score for haplotype pair of detrimental phenotype is 0.81 by Random Forest classifier, which is an impressive performance given that the datasets (Figure 2.8) are quite skewed toward polymorphic phenotype and therefore present a more significant challenge for correctly predicting the detrimental phenotype. The first four-metrics used in Table 2.8 all depend on a fixed threshold for prediction except ROC score. ROC curve and score can evaluate a classifier's predictive power and performance without relying on a specific prediction threshold. In Figure 2.9, ROC curves and scores are shown for haplotype SNP pairs (top panel) and individual SNPs (bottom panel). The two critical observations from Table 2.8 are essentially maintained: a) pairing SNPs in haplotype help improve phenotype prediction (ROC score = 0.95, achieved by RF), as compared to predicting phenotype for individual SNPs (ROC score = 0.81, achieved by SVM-RBF); b) while RF generally performs better, SVM-RBF has a slight edge in predicting individual SNPs and shown in Figure 2.10.

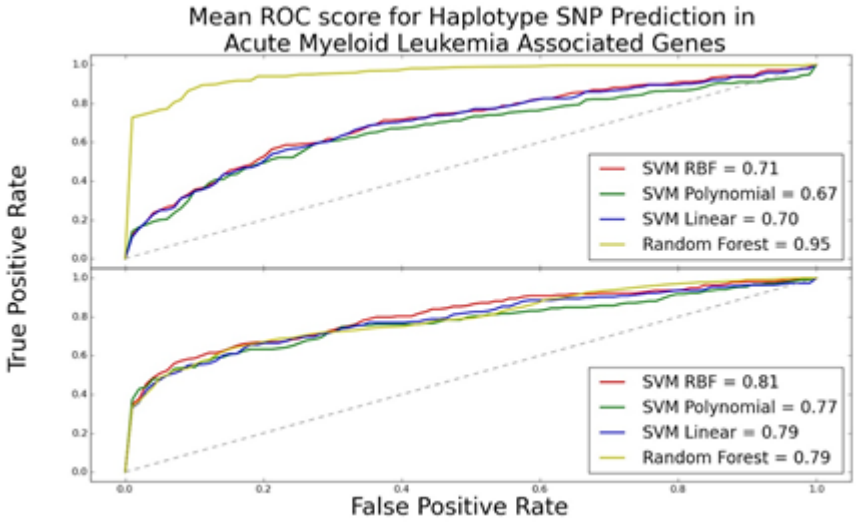


Figure 2.10: Mean ROC score plot for haplotype pair (upper panel) and individual SNP prediction (lower panel).

Again, a t-test was performed on ROC scores from the 10-fold cross validation of the two sets (haplotype and individual SNP) using Random Forest for haplotype pair versus individual SNPs. The p-value is 7.8×10^{-15}, confirming the statistical significance of the difference.

Overall, it suggests that random forest is the better classifier for most of the tasks performed in this study. An exception was observed for task 3, where SVM polynomial is providing better results. Moreover, in that module, the prediction performance is much lower than that of the other tasks. This change can be attributed towards the limited number of instances for each class. If we can get more SNP association with the interacting site, it might improve the results.

2.3 Identification of nsSNP for Comorbid Diseases

2.3.1 Introduction

We have already discussed the importance of SNPs location on protein sites, specifically the interacting sites and haplotype vs individual SNPs. As such, we know that interacting sites are considered as a hotspot for causing diseases. Therefore, we were interested in investigating the association of SNPs at the interacting site of the common genes of the comorbid diseases using the human interactome. This task will help us in identifying the role of SNPs in causing disorders, e.g., SNPs at the interaction sites interrupt protein-protein interaction that might be a common point in signaling pathways involved in the comorbid diseases.

2.3.2 Methods

The data, including Human interactome, disease gene association, network properties of disease pairs and comorbidity data, was used in the study from [22] and was downloaded from their website. There are 605 disease pairs in the dataset that have common genes associated with a disease pair. We used these 605 disease pairs for our study. In this project, we used a similar protocol to identify SNPs at interacting sites and non-interacting sites using supervised learning methods, as mentioned in section 2.2.

2.3.3 Results and Discussion

Our grid search for each gene-SNP association, specifically at interacting and non-interacting sites, was facing the same problem of gathering attributes for feature vector. The absence of attributes is due to unavailable biological data at different levels of data collection, such as genes associated with a SNP with existing 3D structure, as mentioned in the section 2.2. Gene collection for 605 disease pairs resulted in only 129

disease pairs with known 3D structures and associated SNPs at specific protein sites (interacting and non-interacting sites). Further, we found that there are only 37 disease pairs which have more than two SNPs associated. Due to this scarcity of data we did not perform our classification algorithm. The issue of 3D structure might be solved using structure prediction algorithms to construct our feature vector, but the reason here was the lack of missense SNP association with a gene, where our focus was on the common gene of a comorbid disease pair.

2.4 Conclusion

In this work, we carried out comprehensive comparative analysis for predicting SNPs effect associated with four types of cancers, in the context of SNPs being present at protein interacting sites versus non-interacting sites and being paired within a known haplotype versus being unpaired.

Our results confirm that prediction performance improved by using both sequential and structural features together than using them individually. Also, of the two types of classifiers used in the study, random forest outperforms in most cases.

It is found that generic SNP prediction provides better association of SNP to be detrimental or polymorphic SNPs as compared to disease-specific SNPs, although this conclusion does not hold if genes associated with one disease are unique from the other disease. While it is expected that prediction performance will be increased by associating SNPs to the interacting sites, the results show the instead slight decrease in prediction performance. This decrease in predicting accuracy may be caused by the small data set, as many affected proteins in the study do not have known interacting sites.

Compared to individual SNPs, these that appear together in haplotype showed a stronger correlation with one another and with the phenotype and therefore led to better prediction performance. Haplotype SNP prediction provided most promising results. These results could be taken to the next level of improving further accuracy and developing the personalized drugs ultimately. Although currently the haplotype classification and protein site classification were performed for only Acute Myeloid Leukemia, the same protocol can be adapted to perform a similar analysis on other diseases.

Lastly, while this study was performed on cancer diseases, the same protocol could be applied for the prediction of non-cancerous conditions to make this protocol generic for all types of diseases. In this context, we attempted prediction of SNP association with common genes of comorbid diseases. Despite of its promising additive useful information, currently data unavailability presents a major hurdle to construct feature vector for classification.

Chapter 3

PREDICTIONS OF THE MISSING COMMON GENES AND THE INTERACTIONS FOR COMORBID DESEASE PAIR

3.1 Introduction

The genetic cause of diseases is complex and complicated, and it can rarely be attributed to a single gene and its mutations. Instead, often, multiple factors are involved in the manifestation of disease symptoms. Furthermore, genes can take on more than one function, and different pathways and processes are intertwined and can crosstalk to one another. Therefore, typically one gene may be implicated in two or more diseases. Consequently, it is rational to examine not only the associated genes of one disorder to understand its pathology but also the overlap between the sets of associated genes of two diseases of high comorbidity risk to shed lights on the interplay of the two diseases [23, 24, 54]. The knowledge about comorbidity that can be gained from a list of genes, or their product proteins, would be quite limited if not aligned with its biological context, such as the signaling transduction pathways, regulatory and metabolic pathways in which they are involved.

This chapter has three major sections comprising of several contributions towards understanding what bring two diseases “closer”, by finding the missing components specifically using disease module separation and its predictive power. In section 3.2 we will discuss prediction of missing common genes using module separation. In section 3.3 we will represent two case studies to emphasize the

significance and role of biological pathways associated with genes of comorbid diseases to shed a new light that on the linkage or process which leads to cause two diseases at same time. Section 3.4 mainly discusses the prediction of missing interactions based on module separation.

3.2 Prediction of Missing Common Genes (Nodes) in Comorbid Diseases

This section presents a novel method to predict the missing common genes associated with comorbid diseases using the available information in the disease module. Our work starts out with the findings of disease module separation S_{AB} (explained later in 3.2.1) from [22] and explores its utility as a powerful indicator to determine comorbid diseases: smaller S_{AB} indicates that two selected diseases are closely located on the interactome, and hence may show comorbid behavior. It is critical to identify missing common genes to complete the set of genes associated with a disease and contribute towards completing the interactome. The method formulates searching for missing common genes as an optimization problem to minimize a network-based module separation between two subgraphs, formed by mapping the disease-associated genes onto the interactome. In Method section, we will give a complete description of our procedure. We will demonstrate the predictive power of our model, and we also compare our model for accurateness, and reliability by testing using a different method to calculate S_{AB} and with random data. Finally, we conclude by discussing the results and the significance of our approach.

3.2.1 Method

In this section, firstly, we will introduce the various concepts related to disease module on incomplete interactome. Especially a quantity S_{AB}, called module separation,

as given in [22], to measure the relationship between two disease modules A and B. Later, we will explain in detail our method of finding missing common genes for a given pair of disease, which is formulated as an optimization problem to minimize S_{AB}.

Disease Sub-graph on Interactome and Module Separation interactome contains all protein-protein interactions in the cell and can be conveniently represented as a graph (or network), in which proteins are represented as nodes and interaction between two proteins is represented as an edge connecting the two corresponding nodes. Reconstructing the interactome is a central task in systems biology, which studies the cell as a system in a holistic way instead of a simple ensemble of isolated items. Due to the limitation of the current technology, interactome for most organisms, even for model organisms, is incomplete, with missing proteins and their interactions. Nonetheless, the incomplete interactome can already provide valuable insights into many biological processes which cannot be obtained otherwise. In [22], it is shown how to uncover disease-disease relationships through the incomplete interactome. Diseases with genetic causes have been studied widely, often with a focus to identify the culprit gene only, to find that in many cases the reason cannot be attributed to a single gene; instead, it is prevalent that multiple genes involving in numerous cellular processes may be at play. Without putting these pieces in a more significant context, it is difficult to understand the pathological mechanisms fully. Work in [22] presents a systematic study to uncover disease-disease relationships by mapping the associated genes onto the interactome.

As mentioned by [22], given a pair of diseases A and B, the genes known to be associated with them are put into two separate sets G_A and G_B respectively. Let graph G be the interactome, with node set V, and edge set E. Let map the genes in G_A and G_B onto G with two different colors, say, nodes in G corresponding to genes in G_A are

colored red and nodes in G corresponding to genes in G_B are colored blue. For any shared gene, i.e., a gene is known to be associated with both disease A and disease B, then the corresponding node will be colored half red and half blue. Although all the red nodes are genes associated with disease A, indicating relatedness among them, they may not form a single connected component (or subgraph) of graph G of the interactome; often they form several sub-connected graphs. These disconnected graphs are may be due to either incompleteness of the interactome (i.e., missing edges) or unknown associated genes, or a combination of both. However, if the connected components are too fragmented, say not significantly different from what can be formed by randomly mapped genes, then it is difficult to infer valuable relationships reliably. So, in [22], the size of the largest connected component, as a percentage of the total number of genes associated to disease, must be maintained beyond a threshold, which is set based on percolation theory (explained in next paragraph) and the data used in the study. And the largest connected component, meeting the size requirement, is then called module as a representative for the disease. For example, multiple sclerosis (MS) has 69 known associated genes, and the largest connected component, which is qualified as a module with a size of 11, and rheumatoid arthritis (RA) has 51 associated gene and the largest connected component, which is qualified as a module with a size of 9.

According to percolation theory, explained by [22], if p fractions of links are available, then a connected component (disease module) of m nodes undergoes a phase transition under certain conditions. Therefore if p is above p_c^m , then a fraction of nodes will form an observable module but in case of lower p_c^m the module created will be too fragmented to be observable. [22] observed that statistically, if the number of genes

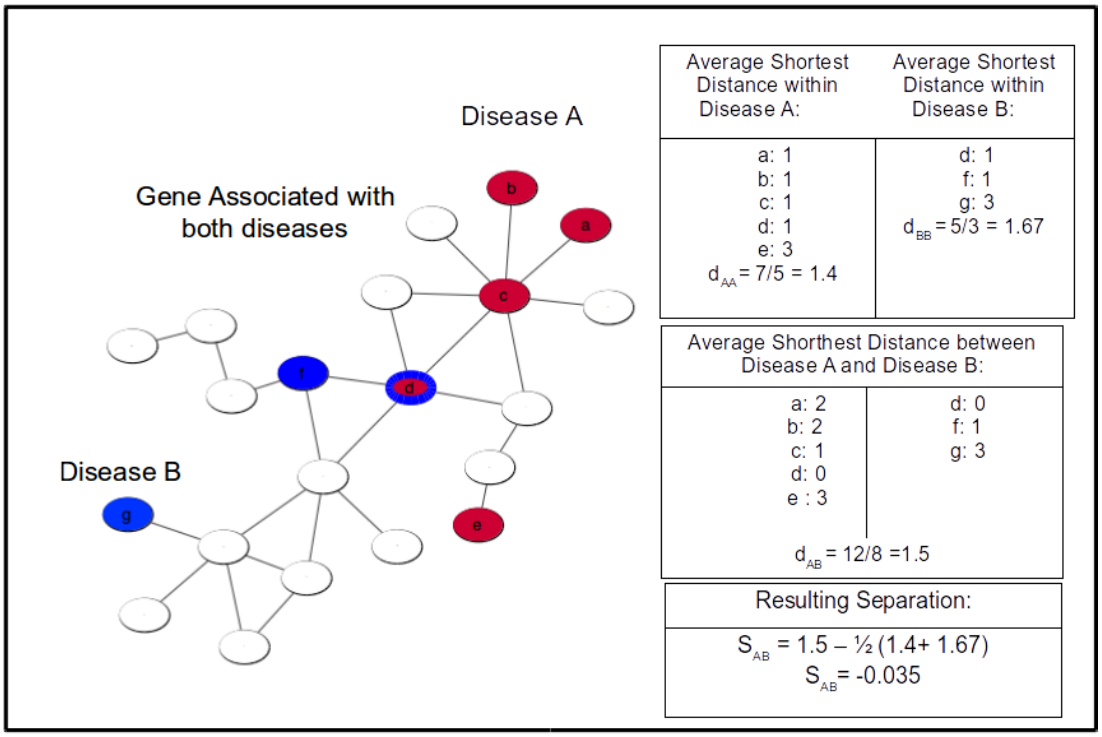


Figure 3.1: Illustration of network separation calculation using toy example

associated with a disease is less than 25 then the disease module formed will be too fragmented.

To uncover disease-disease relationships, a quantity called module separate S_{AB} is introduced as Eq (3.1).

$$S_{AB} \equiv \langle d_{AB} \rangle - \frac{\langle d_{AA} \rangle + \langle d_{BB} \rangle}{2} \qquad 3.1$$

where $<d_{AB}>$ is the average of the shortest distance for every gene of disease A to reach a gene of disease B and vice versa, $<d_{AA}>$ is the average of the shortest distance for every gene in disease A to reach another gene in disease A, and $<d_{BB}>$ the average of the shortest distance for genes of disease B to reach another gene in disease B.

Figure 3.1 shows how S_{AB} is computed for a toy example. Two diseases are considered correlated if the S_{AB} value is positive. Notably, if $Disease_A$ and $Disease_B$ are a complete set or identical set of genes, then the disease pair will have a strong comorbid relationship as shown by equation 3.1 where the average shortest distance between diseases d_{AB} will be zero.

More comprehensive results in [22] demonstrate that this network-based measurement of disease module separation is more indicative of pathological manifestations of disease pairs than simply measuring the overlap between the associated gene sets, such as Jaccard Index as shown in Eq (3.2):

$$J = |G_A \cap G_B| / |G_A \cup G_B| \tag{3.2}$$

It is reported in [22] that, when the disease history of 30 million individuals aged 65 and older is used to determine the relative risk RR of disease comorbidity for each disease pair, the relative risk drops from RR $\geq$ 10 for $S_{AB} < 0$ to the random expectation of RR $\approx$ 1 for $S_{AB} > 0$.

3.2.2 Detection of Missing Shared Genes

To further explore the predictive power of the disease module separation, we set in this work to use it to tackle the incompleteness of the data. Specifically, for disease pairs that are known to share high comorbidity. These disease pairs are expected to have a small, preferably negative module separation S_{AB} value, but instead, have large positive S_{AB}. We hypothesize that this discrepancy is due to some missing pieces of information, such as a missing shared gene, which if recovered should bring the two disease modules closer, i.e., to decrease S_{AB}. Therefore, we formulate the detection of missing common genes between two disease modules as an optimization problem given follows.

$$X^* = argmin\, S_{AB}[+x]$$
$$x \in |G_A \cup G_B| - |G_A \cap G_B| \quad (3.3)$$

where x goes over genes distinctly associated to either disease A or disease B, and $S_{AB}[+x]$ is the module separation when x is added as a shared gene between disease A and B, and x* is the predicted missing shared gene which minimizes the module separation.

The minimization can be achieved either by exhaustive search when the sets G_A and G_B are not very large or by some heuristics when the search space becomes huge. Note that, although Eq (3.3) is formulated for finding a single (most probable) missing common gene, in practice, this method can be applied sequentially multiple times for recovering multiple missing common genes. It is also worthwhile to note that the set of missing common genes recovered by using Eq (3.3) iteratively one gene at a time may likely be different from a set of missing common genes should their candidacy as common gene be evaluated altogether, possibly due to the topology of the interactome and how these genes are located. So, if the number of missing common genes k is known, an alternative formulation of the optimization problem can be defined as follows.

$$X^* = argmin\, S_{AB}[+x]$$
$$X \ |G_A \cup G_B| - |G_A \cap G_B| \quad (3.4)$$

where X* is the optimal set of missing common genes, and X is any subset of size k from the genes that are distinctly associated with either disease A or disease B. This formulation, while theoretically sound and appealing, has two practical issues: a) the number of missing common genes k is not known a priori; and b) the increased

computational complexity due to combinatorial in selecting k out n, where n = $|G_A \cup G_B| - |G_A \cap G_B|$. Because of these issues, we only tested Eq (3.4) for k = 2 and k = 3, while the results reported in the next section are mainly based on Eq (3.3).

3.2.3 Results and Discussion

In this section, we tested our method for identifying missing genes with the data used in [22]. We first describe the dataset briefly and then present the results which are evaluated using a cross-validation scheme.

3.2.3.1 Dataset

The data, including Human interactome, disease gene association, network properties of disease pairs and comorbidity data, was used in the study from [22] and was downloaded from their website. Comorbidity measured as relative risk (RR score) for several diseases using Medicare data from the USA has been calculated by [55]. [22] reported that RR score is computed by 13,039,018 patients diagnosed with one or more diseases over a period of 4 years using the following equation.

$$RR = n_{AB}.n_{tot}/(n_A.n_B) \quad 3.5$$

Where n_{tot} = total number of patients in the data, n_A, n_B = number of patients diagnosed with disease A and B, respectively and n_{AB} = number of patients diagnosed with both diseases A and B.

The goal of the method aims to find de novo missing common genes between a disease pair, for evaluation purpose, the method is tested, in a cross-validation scheme, at recovering known common genes. Therefore, a disease pair must have common genes

to be used in the test. Among 913 disease pairs with known comorbidity, there are 605 disease pairs that that fulfill the requirement of having common gene associated with them. The remaining 308 disease modules which either do not have any common gene or have all the genes common hence are removed from the test dataset.

3.2.3.2 Cross-validation and Performance

The cross-validation scheme is designed as follows. For a disease pairs A and B:

1. Select several common genes and reserve them as positive test examples.
2. Randomly select several non-common genes from G_A and G_B respectively, and reserve them as negative test examples.
3. For each gene x in the test set, run the search algorithm as given in Eq (3.3), and compute $S_{AB}[+x]$, the module separation when x is marked as shared, and x goes over all test examples associated with diseases A and B.
4. Then compute score $s(x) = S_{AB} - S_{AB}[+x]$.
5. Rank all the test examples x's by s(x) in a descending order: the higher the score s(x), the higher that x is ranked and hence more likely to be a common gene. ROC score is computed by comparing the ranked list and the ground truth of the test examples.

While applying the algorithm to the dataset we randomly selected 10 positive and negative examples and if a disease pair has less than 10 common genes then we choose all the available genes for our experiment.

The average ROC score for the dataset is 0.947, as reported in Table 3.1. When Eq (3.4) is used in place of Eq (3.3), where we selected set of genes instead of individual

gene, the average ROC score is 0.976 and 0.979 for k = 2 and 3 respectively shown Figure 3.2. This confirms that considering candidate missing common genes as a subset can indeed achieve better prediction as compared to considering candidate missing common genes individually, though the gain in performance seems to be tapering as the value of k increases.

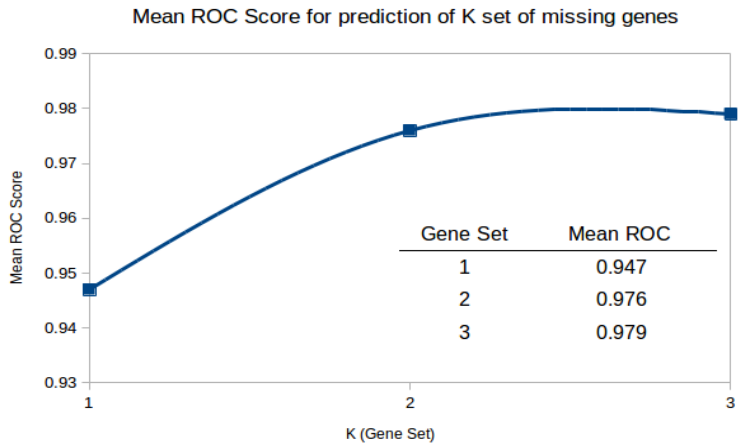


Figure 3.2: Mean ROC Score for prediction of K set of missing genes.

Table 3.1: Average ROC Scores with standard deviation for various comorbidity ranges as relative risk (RR). Higher RR means strongly comorbid

	Comorbidity Range				
	0-8000	**0-1**	**1-2**	**2-3**	**>3**
Number of Disease Pairs	**605**	**133**	**248**	**76**	**148**
Shortest distance ROC score	0.947 ±0.094	0.966 ±0.063	0.950 ±0.089	0.952 ±0.072	0.920 ±0.124
Average distance ROC score	0.491 ±0.279	0.513 ±0.279	0.495 ±0.288	0.508 ±0.269	0.458 ±0.269
Randomization ROC score	0.601 ±0.278	0.606 ±0.282	0.614 ±0.287	0.555 ±0.258	0.599 ±0.247
Shortest distance precision	0.88 ±0.27	0.88 ±0.28	0.85 ±0.31	0.89 ±0.25	0.96 ±0.15
Average distance precision	0.72 ±0.31	0.72 ±0.31	0.71 ±0.32	0.69 ±0.33	0.64 ±0.30
Randomization precision	0.66 ±0.29	0.70 ±0.28	0.63 ±0.29	0.66 ±0.30	0.72 ±0.29
Shortest distance recall	0.91 ±0.13	0.94 ±0.11	0.93 ±0.13	0.93 ±0.09	0.88 ±0.16
Average distance recall	0.69 ±0.30	0.72 ±0.28	0.70 ±0.30	0.70 ±0.31	0.64 ±0.30
Randomization recall	0.78 ±0.26	0.80 ±0.25	0.79 ±0.26	0.73 ±0.26	0.76 ±0.25

Table 3.1 also lists the average ROC score for several cases: a) disease pairs with comorbidity in [0,1], b) disease pairs with comorbidity in [1,2], c) disease pairs with comorbidity in [2,3], and d) disease pairs with comorbidity > 3.0, with case e) being all pairs included. It can be seen clearly that high average ROC scores are achieved for all cases, with case a) achieving marginally the highest. This finding is noteworthy as it suggests that S_{AB} is a useful indicator across all range of relative risk (RR) value whereas in [22] strong correlation was observed between RR drops and S_{AB} switching from negative to positive. Precision and recall reported in Table 3.1 are

computed using a threshold on prediction score s(x) which is set as suggested in [56] and explained next. Essentially, the threshold is set by using ROC curve on the test data to determine the highest peak point of ROC curve from the diagonal line, i.e., the prediction score of the test example that corresponds the peak point is used as the threshold. Average precision and recall are reported as 0.88 and 0.91 respectively for comorbid disease pairs using average shortest distance as method to measure module separation. Figure 3.3 represents a graphical representation of the evaluation metrics (roc score, precision and recall) used for two methods for calculating module separation and when used for randomized data.

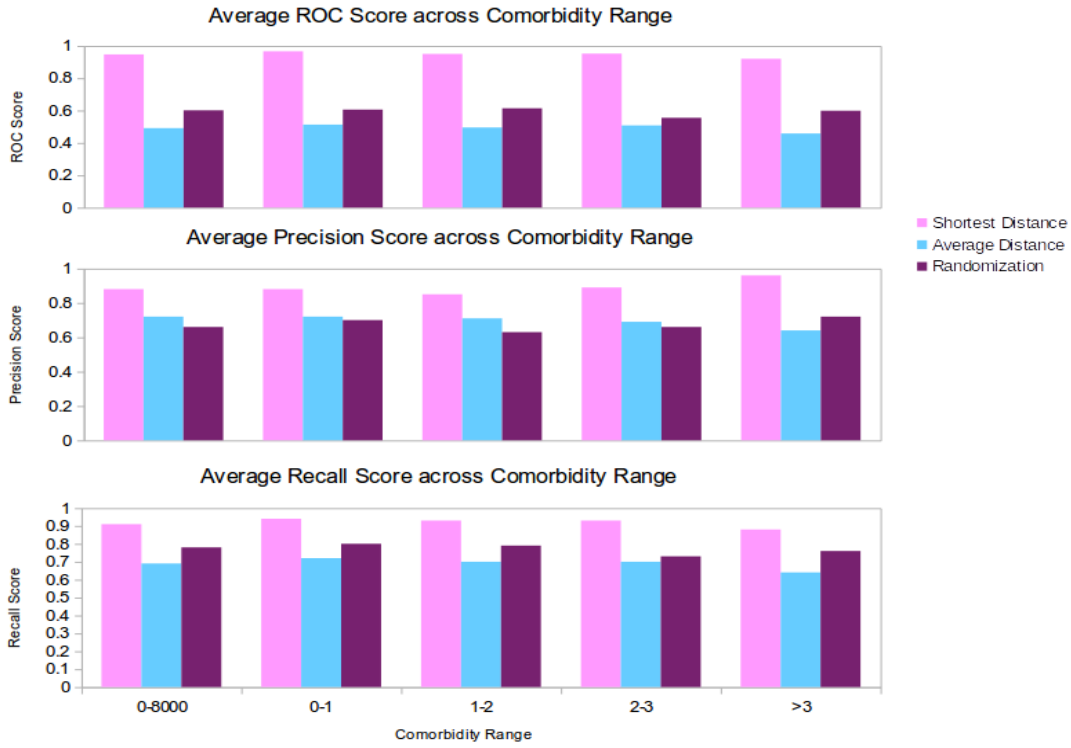


Figure 3.3: Bar chart for average ROC Score, average Precision and average Recall across comorbidity range.

In addition to the average ROC scores, the histogram plot of ROC scores is shown in Figure 3.4. In the histogram, a point in a curve shows in the vertical axis the percentage of disease pairs that have a performance greater or equal than ROC score given in the horizontal axis. It also shows the random ROC score in orange color.

We further examined how the prediction performance is affected by the number of common genes, i.e., the size of the training set. Specifically, we grouped disease pairs based on the range of overlap between associated genes: i) 5 ~ 10 common genes, ii) 10 ~ 15, and iii) 15 or more common genes.

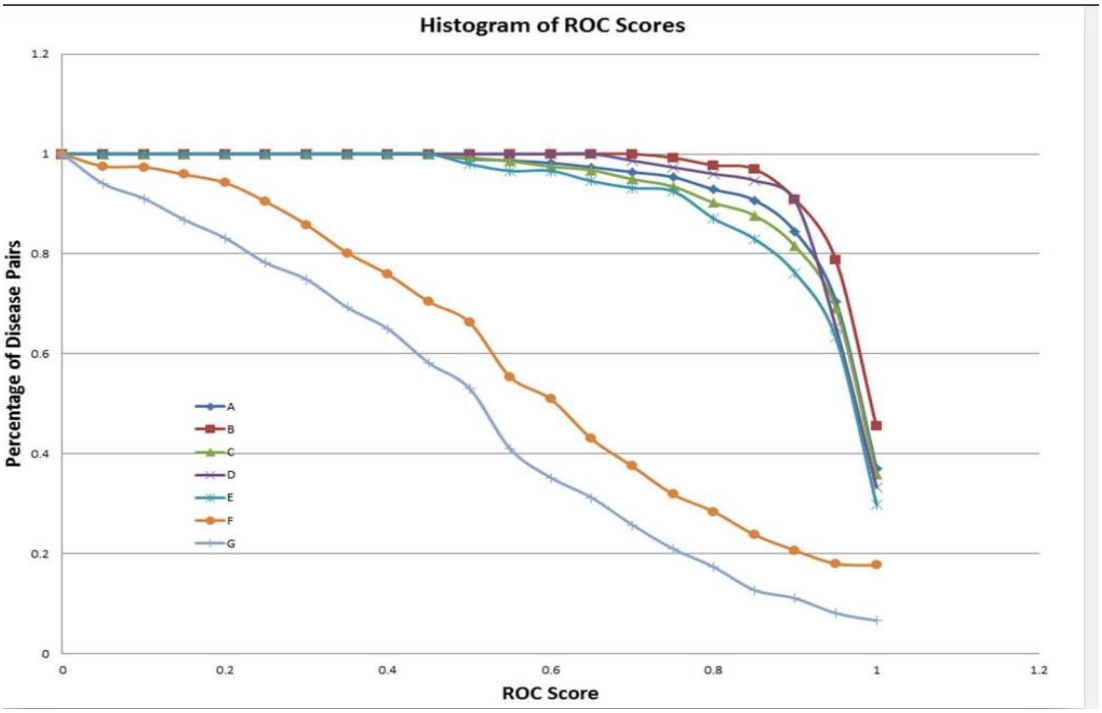


Figure 3.4: Histogram of ROC Scores. A: comorbidity range 0 ~ 1; B: comorbidity range 1~ 2; C: comorbidity range 2 ~3; D: comorbidity range > 3; E: comorbidity range 0 ~ 8000; F: randomized common genes; G: S_{AB} based on average distance.

The effect of the size of training set and the range of RR on prediction performance is reported in Table 3.2, which lists the number of disease pairs achieving

a given ROC score range for different groups under different RR range. For example, 42 pairs with 0~ 5 common genes and RR between 0 and 1.0 have received ROC score in the range (0.9, 1.0) The results show that as the number of common genes increases, the prediction performance in terms of distribution over various ranges is quite stable, with slight improvement, suggesting the method is robust under various conditions. In each case we had all the results above ROC score 0.5. And, more than 80% of the disease modules provide missing gene prediction ROC score between 0.9-1.

Table 3.2: Effect of the size of training set and the range of RR on prediction performance

ROC Score	**Comorbidity Range**				
Range	**0-8000**	**0-1**	**1-2**	**2-3**	**>3**
0 - 5 Common Genes					
0.5-0.6	2	0	2	0	0
0.7-0.8	5	2	3	0	0
0.9-1.0	174	46	81	18	29
Total	181	48	86	18	29
5 - 10 Common Genes					
0.5-0.6	0	0	0	0	0
0.7-0.8	2	0	2	0	0
0.9-1.0	121	36	48	15	22
Total	123	36	50	15	22
10 - 15 Common Genes					
0.5-0.6	0	0	0	0	0
0.7-0.8	1	0	0	0	1
0.9-1.0	46	12	21	4	9
Total	47	12	21	4	10
15 or more Common Genes					
0.5-0.6	10	1	3	0	6
0.7-0.8	24	1	6	4	13
0.9-1.0	220	35	82	35	68
Total	254	37	91	39	87

It should be noted that the missing common gene problem, despite its apparent importance, has not yet been addressed elsewhere in the literature to our best knowledge. Still, to get a sense how well the proposed method does in comparison to a

baseline, we randomize the common genes for each disease pair. Specifically, for each disease pair, the set of common genes is replaced with the same number of genes randomly selected from the whole set of genes in the interactome. The rationale for doing so is to keep the count of common genes for each disease pair and maintain the topology of interactome and the overall relative locations of the two diseases in the pair. When everything else was kept same, it was found that the average ROC score dropped to 0.601 for the 605 disease pairs with their common genes randomized. The detailed results for different comorbidity ranges with respect to the randomized baseline are listed in Table 3.2, and the histogram of ROC scores for the baseline is shown as plot F in Figure 3.4.

For comparison, we also modify method to calculate module separation. Specifically, instead of the shortest distances used in Eq (3.1), we replaced $<d_{AB}>$ with the average distance for all distinct A-B gene pairs, $<d_{AA}>$ is the average distance for all gene pairs within disease module A, and $<d_{BB}>$ the average distance for all gene pairs within disease module B. Using this modified module separation, let's call it all-pair-average based module separation S_{AB}, we get an average ROC score 0.49 for all 605 disease pairs. The histogram of the ROC scores is shown in Figure 3.3 as plot G. One plausible explanation of why the all-pair-average based module separation performs poorly is that the module separation has become much less sensitive to swapping a single gene x's classification in Eq (3.3) from common gene to non-common gene and vice versa.

From comparison to the baseline of randomized data and an alternative definition of module separation, the results show that our proposed method performs very well, suggesting the optimization formulated in Eq (3.3) as a viable solution to

finding missing common genes for a given pair of diseases. In this work, we used brute force to search all genes associated with the disease pair, as our focus is on the viability of using module separation to detect missing common genes, not on the speed. In the dataset, we used for this study, the average number of genes in a disease pair is 168, and it takes 2 min 43 sec to search all genes in the disease pair for putative common genes on a desktop computer: 2.90Ghz intel core i7, 8.00Gb memory. While it is desirable as future work to find a faster heuristic algorithm for search as the number of genes increases, the brute force approach seems to be acceptable for typical cases.

3.3 Case Studies

3.3.1 Introduction

Genetic disorders are caused by dysfunction of a specific gene, disrupting the function of the corresponding protein that is likely involved in some biological pathways and hence creating a disease. We performed a few experiments to investigate genes associated with a comorbid disease and its impact and role in fundamental processes to carry out proper functioning. We also display a few examples where our method predicted a gene associated with one disease from comorbid disease pair and it was ranked higher using the optimization method mentioned in previous section.

3.3.2 Discussion

To begin, we first mapped the common genes of comorbid diseases to biological pathways. We used Reactome database for this purpose [40] [41]. Reactome is an open source database, and it has information of about 2080 human pathways which incorporates 10374 proteins. Mapping the common genes of comorbid diseases onto biological pathways shows that, as expected intuitively, will have linear relation where

when the number of common genes for comorbid disease pair increases the number of pathways associated with that disease pair also increases. To understand this relationship better, we compared it to simulated data. That is, we randomly associated common genes to disease pairs, and then observed the ratio of pathway associated with disease in original and annotated data. Figure 3.5 shows the comparison histogram, displaying frequency of pathways for common genes in annotated vs. original data. This comparison suggests that common genes associated with comorbid disease pair may take effect in causing both diseases simultaneously, possibly in some "coordinated" way, via disrupting fewer pathways than random hit.

To further investigate we performed two case studies. We choose these two cases specifically because of the number of common genes associated, their phenotypic symptoms, network module separation (S_{AB}) and comorbidity (RR) value.

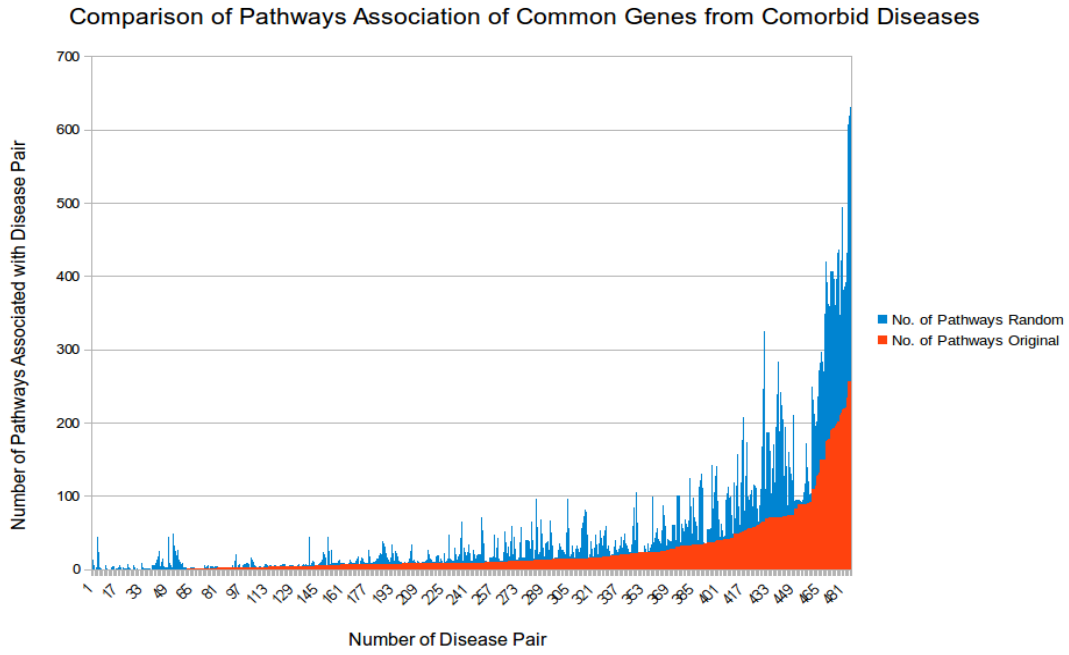


Figure 3.5: Common gene association with number of biological pathways for original and random common genes for comorbid diseases.

3.3.3 Coronary Artery Disease and Arterial Occlusive Disease

This disease pair has network separation value (S_{AB}) as -1.66 and relative risk factor (RR) as 2.04 and the average shortest distance of genes between two diseases d_{AB} as 0.51. Coronary artery disease (CAD) is a group of disorder that include: stable angina, myocardial infarction, and sudden cardiac death and is related to cardiovascular disease [57, 58]. The typical symptom is chest pain or discomfort which may travel into the shoulder, arm, back, neck, or jaw [58]. CAD happens when the arteries that supply blood to heart muscle become hardened and narrowed. While the Arterial occlusive disease (AOD) is a narrowing of the arteries other than those that provide blood to heart and brain [59]. The arterial occlusive disease most commonly affects the legs but can involve other arteries [59]. The most noticeable symptom is leg pain when walking which resolves with rest [59]. Rare diagnosis is due to skin ulcers, poor nail, bluish skin and hair growth in the affected leg [59].

There are 50 genes associated with coronary artery disease, and there are 62 genes related to the arterial occlusive disease. There are 43 genes shared between these two diseases. Figure 3.6 represents the gene association for the comorbid disease pair as venn-diagram. Genes in arterial occlusive disease and coronary disease are blue and yellow respectively. The shared genes for this disease pair are shown in green.

We performed a detailed study to get an insight of how biologically these genes are associated with the two diseases. We found that there are 54 distinct pathways where the genes related to this disease pair play their role. We also noticed that these genes were mostly from the shared gene set of the disease pair but in 9 different pathways, we found that there are genes associated to arterial occlusive disease only along with shared genes are performing their role. In this context, this disease pair becomes interesting to discuss in detail and display the evident to support our optimization algorithm to

pinpoint potential missing common gene. We will shed light on two pathways in detail. We consider these pathways significant to discuss due to two reasons: one, there is a non-shared gene, playing a role along with several shared genes and second, this uncommon gene is ranked higher while using our optimization method for searching missing common gene.

There are 13 different pathways where two shared genes: Collagen alpha-1 (IV) chain (CO4A1, gene ID:1282) and Collagen alpha-2 (IV) chain (CO4A2, gene ID: 1284) are performing their function. These genes mainly perform similar function, as they both possess anti-angiogenic and anti-tumor cell activity. It inhibits proliferation and migration of endothelial cells, reduces mitochondrial membrane potential, and induces apoptosis. These genes with another gene Apolipoprotein E (APOE, Gene ID: 348) associated arterial occlusive diseases play their role in pathway R-HSA-3000480 and it plays role in scavenging by class A receptors. Scavenger receptors are responsible for recognizing modified low-density lipoprotein [60] (LDL) by oxidation or acetylation. APOE plays roles in mediating the binding, catabolism of lipoprotein particles and can serve as a ligand for the LDL (apo B/E) receptor. Due to incompetence

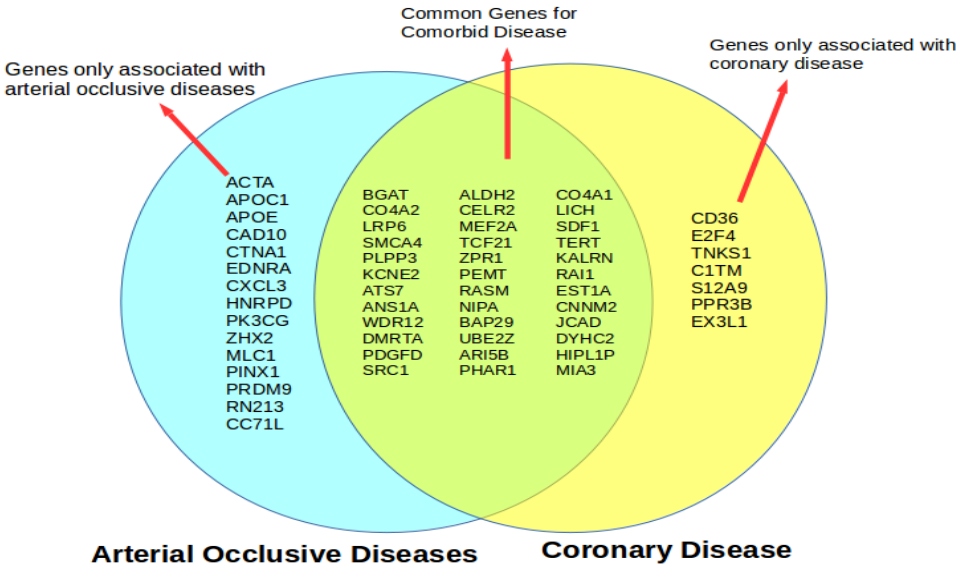


Figure 3.6: Gene association of arterial occlusive and coronary diseases

of these genes accumulation of LDL will become cause of not only occlusive diseases but also coronary diseases.

The gene APOE is also playing role in another pathway R-HSA-8957275 with one shared gene of disease pair. The common gene of the disease pair is transport and golgi organization protein 1 homolog (MIA3, Gene ID: 375056). The pathway under discussion has main role in Post-translational protein phosphorylation. This means that maturation of proteins after converting from gene to protein is done by this process. MIA3 plays a role in the transport of cargos that are too large to fit into COPII-coated vesicles and require specific mechanisms to be incorporated into membrane-bound carriers and exported from the endoplasmic reticulum. This gene is also specifically required for the secretion of lipoproteins.

These two pathways establish the fact that genes not only crosstalk and play role in different pathways but also have implication in causing multiple diseases simultaneously due to their interplay in several pathways. The detailed study of these two pathways reflect this finding and help in filling the dots as this gene is not only ranked higher but also cross-talk between two pathways critical for the disease pair. While considering APOE gene as common gene, the s_{AB} value is decreased to -1.7 with the d_{AB} value of 0.47. We conclude that if this gene is considered as shared gene instead then not only disease module will complete the missing information, but it also helps in drug designing considering both pathways and their function. Figure 3.7 shows these two pathways with the association of the genes.

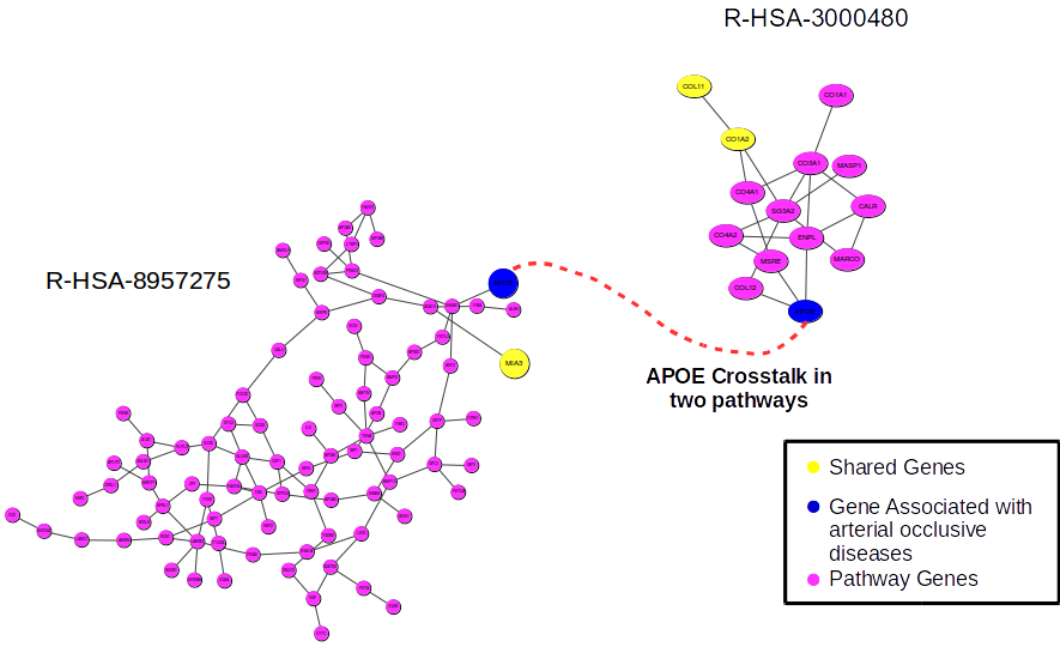


Figure 3.7:Pathways and gene association for comorbid diseases.

3.3.4 Diabetes Mellitus Type 1 and Lung Disease Obstruction

We have selected this disease pair due to the distinct symptomatic behavior. We emphasize the comorbid behavior of diseases due to their mutual gene sharing whether the phenotype of these disorders is different. The typical symptoms for lung disease include a dry cough and shortness of breath at rest or with exertion while the major symptoms of diabetes mellitus type 1 are increased thirst, frequent urination, extreme hunger and unintended weight loss.

Beside these entirely different phenotypes, both diseases have 13 genes in common, and these genes take part in 14 different pathways. We also found that there are several unique genes of this disease pair which are performing their functions together to complete one single process. We found that there are 25 different pathways which have no single common gene but have several genes associated solely with single disease and playing their respective role. There are 120 and 53 genes associated with diabetes mellitus type 1 and lung diseases respectively. It was found that these diseases are comorbid with network relative risk (RR) as 0.958 and the module separation is separation (S_{AB}) -0.00081 and average shortest distance between two diseases d_{AB} is 1.95. Figure 3.8 explains this comorbid disease pair in detail.

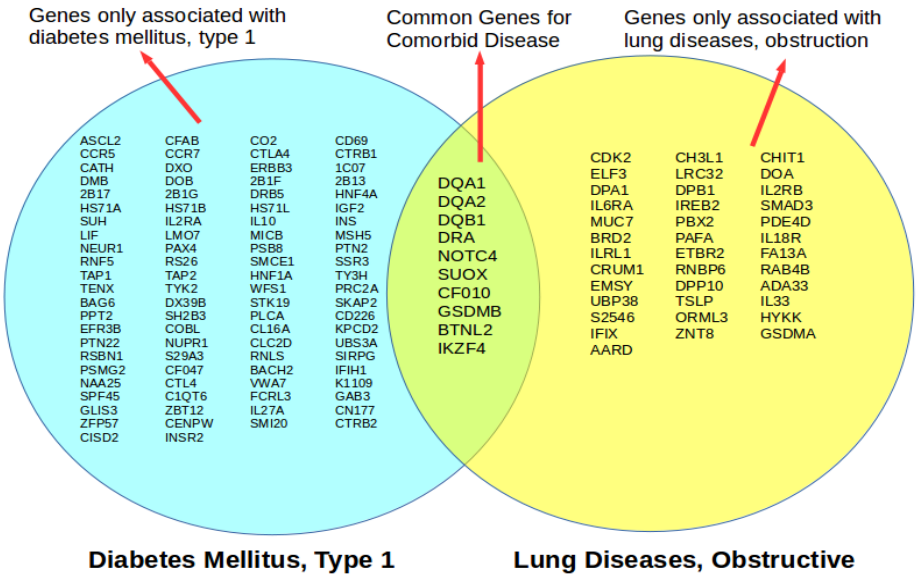


Figure 3.8:Gene association with diabetes mellitus, type 1 and lung diseases.

We will discuss a single pathway here and explain the integrated closeness of genes associated with each disease. There is one crucial pathway, R-HSA-202427, phosphorylation of CD3 and TCR zeta chains. This pathway is a chain of reaction to

activate T cell receptor (TCR) stimulation. T Cell Receptor (TCR) promotes some signaling cascades that ultimately determine cell fate through regulating cytokine production, cell survival, proliferation, and differentiation. Dysfunction in this pathway can disrupt the cell cycle eventually causing several diseases at a time. We found that this pathway shows a strong association between disease pair diabetes mellitus type 1 and Lung disease obstruction. There are four genes which are common and lie on the same pathway from the disease pair. These genes are: HLA class II histocompatibility antigen, DQ Alpha 1 chain (HLA-DQA1, gene ID: 3117), HLA class II histocompatibility antigen, DQ alpha 2 chain (HLA-DQ2, gene ID: 3118), HLA class II histocompatibility antigen, DQ beta 1 chain (HLA-DQB1, gene ID: 3119), and HLA class II histocompatibility antigen, DR alpha chain (HLA-DRA, gene ID: 3122). All these genes assist in binding peptides derived from antigens that access the endocytic route of antigen presenting cells (APC) and presents them on the cell surface for recognition by the CD4 T-cells.

In addition to these common genes we found four genes solely associated with Diabetes Mellitus Type 1 namely: HLA class II histocompatibility antigen, DQ beta 2 chain (HLA-DQB2, gene ID: 3120), HLA class II histocompatibility antigen, DRB1-15 beta chain (HLA-DRB1, gene ID: 3123), HLA class II histocompatibility antigen, DR beta 5 chain (HLA-DRB5, gene ID: 3127) and Tyrosine-protein phosphatase non-receptor type 22 (PTPN22, gene ID: 26191). The first three genes have a similar function like its sister genes to present peptide on the cell surface for recognition while the last gene acts as negative regulator of the T-cell receptor (TCR) signaling by direct dephosphorylation of the SRC family kinases.

There are also two genes associated with Lung diseases that play their role in this pathway. These genes are HLA class II histocompatibility antigen, DP alpha 1 chain (HLA-DPA1, gene ID: 3113) and HLA class II histocompatibility antigen, DP beta 1 chain (HLA-DPB1, gene ID: 3115) from lung disease obstruction which is associated with the same pathway. These genes also help in transportation of peptide on the cell surface for recognition.

All these genes perform their function in cascade, and therefore disruption of one gene can lead to dysfunction of the whole process. Since these genes are associated with two different diseases regardless of their phenotype, these diseases have a high rate of occurring together. Also, notice that the RR value is below one, but still a strong association of these diseases can be seen through their genes intervention in pathways. Figure 3.9 represents the gene association for this pathway. Our optimization method

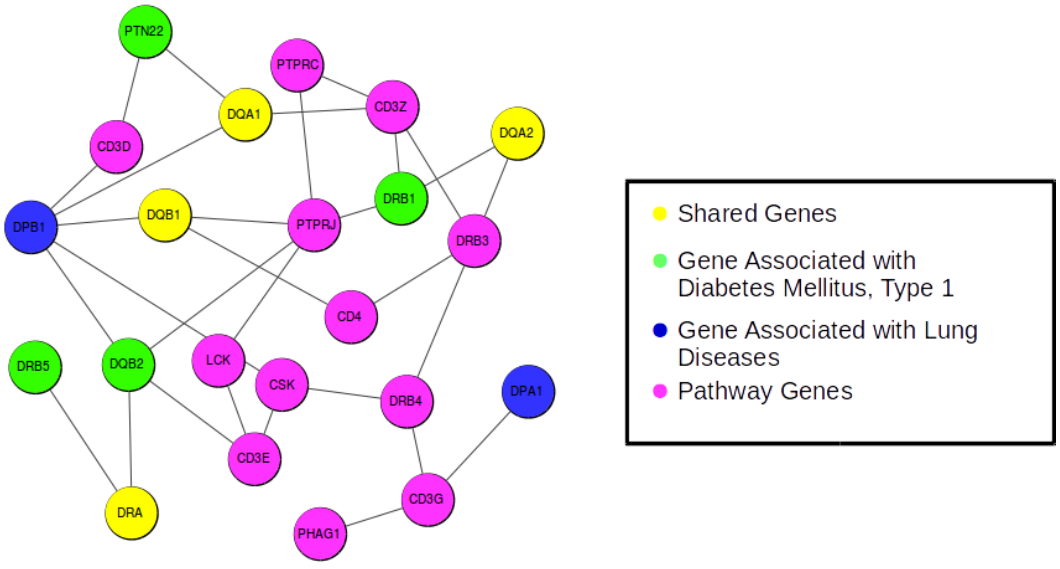


Figure 3.9: Pathway association with Diabetes mellitus, type 1 and Lung disease obstruction.

for finding missing common genes also rank few of these genes higher and while

considering each of these genes as a common gene we found that module separation is decreased for all genes except for one. And if we consider all non-common genes discussed above as common genes for the disease pair then there is a significant decrease in the average shortest distance between diseases from d_{AB} 1.95 to 1.80 as well as module separation s_{AB} from -0.000806 to -0.122.

3.4 Prediction of Missing Common Interactions (Edges)

3.4.1 Introduction

In this section, we will discuss our implementation and results for prediction of missing interactions to contribute towards completion of the human interactome. For this purpose, we used the same algorithm as described for prediction of missing common genes.

Prediction of missing interactions associated with comorbid diseases is equally significant as this will not only contribute towards the more substantial task of completing interactome but also help in achieving the goal of complete genes associated with a set of the interactions related to the comorbid disease.

3.4.2 Method

We formulate the detection of missing common edges between two disease modules as an optimization problem given follows.

$$e^* = \underset{e}{\operatorname{argmin}}\, S_{AB}[+e] \qquad (3.5)$$

$$|E_A \cup E_B| - |E_A \cap E_B|$$

where e goes over interactions (edges) from the human interactome available which are distinctly associated with either disease A or disease B, and $S_{AB}[+e]$ is the module separation when e is added as a shared edge between disease A and B, and e* is the predicted missing shared edge which minimizes the module separation.

3.4.3 Results and Discussions

We tested our method for identifying missing interactions with the data used in [22]. We first describe the dataset briefly and then present the results which are evaluated using a cross-validation scheme.

3.4.3.1 Dataset Used for Missing Interaction Prediction

The data included human interactome, disease gene association, network properties of disease pairs and comorbidity (RR score) for disease pairs from [22] and was downloaded from their website. Comorbidity (RR score) for several diseases using Medicare data from the USA has been calculated by [55]. We started with a smaller data set by selecting 20 different disease pairs having comorbidity range between 0-1.

While the algorithm is ultimately aimed at finding de novo missing common interactions between a disease pair, for evaluation purpose, the method is tested, in a cross-validation scheme, at recovering known common interactions.

3.4.3.2 Cross-validation and Performance

The cross-validation scheme is designed as follows. For a disease pairs A and B:

1. Select several common genes and delete interactions (now onward called as: edges) from these genes. Such that in the edge (v,u) either v or u belongs to common gene set, reserve them as positive test examples.

2. Randomly, select several common genes and add an edge from these genes. Such that in the edge (v,u) either v or u belongs to common gene set, and reserve them as negative test examples.
3. For each edge e in the test set, run the search algorithm as given in Eq (2.5), and compute $S_{AB}[+e]$, the module separation when e is marked edge added as shared, and e goes over all test examples associated with diseases A and B. Then compute score $s(x) = S_{AB} - S_{AB}[+e]$.
4. Rank all the test examples e's by s(e) in a descending order: the higher the score s(e), the higher that x is ranked and hence more likely to be a common edge. ROC score is computed by comparing the ranked list and the ground truth of the test examples.

3.4.4 Discussion

We used smaller data set consisting of 20 disease pairs to test our formulated method. These disease pairs are listed in Table 3.3 along with ROC score.

Table 3.3: Selected Disease pairs for missing interaction (edge) prediction with their ROC score

Disease Pairs		ROC	RR	S_{AB}
Endocrine system	Immune system	0.50	1.02	-0.32
Exophthalmos	Thyroid diseases	0.44	1.02	-1.26
Crohn disease	Demyelinating diseases	0.57	0.97	-0.01
Exophthalmos	Glomerulonephritis	0.25	0.99	-0.19
Cerebellar ataxia	Leukemia, b-cell	0.50	1.02	-0.01
Dementia	Parkinsonian disorders	0.50	1.02	-0.19
Ankylosis	Celiac disease	0.50	1.03	-0.29
Glomerulonephritis	Sleep disorders	0.25	0.96	-0.09
Breast neoplasms	Ovarian neoplasms	0.50	1.01	-0.08
Adenocarcinoma	Goiter	0.17	0.99	-0.02
Cholestasis	Rheumatic diseases	0.63	1.01	0.00
Bechet syndrome	Spondyloarthropathies	0.50	1.00	-0.22
Arrhythmias, cardiac	Death, sudden	0.50	1.00	-0.11
Bechet syndrome	Spondylitis	0.50	1.00	-0.22
Hyperthyroidism	Uveitis	0.25	1.00	-0.35
Ankylosis	Multiple sclerosis	0.50	1.02	-0.09
Bone diseases, metabolic	Metal metabolism, inborn errors	0.44	1.02	-0.58
Head and neck neoplasms	Stomatognathic diseases	0.40	1.00	-0.24
Diabetes mellitus type 1	Respiratory tract diseases	0.50	0.97	-0.08
Diabetes mellitus type 2	Pancreatic diseases	0.50	1.02	-0.04

We found that average ROC score for these 20 pairs is 0.44. This low performance can be explained by the topology of network. Adding/deleting only one edge from the network does not affect its total performance as compared to adding/removing gene from the network.

Based on this result we considered to reformulate our method. We increase the number of edges removed/added for each disease pair by using the Eq (3.6) as given below. Consider the number of missing common edges as: k is known then,

$$E = argmin\, S_{AB}[+E] \qquad (3.6)$$
$$E(E_A \cup E_B) - (E_A \cap E_B)$$

where E* is the optimal set of missing common edges, and E is any subset of size k from the edges that are distinctly associated with either disease A or disease B. We used cross validation method with increasing the number of edges added/deleted. This method provided results as shown in Table 3.4. The average ROC score is calculated for 20 disease pairs given in Table 3.4.

Table 3.4:Average ROC to recover K set of common Interactions.

K (Interaction)	**Average ROC Score**
1	0.44
2	0.39
3	0.45
4	0.32
5	0.46

The approach to use a set of edges instead of singleton edge was also unable to provide promising outcome. This effect could be explained by the sub-graph/ module associated with one disease. We know that disease module is not a connected graph, but it consists of several sub-graphs and even singletons. By adding/removing edges from module might not bring nodes close enough to make them comorbid and result in such lousy performance.

3.5 Conclusion

In this work, we developed a novel method to predict missing common genes for a given disease pair. The algorithm formulates the task as an optimization problem of minimizing network-based module separation for subgraphs formed by associated genes on the interactome, with the hypothesis that correctly identified missing common genes would bring the two-module closer. The results of cross-validation from a benchmark dataset of more than 600 disease pairs show high prediction accuracy on average, measured as ROC score. The method provides a useful tool to infer a better understanding of disease-disease interaction regarding related genes. While the methodology is tested in cross-validation mode in this study, it can be easily deployed to predict de novo missing genes, i.e., those genes that are not associated with any disease but have an impact on the phenotype of both disorders. We also showed that biological pathways are associated with genes associated to specific disease. Disruption in such pathway can be a plausible cause of comorbid disease in human. To sum up, we have found that genes associated with comorbid disease pair have some underlying biological mechanism which when get disrupted, becomes the reason of causing multiple diseases. We have seen that one pathway can be associated with one or more

diseases and therefore, one dysfunction of pathway may lead to one or more diseases in human simultaneously. Lastly, we also tried to predict missing interactions, but our method is not providing promising results for this area of research.

Chapter 4

COMORBID DISEASE PREDICTION WITH GEOMETRIC SPACE EMBEDDING

4.1 Introduction

In this chapter, we develop a new model to predict comorbid diseases for large dataset. Specifically, inspired by the correlation between the disease module separation S_{AB} and comorbidity in [19], our method exploits the idea of embedding the PPI network into a high dimensional geometric space to better characterize and incorporate interactome structural information to distinguish comorbid diseases from non-comorbid diseases. Instead of using module separation as a sole means to directly predict comorbidity, our method first projects disease module into various dimensions to "fingerprint" the module and then trains a classifier to discriminate comorbid disease pairs from non-comorbid pairs. We use human interactome to collate the proteins associated with diseases of interest and then we transform the PPI network into a high dimensional space, which we believe can offer multiple perspectives to capture the relative positioning of disease modules. Furthermore, we include gene-disease association information from GWAS and OMIM as reported in [22]. We then apply supervised machine learning techniques to train a classifier to differentiate comorbid diseases from non-comorbid diseases based on the features/fingerprints extracted from the high dimension space. Our dataset comprises of 10743 disease pair for classification having known gene-disease association and comorbidity values, making it a much larger dataset for the given problem than previous studies.

4.2 Methods

4.2.1 Overview

We consider PPI network as a graph G= (V, E) where V is a set of nodes and E is a set of edges. The graph is consider as connected if for all pairs of nodes there is a path between them comprise of edges from E. In general PPI networks are comprised of several subgraphs with usually one large connected component which includes maximum information in term of proteins and their interactions. For example, we use human interactome in this study provided by [22] which has 13460 proteins in total, and the largest connected component has 13329 proteins which comprise 99% of the total proteins in the network. In this study, we use only the largest connected component, due to the limitation of embedding in geometric space where disconnected components of a graph when converted into low dimensional space may result in undefined spatial overlap. We started with embedding the largest connected component of a PPI network into low dimensional space to compute spatial distances between the embedded nodes.

4.2.2 The Embedding Algorithm

We used the embedding algorithm which is based on Multi-Dimensional Scaling (MDS) [61]. MDS is a spectral method based on eigenvalues and eigenvectors. It is a classical nonlinear dimensionality reduction algorithm, and it is based on Euclidean distance. Since human interactome is represented as a graph where coordinates of nodes are unknown, therefore an extension called isometric feature mapping based on geodesic distance is applied to PPI networks [62]. Geodesic distance is the distance between two vertices in a graph calculated as the number of edges in a shortest path connecting

them. The only difference between MDS(Classical) and MDS (Isomap) is the distance matrices used; otherwise they are equivalent. We choose MDS (Isomap) embedding technique for our method due to its ability to calculate distance for a graph.

The primary processing of Isomap is such that: Given a set of n nodes and a distance matrix whose elements are shortest paths between all node pairs, find coordinates in a geometric space for all the nodes such that the distance matrix derived from these coordinates approximates the original geodesic distance matrix to its possible extent. Detailed procedure for embedding task given below and also explained in:

1. Construct a network of using protein-protein interactions.
2. Find the largest connected component and compute minimum spanning tree
3. Compute the shortest paths of all node pairs to form matrix D
4. Get a symmetric, positive semi-define matrix by applying double centering to matrix D: $A = -\frac{1}{2}JD^2J, J = I - n - 1, 11'$ **J** = **I** − $n-$**11**′, where **I** is the identity matrix that has the same size as **D**; and **1** is a column vector with all one, and **1**′ is the transpose of **1**.
5. Extract the *m* largest eigenvalues $\lambda_1 \ldots \lambda_m$ of **A** and the corresponding *m* eigenvectors $e_1 \ldots e_m$, where *m* is the dimensions of target geometric space.
6. Then, a *m*-dimensional spatial configuration of the *n* nodes is derived from the coordinate matrix $X = E_m \Lambda_m^{1/2}$, where **E***m* is the matrix with *m* eigenvectors and Λ*m* is the diagonal matrix with *m* eigenvalues of **A**.

Geometric embedding is explained in Figure 4.1 using a toy example. There are several other embedding algorithms, such as Stochastic Neighbourhood Embedding (SNE) [63] and tSNE [64], Minimum Curvilinearity Embedding (MCE), non-centered MCE (ncMCE) proposed by Cannistraci et al.[65, 66]. The comparison between them

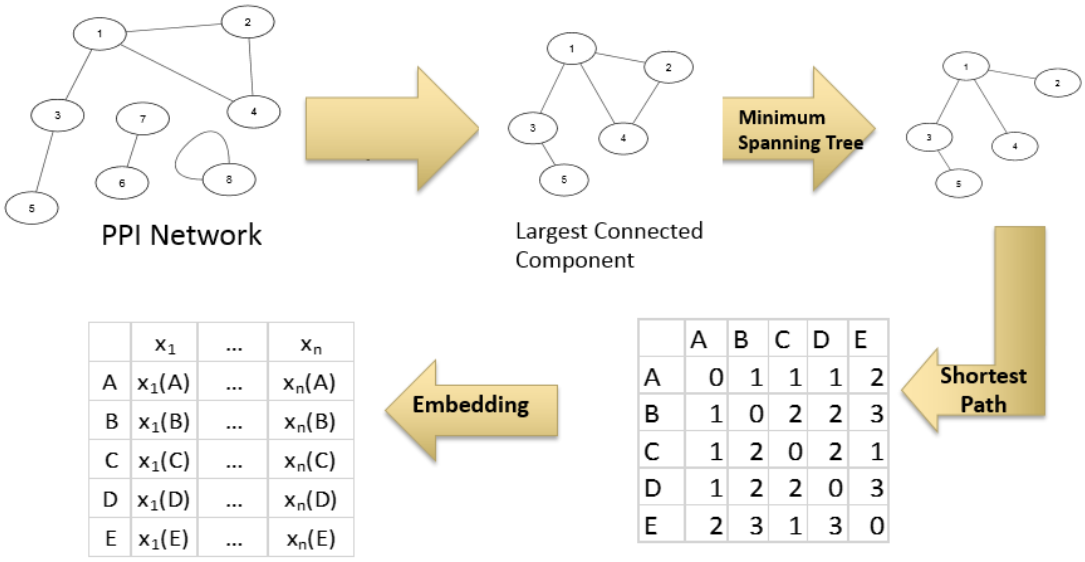


Figure 4.1: Process to compute geometric embedding using toy example.

is beyond the scope of our project, i.e., find a better way to predict comorbid diseases. Therefore, we used most recent MCE [66], ncMCE [65], and a method proposed by Kuchaiev et al.[67].

4.2.3 Disease Comorbidity Prediction

Our comorbidity prediction method exploits the key idea that a high dimensional geometric space provides multi facets (or angles) to capture and characterize the proteins' relative positions in the interactome and hence makes it easier to distinguish the comorbid diseases from non- comorbid diseases by the distribution of the associated

proteins on the interactome. The steps developed to implement this idea are given as follows and schematically represented in Figure 4.2:

1. Embed the human interactome network into a geometric space of dimension m, and extract feature vectors.
2. Choose a threshold for comorbidity
3. Train the data using a supervised learning classifier such as Support Vector Machine or Random Forest
4. Test the model for disease comorbidity prediction.
5. Evaluate the model using several evaluation metrics

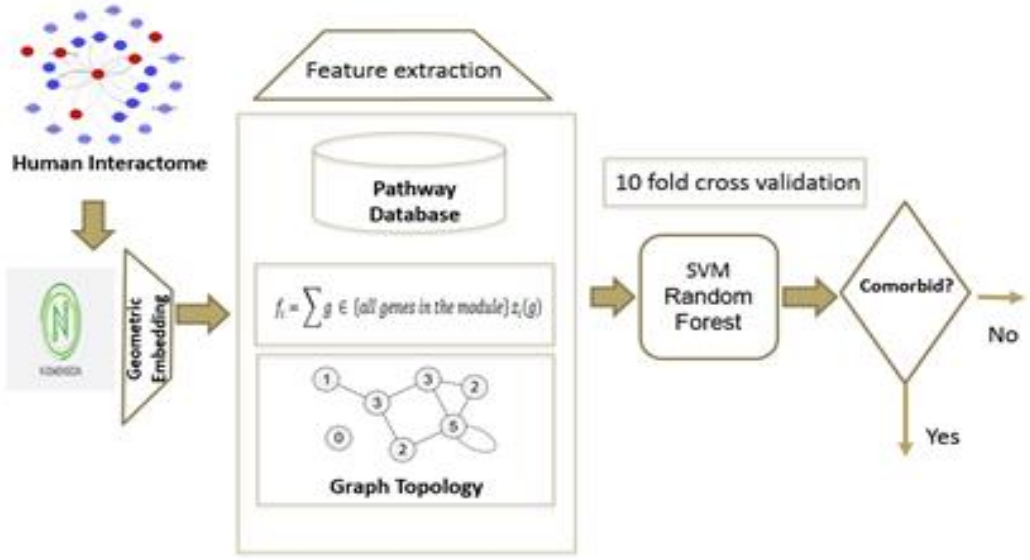


Figure 4.2: Schematic form of algorithm to predict a disease pair as comorbid or non-comorbid disease.

4.2.4 Classification

As mentioned above, we formalize the prediction of comorbid disease as a classification problem and adopt supervised learning approach. Specifically, this problem is considered as a binary classification problem where either a disease pair is comorbid or non-comorbid, corresponding to the output y of the binary classifier, where y = 1 for comorbid disease pair and 0 for a non-comorbid disease. The classifier is to learn the actual mapping from input vector x to output: y = F(x), with a hypothesis function H (x, ɵ) as explained in section 2.2 in equation 2.1. Once the classifier is trained, it is used to make prediction/classification on unseen data, i.e., disease pair whose comorbid property is not known a priori. In this study, two powerful classifiers, Random Forest [43] and Support Vector Machines [44], are selected as explained in 2.2.

For SVM, three kernel functions were adopted and assessed: Linear, Radial Basis Function, where the parameter $C = 3.5, \gamma = 1.06$ and Polynomial where the degree $d = 4$. These values of C and degree of polynomial d were optimized by using Opunity 1.1.1, a python package.

4.2.5 Data and Feature Characterization

The dataset used in this study is adapted from [22], which consists of 10743 disease pairs with comorbidity measured as relative risk RR based on clinical data; RR>1 for a disease pair indicates that the diseases are diagnosed more often in the same patients that expected by chance given their prevalence. This comorbidity value is considered as ground truth to determine disease pair and their association regarding comorbidity.

We used various values of geometric space of m for this study. Therefore, the feature vector for this study is comprised of m+3 features in total. The feature vector for

any disease pair module includes m features from the geometric space <f_1, …, f_i, …, f_m>, where f_i is the projection of the disease pair onto the i-th dimension, i.e., the sum of i-th coordinate z for all genes in the given disease module.

$$f_i = \sum g \in \{all\ genes\ in\ the\ module\}\ z_i(g) \tag{4.4}$$

where z_i (g) is the i-th coordinate z of gene g and the other three features are:

1. The average degree of nodes by calculating the number of edges connecting to each node. We calculated the average of all the proteins associated with a disease pair.
2. Average centrality used to measure how often each graph node appears on the shortest path between two nodes in the graph. Since there can be several shortest paths between two graph nodes s and t, the centrality of node u is:

$$c(u) = \sum_{s,t \neq u} \frac{n_{st}(u)}{N_{st}}$$

Where $n_{st}(u)$ is the number of shortest paths from s to t that pass-through node u, and N_{st} is the total number of shortest paths from s to t. We computed the average of all the nodes associated with both diseases taking part in disease pair under consideration.

3. The average number of pathways associated with genes of a disease pair. This pathway count is collected from Reactome database [68, 69].

Reactome is an open source database and contains information of about 2080 human pathways which incorporates 10374 proteins.

4.2.6 Cross-Validation and Evaluation

To assess the prediction performance, we adopt the widely accepted cross-validation scheme. Specifically, we used 10-fold cross-validation. We used some commonly used measurements to indicate the performance, which includes accuracy, precision, recall, F1 score, and ROC score. These methods are explained in detail in section 2.2.5.

4.3 Results and Discussion

4.3.1 Dataset

The data used for this study including the human interactome, disease gene association and comorbidity values RR is adapted from [22]. The dataset contains 10743 disease pairs. We used comorbidity values computed and reported in [55] for the classification purpose. Comorbidity RR value ranges from 0 to 9000 for our data. There are 6269 disease pairs with comorbidity value RR >=1, which is more than 50% of our dataset.

Among these disease pairs, there are 1868 disease pairs with comorbidity value RR= 0, comprising 17% of the dataset. The other disease pairs are spread out to the max RR=8861.6 and there are only 854 disease pair with comorbidity value > 4. Also to setting RR = 1 as the comorbidity threshold like in [22], in this study we even tested with a relaxed threshold at RR = 0, namely, any disease pairs with non-zero RR value is considered disease pairs and only these pairs with zero RR value are considered non-

comorbid. So correspondingly we prepare two sets of training and testing data (Comorbidity_0 and Comorbidity_1) to evaluate the performance of our method.

4.3.2 Geometric Space

The first crucial task of our method is to embed the interactome into a geometric space of dimension m. Initially, we picked only one geometric embedding method and tested with different dimension space values from m=2 to m=13 [67]. We used Kuchaiev et al. [67] and noticed that as the dimension increases, the prediction performance ROC score roughly increases as well. The Table 4.1 and Figure 4.3 represents the details about the performance where we can see that after dimension 8 random forest have consistent performance. SVM RBF, our best classifier has increased performance until dimension 12, and later it starts decreasing. This method uses a subspace iteration to compute eigenvalues, but this method as reported [67] has time complexity issue with a graph having edges >20000.

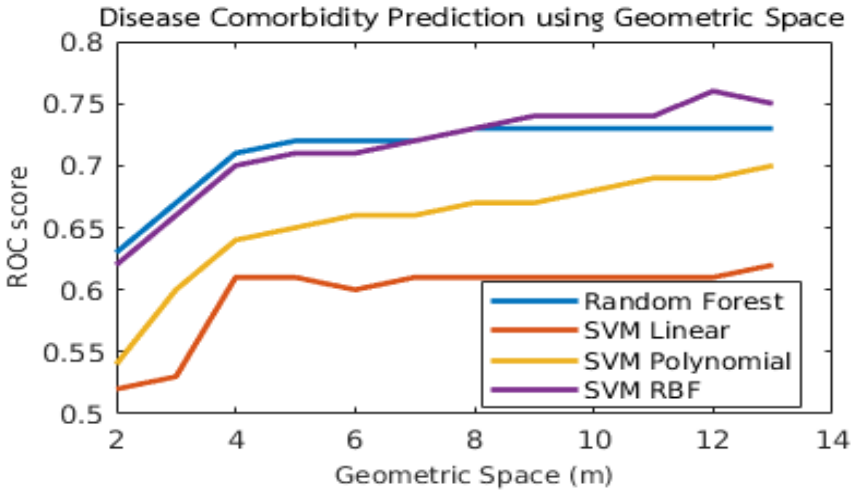


Figure 4.3: ROC score vs Geometric space to show comorbid disease prediction.

Table 4.1: Geometric space performance using different dimension values

Dimensions	Random Forest	SVM RBF	SVM Polynomial	SVM Linear
2	0.63	0.62	0.54	0.52
3	0.67	0.66	0.60	0.53
4	0.71	0.70	0.64	0.61
5	0.72	0.71	0.65	0.61
6	0.72	0.71	0.66	0.60
7	0.72	0.72	0.66	0.61
8	0.73	0.73	0.67	0.61
9	0.73	0.74	0.67	0.61
10	0.73	0.74	0.68	0.61
11	0.73	0.74	0.69	0.61
12	0.73	0.76	0.69	0.61
13	0.73	0.74	0.70	0.62

These results indicate that there is a linear relationship between dimensionality and prediction performance using ROC score. We have also noticed that increasing dimension space more than m=8, random forest and SVM linear does not improve performance. We encountered that by increasing the dimension space more than m=13 the computational time increases drastically. Therefore, we also used several other embedding methods to compare the performance regarding prediction performance and time complexity. We used three more embedding algorithms to compute high dimensional space from MCE [66], ncMCE [65] and MDS. For this purpose, we calculated dimension space at m=13 and performed the experiment the ROC score for each classifier is given below in Table 4.2.

.

Classifier	Centered MCE	Non-centered MCE	MDS	Kuchaiev et al.
Random Forest	0.73	0.73	0.72	0.73
SVM_RBF	0.74	0.74	0.73	0.74
SVM_Polynomial	0.70	0.69	0.69	0.70
SVM_Linear	0.60	0.59	0.60	0.62

Table 4.2: ROC Score for several geometric embedding algorithms

Since MCE, nMCE and Kuchaiev et al methods provided us the equivalent performance we selected MCE for our further analysis. We then attempted to increase the dimension space up to 50 to observe the behavior, and it turned out that the improvement in performance is consistent with 20 dimensions. except SVM using linear kernel but still the performance of SVM_linear is not better than other classifiers as shown in Table 4.3 and Figure 4.5. This linear relationship of dimensional space and SVM using linear kernel suggest that by increasing the dimensions eigenvalues for each node gets so close that the topology of graph tends to be linear. Therefore, for the rest of the study, we selected dimension space m=20 using MCE geometric embedding method. The Figure 4.4 shows the distribution of positive and negative examples of the training set at selected dimensions. We can see that there is not a single dimension standing out to contribute solely to comorbid behavior, but it is a respective contribution of every individual projection angle.

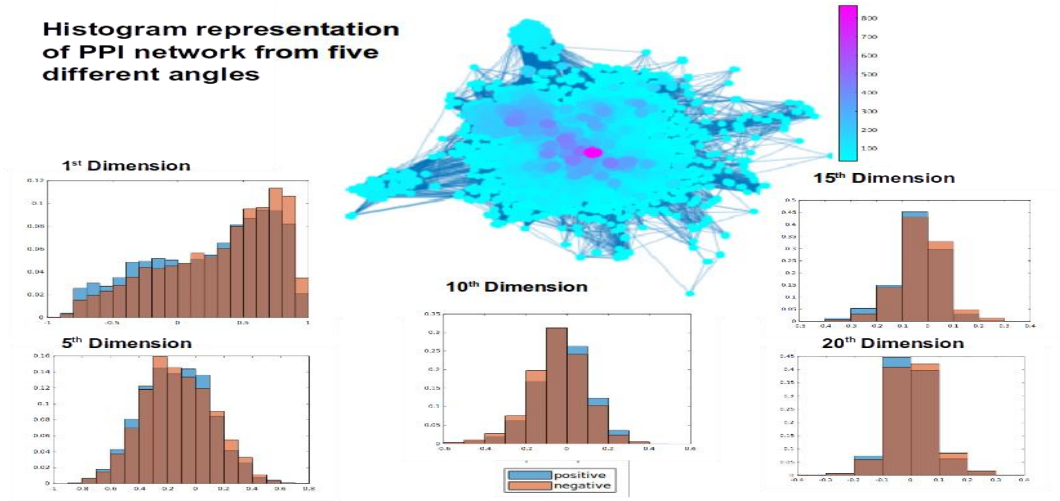


Figure 4.4: Histogram representation of PPI networks from five different angels.

Table 4.3:Comorbid disease prediction of MCE centered method at various dimension values.

	10	20	30	40	50
Random Forest	0.726	0.739	0.739	0.739	0.741
SVM RBF	0.720	0.762	0.761	0.764	0.762
SVM Polynomial	0.682	0.725	0.725	0.731	0.731
SVM Linear	0.582	0.618	0.623	0.644	0.667

Our results mainly focus on comorbidity value RR = 0 and RR = 1 but we also collected results for comorbidity value RR = 2 and 3. The average precision, recall, F-measure and ROC score for each threshold is shown in We compared our results with [22] since this is the only study which used a large amount of data for their analysis. With our method we showed that only common genes between the disease are not the fact of being comorbid, but their spatial location has an impact on their co-occurrence. We can see that regardless of any specific threshold our model performs better the than the already existing method to predict comorbidity using any of the classifier.

Table 4.4 Table 4.4 we can see that RR= 0 and 1 are thresholds where we can predict comorbid diseases from non-comorbid diseases. The results show that if we increase the comorbidity value RR more than one we mislead the classifier towards false prediction. The high score at RR=3 for average precision and recall with low ROC score indicates that most of the disease pairs are falsely predicted as non-comorbid disease pairs.

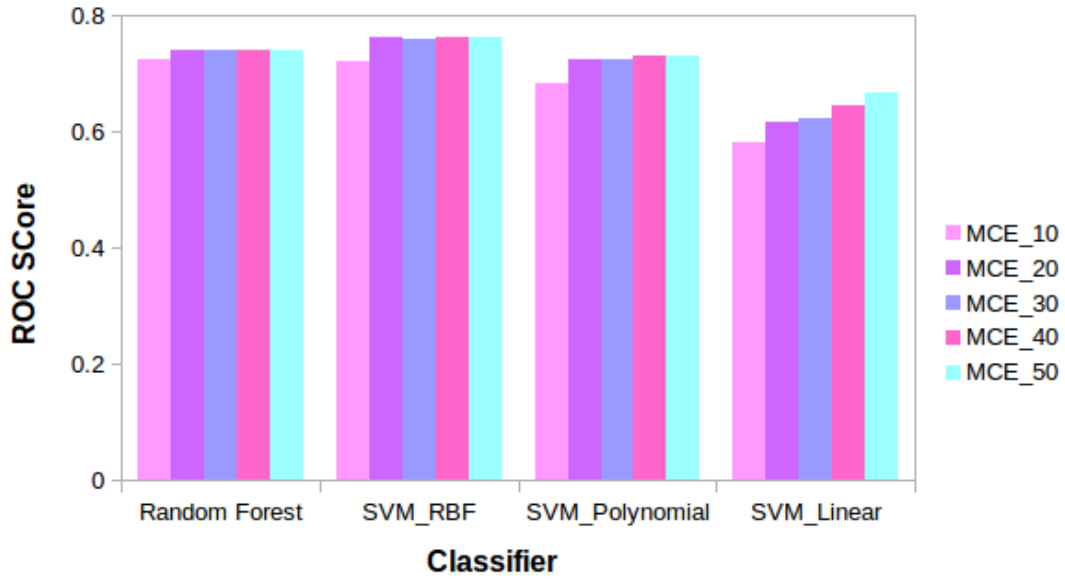


Figure 4.5: Comorbid disease prediction using cantered MCE at higher geometric space.

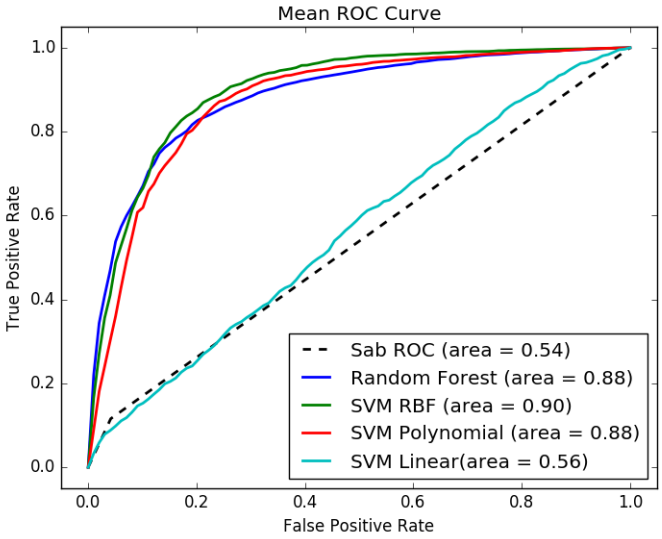
We compared our results with [22] since this is the only study which used a large amount of data for their analysis. With our method we showed that only common genes between the disease are not the fact of being comorbid, but their spatial location has an impact on their co-occurrence. We can see that regardless of any specific threshold our model performs better the than the already existing method to predict comorbidity using any of the classifier.

Table 4.4: Comorbid disease prediction evaluation metrics score at various comorbidity threshold values.

	Precision	Recall	F-measure	Accuracy	ROC
Comorbidity 0					
SVM linear	0.68	0.83	0.75	0.83	0.56
SVM RBF	0.9	0.9	0.89	0.9	0.9
SVM Polynomial	0.87	0.88	0.86	0.88	0.88
Random Forest	0.86	0.86	0.83	0.86	0.88
Module Separation	0.92	0.26	0.31	0.26	0.55
Comorbidity 1					
SVM linear	0.59	0.6	0.56	0.6	0.62
SVM RBF	0.7	0.7	0.69	0.7	0.76
SVM Polynomial	0.68	0.68	0.67	0.68	0.72
Random Forest	0.69	0.7	0.69	0.7	0.74
Module Separation	0.89	0.47	0.57	0.47	0.54
Comorbidity 2					
SVM linear	0.63	0.79	0.7	0.79	0.52
SVM RBF	0.72	0.78	0.73	0.78	0.65
SVM Polynomial	0.72	0.77	0.73	0.77	0.66
Random Forest	0.74	0.79	0.74	0.79	0.65
Module Separation	0.54	0.86	0.77	0.81	0.77
Comorbidity 3					
SVM linear	0.78	0.88	0.83	0.88	0.5
SVM RBF	0.81	0.88	0.83	0.88	0.65
SVM Polynomial	0.81	0.87	0.83	0.87	0.67
Random Forest	0.83	0.88	0.84	0.88	0.65
Module Separation	0.54	0.86	0.83	0.86	0.84

Our method significantly outperforms the baseline method, which is based on the module separation S_{AB} to predict whether a pair of disease is comorbid [22]. We compared our results with [22] since it is to our best knowledge the only study which used a large amount of data for their analysis. For these variants our method, SVM_RBF is the best performer in all datasets Comorbidity_0 (with ROC score = 0.90) and Comorbidity_1 (with ROC score = 0.76) as shown in Figure 4.6, which correspond 64% improvement and 41% improvement respectively from the baseline method. It is also noticed that, on average, better performance is achieved for the dataset Comorbidity_0, which has a more relaxed RR threshold as shown in Table 4.5..

It is important to emphasize here that these results ultimately lead to an area of investigation always heated in the community that the information about Protein-protein interaction and unknown proteins is not complete. If these pieces of information get improved, we might have even better performance towards prediction since we know the interactome we used to be not complete yet but still, this incomplete network has the power to provide useful information.



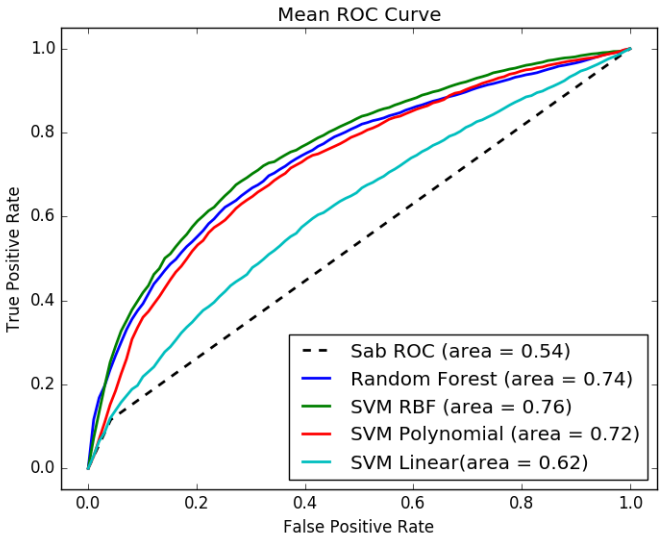


Figure 4.6: ROC Score of comorbidity prediction at (a) RR = 0 and (b) RR = 1 compared with baseline.

We also compared our results by randomizing the genes associated with a disease pair. We retained the gene count associated with each disease and the number of common genes related to a disease pair to maintain the overall topology of a disease pair sub-graph. This experiment shows that even the random data performs better than module separation method but has poor performance when compared with our approach as shown in Figure 4.7. This better performance of our method is due to the spatial arrangement of proteins, which in low dimensional space captures the precise localization of proteins and its association with other proteins in a way that was not achievable by two-dimensional PPI network.

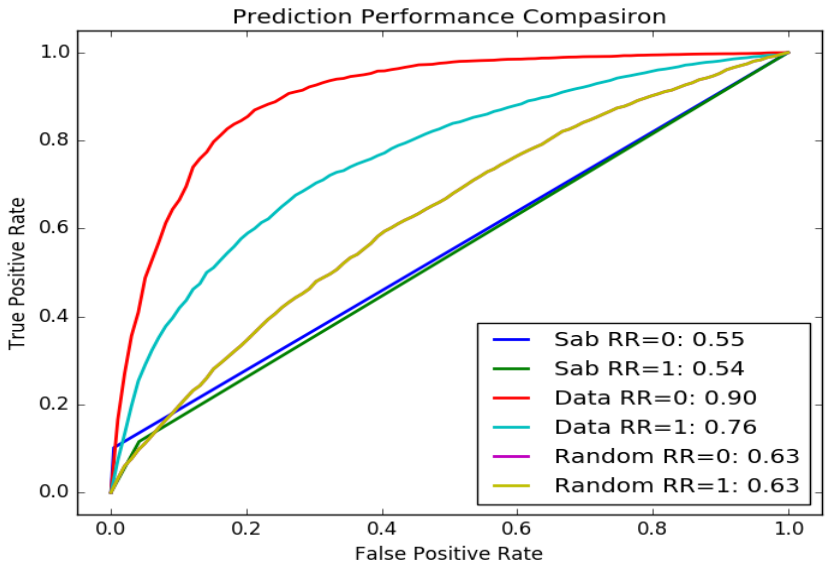


Figure 4.7: ROC Score of comorbidity prediction at RR=0 and RR=1 compared with random data and baseline using SVM_RBF.

Table 4.5: Evaluation metrics for comorbidity 0 and comorbidity 1 using different classifiers

	Precision	Recall	F-measure	Accuracy	ROC
Comorbidity 0					
SVM_linear	0.56	0.59	0.55	0.59	0.59
SVM_RBF	0.60	0.61	0.60	0.61	0.59
SVM_Ploynomial	0.59	0.60	0.59	0.60	0.58
Random Forest	0.62	0.62	0.62	0.62	0.61
Comorbidity 1					
SVM_linear	0.57	0.59	0.54	0.59	0.59
SVM_RBF	0.60	0.61	0.60	0.61	0.59
SVM_Ploynomial	0.59	0.60	0.59	0.60	0.58
Random Forest	0.62	0.61	0.62	0.61	0.61

We also performed a t-test to verify the null hypothesis using 10-fold-cross validation data of original data and the random data. The p-value of 0.0176 validates the statistical significance of our results.

Given that genes are not randomly associated with diseases and there is an underlying rewiring which connects these genes with one another to perform the proper concerned function, disruption of any gene is not damage restricted to itself but related to all the connections it made. These observations supported us to construct a network where we can observe gene related disruption easily. We created a weighted graph using the pathway information from Reactome database. We assign a weight to an edge if both the genes connected are involved in a pathway. Further, we used this weighted network to obtain the matrix D of shortest paths of all node pairs for step two of our protocol.

With the use of the weighted network, we were able to improve the prediction performance with 1% increase for 20 dimensions with p-value 0.93 using ROC score of 10-fold cross-validation. We suspected that might be 10-fold cross validation does not

provide enough data to produce substantial results for such a small increase. Therefore, we also increased the number of cross-validation as 20, 30 and 100, the p-values were 0.311 and 0.29 and 0.15 respectively.

We also attempted to reduce the dimensions and observed the performance. We found that at dimension m=13 the prediction improvement was even 1%, but the p-value was 0.009. This outcome provides a statistically significant improvement over the unweighted graph. The behavior that the performance peaks at some dimension rather than keeps going up as the dimension increases is conceivably due to the possibility that noise is also introduced. We also looked at the minimum spanning tree to see the difference in the edge selection and found that 78% of the edges are similar between the two minimum spanning tree and thus only 22% of the edges made an improvement of 1% in the performance.

We also identify several disease pairs to showcase the significance and better performance ability of our protocol. We are showing three cases where module separation S_{AB} was unable to find an association in disease pair despite higher comorbidity value, but projecting genes on the higher dimension captured it closely. It might be due the disruption of these pathways associated with the disease pairs became a cause for the comorbid behavior of disease pair. Specifically, first disease pair shows the overlap in genes related to disease pair. Second disease pair does not have any common gene, but we found that there is a direct link of pathway associated gene to connect these genes. Third disease pair represents the importance of weighted graph where not only module separation but also our unweighted graph was unable to capture comorbidity, but weighted graph did its job in finding a comorbid association in disease pair.

4.4 Case Studies

4.4.1 Leprosy and Lymphoma

Leprosy has affected human health for decades. It is a chronic infectious disorder caused by a bacterium, Mycobacterium leprae, that affects the skin and peripheral nerves [71]. Whereas, Lymphoma is a group of blood cancer developed from lymphocytes [72]. Using our disease network, we found that there are 13 genes associated with Leprosy and 24 genes related to Lymphoma. This disease pair also has three common associated genes HLA-DQA2, HLA-DQB1, and HLA-DRB5. This disease pair has comorbidity value RR=1.43 and module separation S_{AB}= 0.105. These three genes are associated with several pathways as shown in Figure 4.8.

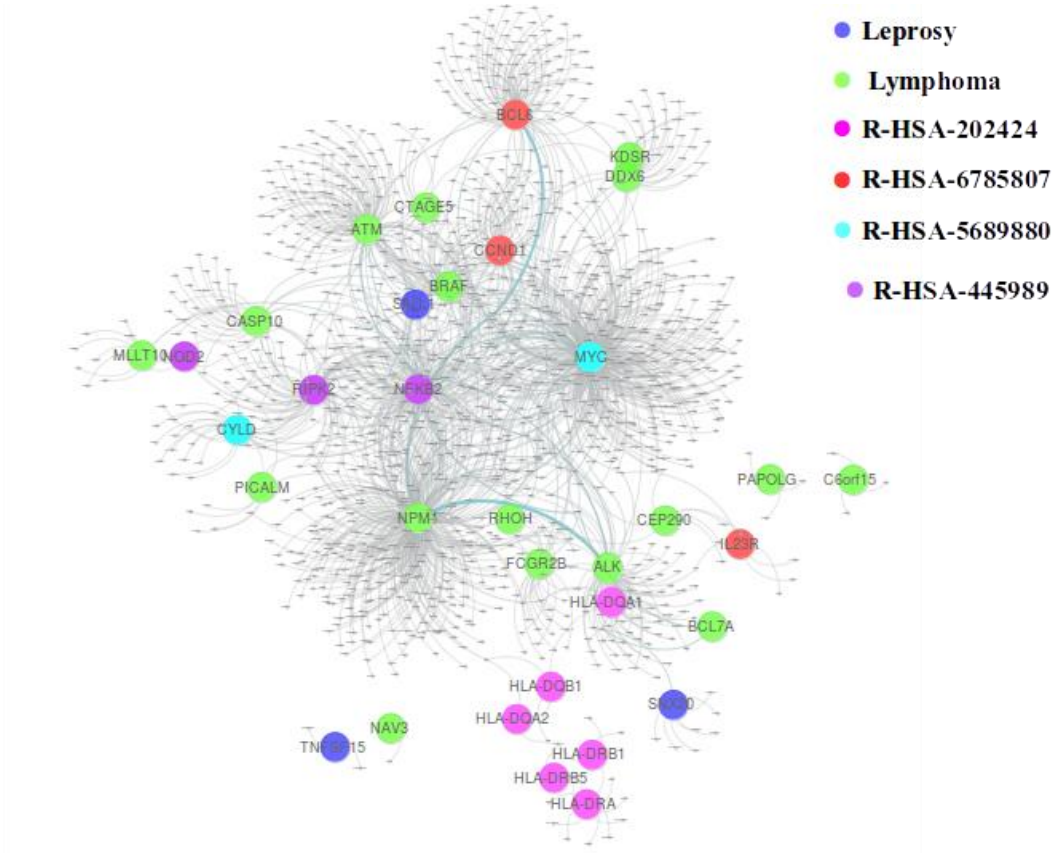


Figure 4.8: Subgraph of leprosy and lymphoma diseases.

With data collection from Reactome database, we found that there are eight different pathways associated with these genes. Specifically, R-HSA-202424 has seven genes from leprosy and three genes from lymphoma taking part together. Among these genes, there are three common genes. This pathway of downstream TCR signaling has a crucial role in gene expression changes which is required for the T cell to gain full proliferative competence and to produce effector cytokines. There are three transcription factors found to play a vital role in TCR-stimulated changes in gene expression, namely NF-kB, NFAT, and AP-1.

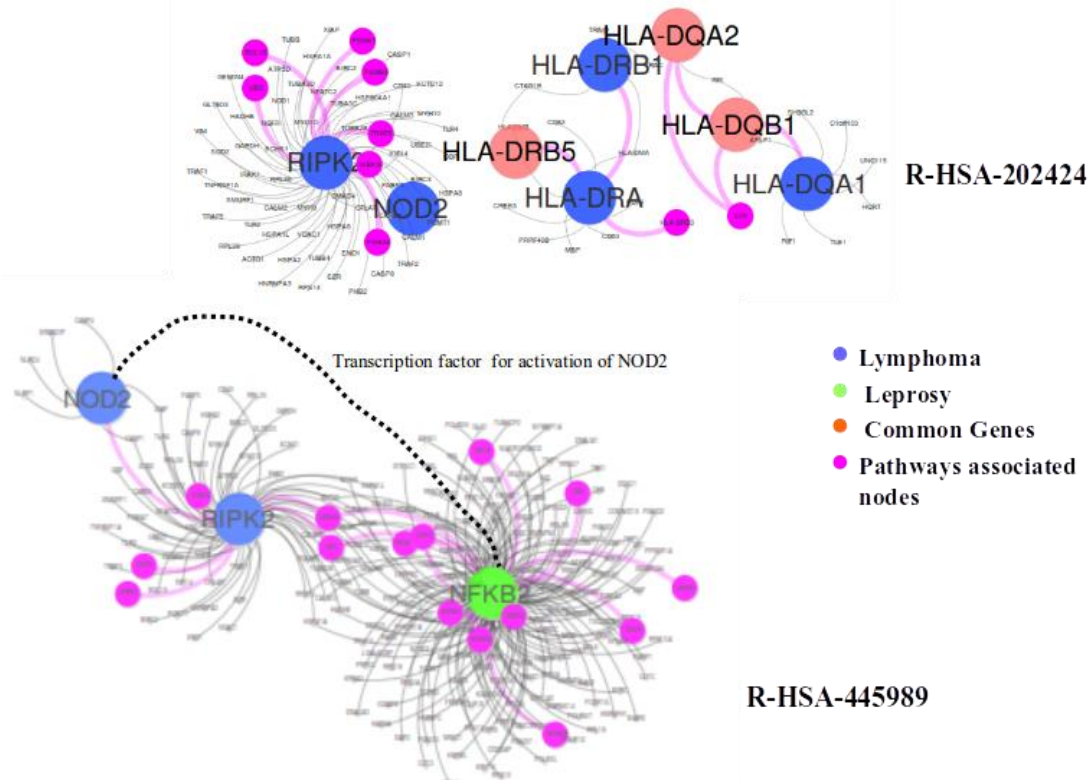


Figure 4.9: Pathway relation to genes associated with leprosy and lymphoma.

We found that among these three transcription factors, NF-kB is associated with lymphoma. Interestingly, this transcription factor with two more genes related to leprosy

is part of another pathway R-HSA-445989. This pathway deals with NFkB activation by TAK1 by phosphorylation and with activation of IkB kinase (IKK) complex. Phosphorylation of IkB results in dissociation of NF-kappaB from the complex allowing translocation of NF-kappaB to the nucleus where it regulates gene expression. The genes associated with leprosy and pathway R-HSA-445989 have a significant role in NFkB activation which is the precursor of the TCR signaling pathway R-HSA-202424 as shown in Figure 4.9.

Two more pathways: R-HSA-6785807 and R-HSA-5689880 have common gene MYC from lymphoma and two separate genes IL23R and CYLD from leprosy associated with pathways respectively. R-HSA-6785807 also has genes BCL6, CCND1 associated with lymphoma, taking their part in the process.

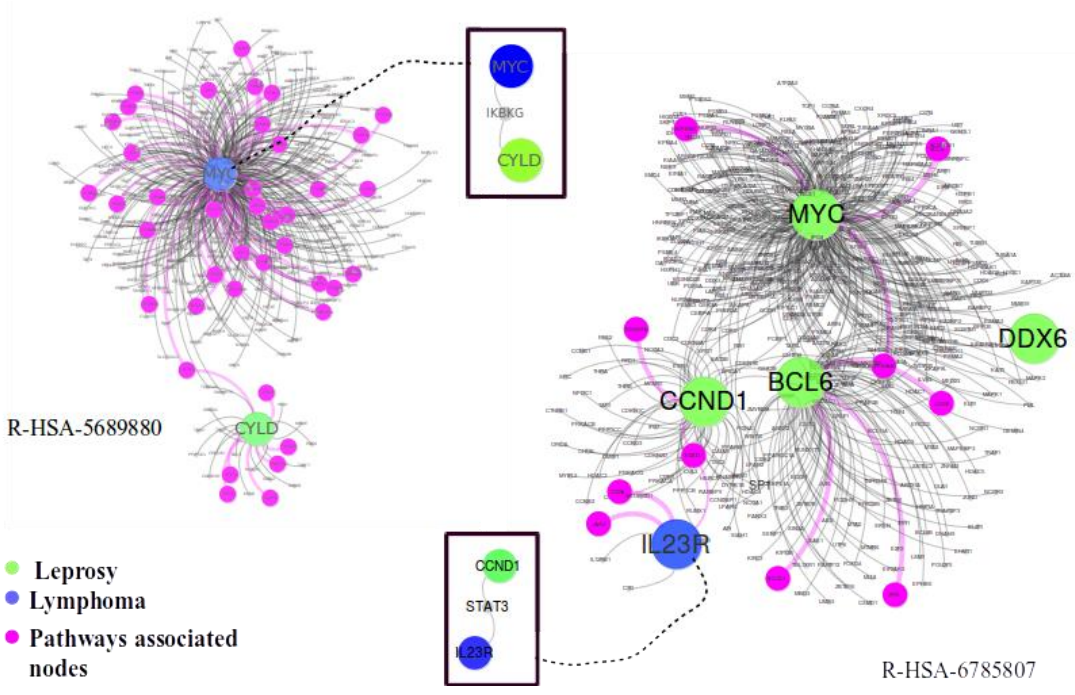


Figure 4.10: Pathway association with leprosy and lymphoma.

R-HSA-5689880 is a pathway associated with Ub-specific processing proteases (USPs). They recognize their substrates by interactions of the variable regions with the substrate protein directly, or via scaffolds or adapters in multiprotein complexes. Whereas R-HSA-6785807 is Interleukin-4 and 13 signaling pathway, where Interleukin-4 (IL4) is a principal regulatory cytokine during the immune response [73]. Another interesting fact about these two pathways is that both have a direct link with gene associated with disease pair and pathway associated gene as shown in Figure 4.10.

Due to this close association of genes with pathways, already demonstrated by high comorbidity, and the presence of the shared gene associated with disease pair. We can safely say that Module separation algorithm failed to capture comorbid nature of disease.

4.4.2 Bechet Syndrome and Osteoporosis

Bechet's syndrome is a type of inflammatory disorder which affects multiple parts of the body causing mouth or genital sores and inflammation of parts of the eye [ref]. OMIM and GWAS associate 13 genes with Behcet's syndrome. While Osteoporosis is a bone disease that occurs when the body either loses too much bone or makes too little bone, or both. In this conditions bones become weak and may break from a fall or, in severe cases, from sneezing or minor bumps [74]. There are 14 genes associated with osteoporosis. There is no known gene overlap between the disease pair. The comorbidity value is RR=2.48 and module separation S_{AB}= 0.5 as shown in Figure 4.11.

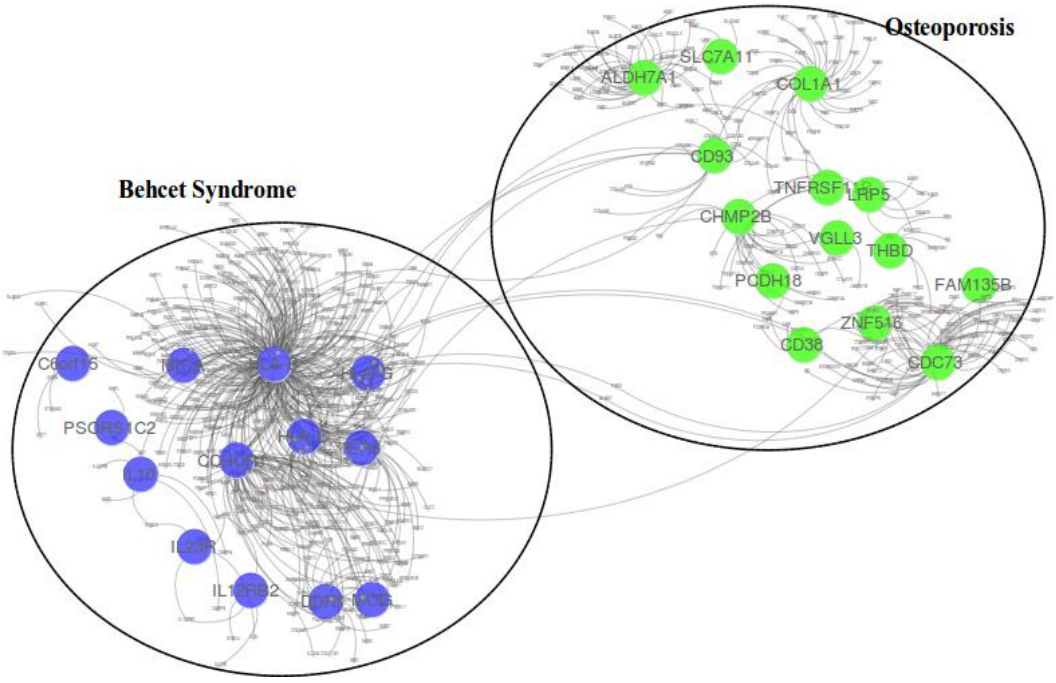


Figure 4.11: Gene disease relationship of behcet syndrome and osteoporosis.

We found that there are three pathways associated with these two diseases collectively. Interestingly each pathway had a gene associated with another pathway to support the link of potential disruption of the system level biology of this disease pair.

As shown in R-HSA-3000171 and R-HSA-198933 have COL1A1 common, and R-HSA-198933 and R-HSA-6798695 have HLA-G and HLA-B common. R-HSA-3000171 plays a role in non-integrin membrane-ECM interactions while Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell. R-HSA-198933 has a role in neutrophil degranulation. The tasks performed by these pathways are closely related to each other and inactivation of one gene can lead to a cascade of non-functionality resulting in causing both diseases at the same time. Geometric embedding of PPI to higher dimension was able to capture this correlation between

disorders which was not seen using module separation technique as shown in Figure 4.12.

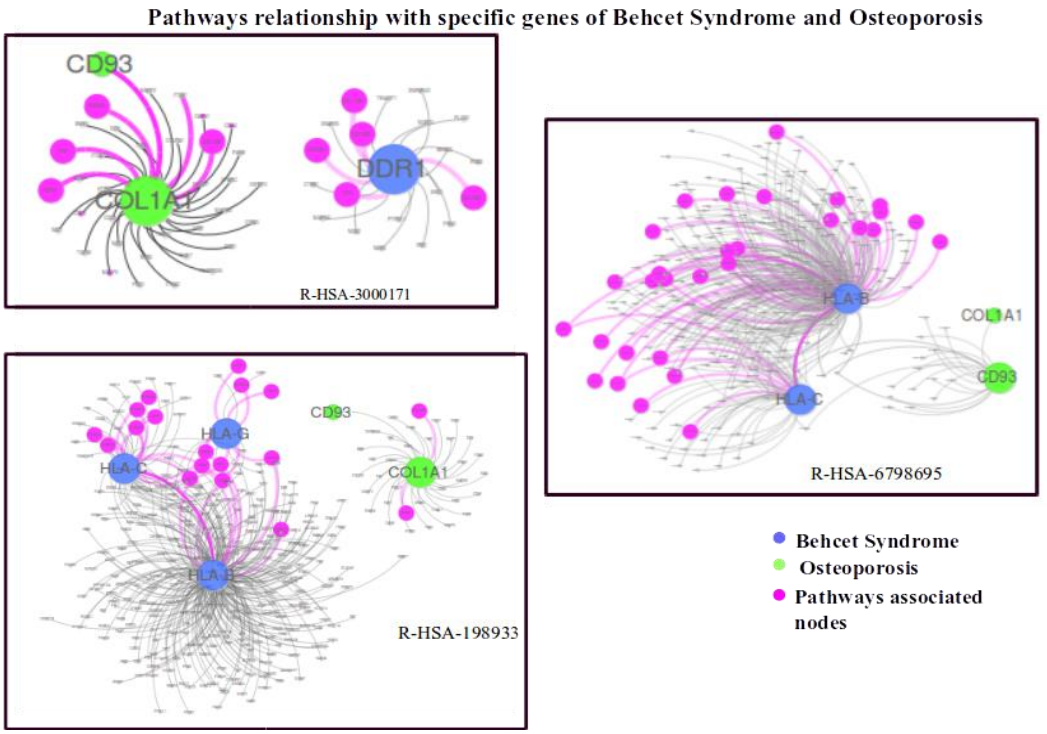


Figure 4.12: Gene pathway association of Bechet syndrome and osteoporosis.

4.4.3 Epilepsy and Glioma

Epilepsy is a group of neurological disorders characterized by episodes that can vary from brief to long periods of vigorous shaking. These episodes can result in physical injuries, including broken bones [75]. A glioma is a type of tumor that starts in the glial cells of the brain and spine causing 30% of all brain tumors and 80% of malignant brain tumors [76]. There are 25 genes associated with epilepsy and 17 genes associated with glioma. Even though both diseases are associated with the brain there is no single common gene associated with the disease pair, besides having high comorbidity RR=10.69 as shown in Figure 4.13.

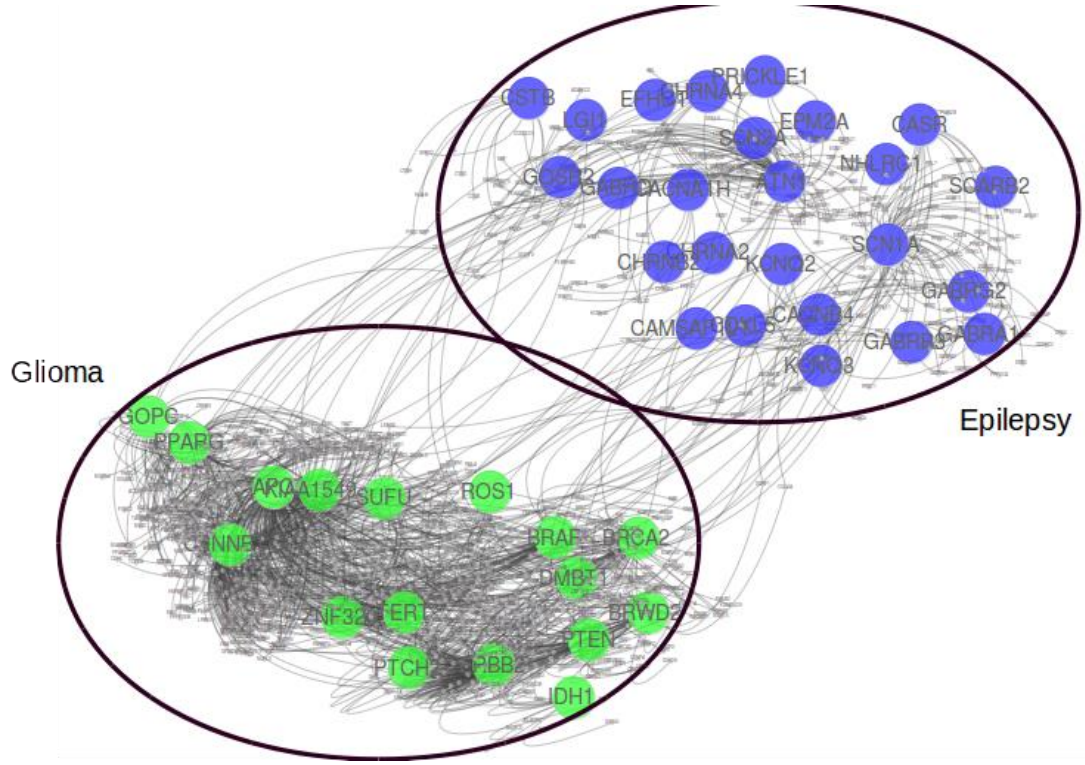


Figure 4.13: Gene Disease relation of Epilepsy and Glioma.

Interestingly module separation was also calculated the disease genes laying apart from one another giving S_{AB}=0.29. It was also observed that our method was also unable to predict it as a comorbid disease. But when we plugged in the weights to the genes due to their pathway association, we found that this disease pair was predicted as a comorbid disease pair. Further incorporation of pathway analysis also shows that there is a link which might cause co-occurrence of these diseases.

We found that there are two pathways R-HSA-6798695 and R-HSA-8943724 associated with disease pair. R-HSA-6798695 is related to Neutrophil degranulation while R-HSA-8943724 is related to Regulation of PTEN gene transcription as shown in Figure 4.14 PTEN gene helps in regulating cell division by keeping cells from growing and dividing too rapidly or in an uncontrolled way. On top of that, if there is any disruption in Neutrophil degranulation, it also affects the defense mechanism of the body. Literature also supports this claim that genes involved in the immune response might play a role in the pathogenesis of tumor growth as well as epileptic symptoms in patients with gliomas [77].

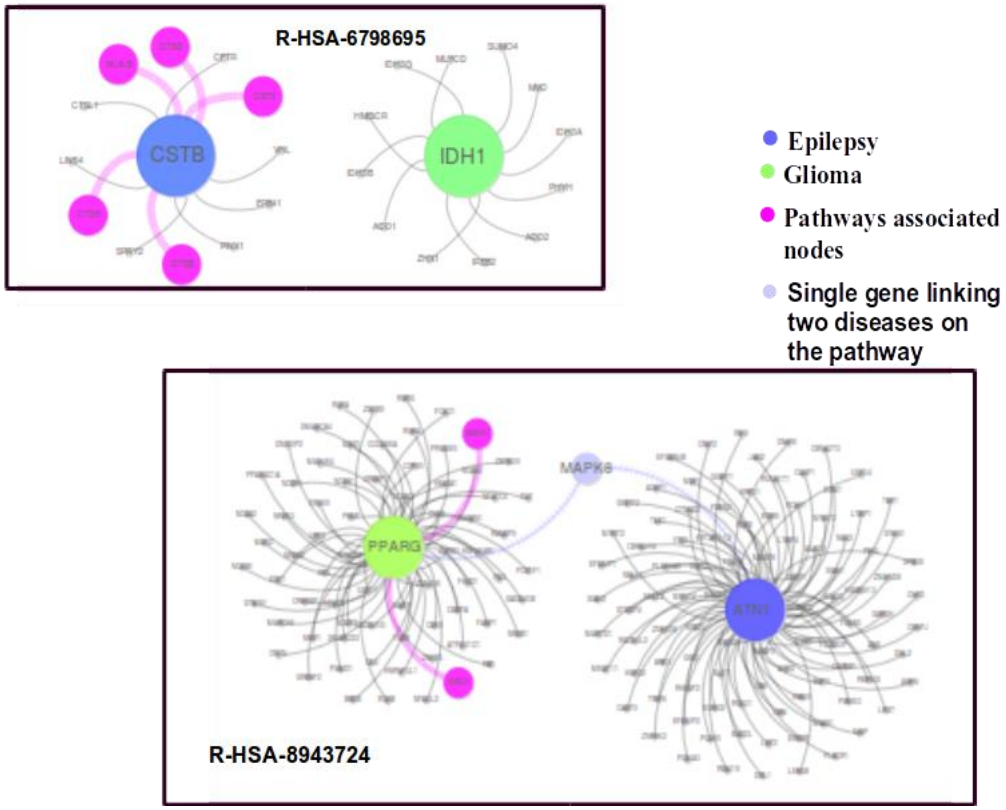


Figure 4.14: Pathways relationship with specific genes of Epilepsy and Glioma.

Our study indicates that the conversion of human interactome to a low dimensional space using any of the embedding algorithms provides a utility to zoom-in the human interactome precisely and provide relevant information which is not captured otherwise. It also emphasizes the problem: Given the comorbidity value RR of disease pairs, we can classify disorders into comorbid or non-comorbid pairs with significant improvement over the previously available method for a large dataset.

4.5 Conclusion

In this work, we developed a computational method to effectively predict comorbid diseases on a large scale. While intuitively the chance for two conditions to be comorbid should go up as they have more associated genes in common, previous studies show that module separation -- how these associated genes of two diseases are distributed on the interactome plays a more critical role in determining the comorbidity than does the number of common genes alone. Our fundamental idea in this work is to embed the two-dimension planar graph of human interactome into a high dimensional geometric space so that we can characterize and capture disease modules (subgraphs formed by the disease-associated genes) from multiple perspectives. Hence it provides enhanced features for a supervised classifier to discriminate comorbid disease pairs from non-comorbid disease pairs more accurately than based on merely the module separation. The results from cross-validation on a benchmark dataset of more 10,000 disease pairs show that our method significantly outperforms the technique of using module separation for comorbidity prediction.

Chapter 5

CONCLUSIONS AND FUTURE WORK

In this doctoral thesis, we mainly focused on the study of comorbid diseases by developing and utilizing the state-of-art machine learning methods to identify and predict key elements related to disease comorbidity, at the sequence level, gene cluster level, and disease level. We developed novel computational methods to tackle to the identification/prediction problems and compared our results with existing methods where applicable. Figure 5.1 depicts the work done in this thesis in a nutshell.

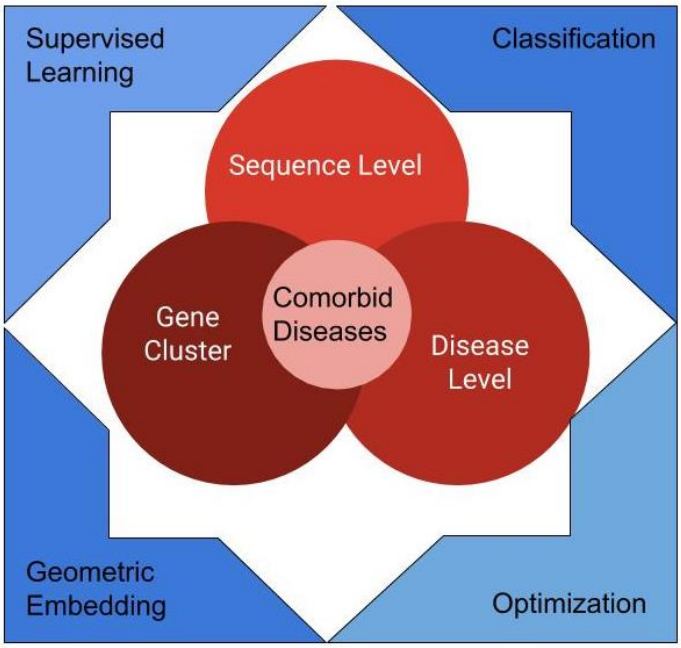


Figure 5.1:Comorbid disease analysis at different levels using machine learning algorithms.

In chapter 2, we carried out a comprehensive comparative analysis for predicting SNPs effect associated with cancers, in the context of SNPs being present at protein interacting sites versus non-interacting sites and being paired within a known haplotype versus being unpaired. We also performed a comparative analysis of using sequential features only, structural features only and combined features.

Our results confirm that prediction performance is improved by using both sequential features and structural features than using them separately. Also, of the two types of classifiers used in the study, random forest outperforms in most cases.

It is found that generic SNP prediction provides a better association of SNP to be detrimental or polymorphic SNPs as compared to disease-specific SNPs, although this conclusion does not hold if genes associated with one disease are unique from the other disease. We also find that prediction performance was increased by associating SNPs to the interacting sites than non-interacting sites. Compared to individual SNPs, these that appear together in haplotype showed a stronger correlation with one another and with the phenotype and therefore led to better prediction performance.

Haplotype SNP prediction provided most promising results. This task could be taken to the next level for improving accuracy further. The possible improvement is by using semi-supervised learning to improve the prediction performance. Currently the performance measured as roc score is 0.95 and this prediction is pretty good to move on to develop customize drugs. The same strategy of using semi supervised learning might also help in prediction of improved performance of residues at interacting sites. Although currently the haplotype and protein site classification were performed for only Acute Myeloid Leukemia, the same protocol can be adapted to perform a similar

analysis on other diseases. Also, the main goal of this project was to apply state-of-art algorithms to establish predictive performance, but one can choose other classifiers do another round of comprehensive comparison depending on the specific requirements.

We also collected the SNPs associated with common genes of the comorbid diseases focusing on the interacting sites at first, and we faced a massive deficit of available data regarding this problem at two levels one at 3D structure and second at the reported SNPs. We only found 129 pairs with SNP association. This crucial task can get accomplished using the same computational pipeline although currently lack of sufficient data samples presents a major hurdle to generate a good learned model. An immediate task in this regard will be associating SNP with diseases and create a single database. SNPedia is a website which specifically associate diseases with SNP and have largest corpus but it is not in a downloadable format. A consolidated SNP disease association may help in SNP and comorbid disease association, leading to several new exploration in comorbid disease and their genetic implications. These developments will not only provide necessary information but also help in bringing together the genetic causes of the phenotypical conditions.

In Chapter 3, we developed a novel method to predict missing common genes for a given disease pair. The method formulates the task as an optimization problem of minimizing network-based module separation for subgraphs formed by associated genes on the interactome, with the hypothesis that correctly identified missing common genes would bring the two-modules "closer". The method provides a useful tool to infer a better understanding of disease-disease interaction regarding related genes. While the method is tested in cross-validation mode in this study, it can be easily deployed to predict *de novo* missing genes, i.e., those genes that are not associated with any disease

but have an impact on the phenotype of both diseases. We also showed that biological pathways are associated with genes which have a known relationship with comorbid diseases and disruption in such pathway can be a plausible cause of comorbid diseases in human. Lastly, we also explored to predict missing interactions using the similar method which resulted in the poor prediction performance. Another way to solve missing common edge problem is given: an attempt to predict and complete the scattered disease module subgraphs might result in the prediction of missing common edges. We know that disease module is not an integrated subgraph but consist of several subgraphs. The largest of these subgraphs is known as Least connected component (LCC). The idea is to find common genes in each subgraph and connect them by adding an edge. In this case, we will be able to identify not only the missing edges in general but also contribute towards making disease module closer and integrated, and ultimately towards completing the human genome.

In Chapter 4, we developed a computational method to predict comorbid diseases on a large scale effectively. While intuitively the chance for two conditions to be comorbid should go up as they have more associated genes in common, previous studies show that module separation -- how these associated genes of two diseases are distributed on the interactome plays a more critical role in determining the comorbidity than does the number of common genes. Our fundamental idea in this work is to embed the two-dimension planar graph of human interactome into a high dimensional geometric space so that we can characterize and capture disease modules (subgraphs formed by the disease-associated genes) from multiple perspectives, and hence provide enhanced features for a supervised classifier to discriminate comorbid disease pairs

from non-comorbid disease pairs more accurately than based on merely the module separation.

We may improve the classification performance for comorbid disease prediction by using transductive learning methods. This method will allow to incorporate disease pairs with unknown comorbid values to be part of training set to improve the prediction performance. In our study we only considered the graph related properties of a protein like its degree and its centrality but by including the specific properties like gene co-expression, disease symptoms, cellular component as features for the classifier.

There are a few tools available to compute comorbid behavior of a disease pair, each of which has some limitation either in terms of functionality or size of the corpus used for learning a model. Since our method provides promising results, as shown in previous chapter, this computational pipeline could be transformed into a software available to predict unknown disease pair by providing the genes association to diseases.

There is a wide range of potentials to further the computational study on comorbid diseases. One aspect is to incorporate the cause-effect mechanisms due to drug-drug interaction leading to comorbid diseases. This area of research mainly focuses on non-genetic reasons of comorbid diseases.

REFERENCES

[1] N. Basit, H. Wechsler, Prediction of enzyme mutant activity using computational mutagenesis and incremental transduction, Advances in bioinformatics 2011 (2011).

[2] J. Wu, M. Gan, R. Jiang, Prioritisation of candidate Single Amino Acid Polymorphisms using one class learning machines, International journal of computational biology and drug design 4(4) (2011) 316-331.

[3] A. David, R. Razali, M.N. Wass, M.J.E. Sternberg, Protein-Protein Interaction Sites are Hot Spots for Disease-Associated Nonsynonymous SNPs, Human Mutation 33(2) (2012) 359-363.

[4] T.S. Lee, D.M. York, Computational Mutagenesis Studies of Hammerhead Ribozyme Catalysis, Journal of the American Chemical Society 132(38) (2010) 13505-13518.

[5] M. Masso, Vaisman, II, Knowledge-based computational mutagenesis for predicting the disease potential of human non-synonymous single nucleotide polymorphisms, Journal of Theoretical Biology 266(4) (2010) 560-568.

[6] R.T. Bradshaw, B.H. Patel, E.W. Tate, R.J. Leatherbarrow, I.R. Gould, Comparing experimental and computational alanine scanning techniques for probing a prototypical protein-protein interaction, Protein Engineering Design & Selection 24(1-2) (2011) 197-207.

[7] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, A.S. Kondrashov, S.R. Sunyaev, A method and server for predicting damaging missense mutations, Nature Methods 7(4) (2010) 248-249.

[8] M.X. Li, J.S.H. Kwan, S.Y. Bao, W.L. Yang, S.L. Ho, Y.Q. Song, P.C. Sham, Predicting Mendelian Disease-Causing Non-Synonymous Single Nucleotide Variants in Exome Sequencing Studies, Plos Genetics 9(1) (2013) 11.

[9] F. Gnad, A. Baucom, K. Mukhyala, G. Manning, Z.M. Zhang, Assessment of computational methods for predicting the effects of missense mutations in human cancers, Bmc Genomics 14 (2013) 13.

[10] B. Reva, Y. Antipin, C. Sander, Predicting the functional impact of protein mutations: application to cancer genomics, Nucleic Acids Research 39(17) (2011) E118-U85.

[11] Y. Dehouck, J.M. Kwasigroch, M. Rooman, D. Gilis, BeAtMuSiC: prediction of changes in protein-protein binding affinity on mutations, Nucleic Acids Research 41(W1) (2013) W333-W339.

[12] P. Kumar, S. Henikoff, P.C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, Nature Protocols 4(7) (2009) 1073-1082.

[13] Y. Dehouck, J.M. Kwasigroch, D. Gilis, M. Rooman, PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality, Bmc Bioinformatics 12 (2011) 12.

[14] Y.X. Song, X. Zhou, Z.N. Wang, P. Gao, A.L. Li, J.W. Liang, J.L. Zhu, Y.Y. Xu, H.M. Xu, The Association between Individual SNPs or Haplotypes of Matrix Metalloproteinase 1 and Gastric Cancer Susceptibility, Progression and Prognosis, Plos One 7(5) (2012) 10.

[15] K. Sun, J.P. Goncalves, C. Larminie, N. Przulj, Predicting disease associations via biological network analysis, BMC Bioinformatics 15 (2014) 13.

[16] J. Loscalzo, I. Kohane, A.L. Barabasi, Human disease classification in the postgenomic era: A complex systems approach to human pathobiology, Molecular Systems Biology 3 (2007) 11.

[17] N. Gulbahce, H. Yan, A. Dricot, M. Padi, D. Byrdsong, R. Franchi, D.S. Lee, O. Rozenblatt-Rosen, J.C. Mar, M.A. Calderwood, A. Baldwin, B. Zhao, B. Santhanam, P. Braun, N. Simonis, K.W. Huh, K. Hellner, M. Grace, A. Chen, R. Rubio, J.A. Marto, N.A. Christakis, E. Kieff, F.P. Roth, J. Roecklein-Canfield, J.A. DeCaprio, M.E. Cusick, J. Quackenbush, D.E. Hill, K. Munger, M. Vidal, A.L. Barabasi, Viral Perturbations of Host Networks Reflect Disease Etiology, Plos Computational Biology 8(6) (2012) 10.

[18] D.S. Lee, J. Park, K.A. Kay, N.A. Christakis, Z.N. Oltvai, A.L. Barabasi, The implications of human metabolic network topology for disease comorbidity, Proceedings of the National Academy of Sciences of the United States of America 105(29) (2008) 9880-9885.

[19] K.I. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, A.L. Barabasi, The human disease network, Proceedings of the National Academy of Sciences of the United States of America 104(21) (2007) 8685-8690.

[20] B. Linghu, E.S. Snitkin, Z.J. Hu, Y. Xia, C. DeLisi, Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network, Genome Biology 10(9) (2009).

[21] M. Zitnik, V. Janjic, C. Larminie, B. Zupan, N. Przulj, Discovering disease-disease associations by fusing systems-level molecular data, Scientific Reports 3 (2013).

[22] J. Menche, A. Sharma, M. Kitsak, S.D. Ghiassian, M. Vidal, J. Loscalzo, A.L. Barabasi, Uncovering disease-disease relationships through the incomplete interactome, Science 347(6224) (2015).

[23] E. Almaas, Biological impacts and context of network theory, Journal of Experimental Biology 210(9) (2007) 1548-1558.

[24] U. Alon, Network motifs: theory and experimental approaches, Nature Reviews Genetics 8(6) (2007) 450-461.

[25] E. Capobianco, Comorbidity: a multidimensional approach, Trends in molecular medicine 19(9) (2013) 515-521.

[26] C.A. Hidalgo, N. Blumm, A.L. Barabasi, N.A. Christakis, A Dynamic Network Approach for the Study of Human Phenotypes, Plos Computational Biology 5(4) (2009).

[27] M.A. Moni, P. Liò, comoR: a software for disease comorbidity risk assessment, Journal of clinical bioinformatics 4(1) (2014) 8.

[28] R. Gijsen, N. Hoeymans, F.G. Schellevis, D. Ruwaard, W.A. Satariano, G.A.M. van den Bos, Causes and consequences of comorbidity: a review, Journal of clinical epidemiology 54(7) (2001) 661-674.

[29] B. Starfield, K.W. Lemke, T. Bernhardt, S.S. Foldes, C.B. Forrest, J.P. Weiner, Comorbidity: implications for the importance of primary care in ‘case’management, The Annals of Family Medicine 1(1) (2003) 8-14.

[30] L.F. Drager, P.R. Genta, R.P. Pedrosa, F.B. Nerbass, C.C. Gonzaga, E.M. Krieger, G. Lorenzi-Filho, 249 Characteristics And Predictors Of Obstructive Sleep Apnea In Consecutive Patients With Hypertension, Sleep Medicine 10 (2009) S67.

[31] A. Levin, O. Djurdjev, B. Barrett, E. Burgess, E. Carlisle, J. Ethier, K. Jindal, D. Mendelssohn, S. Tobe, J. Singer, C. Thompson, Cardiovascular disease in patients with chronic kidney disease: Getting to the heart of the matter, American Journal of Kidney Diseases 38(6) (2001) 1398-1407.

[32] C.H. Zheng, L. Zhang, V.T.Y. Ng, S.C.K. Shiu, D.S. Huang, Molecular Pattern Discovery Based on Penalized Matrix Decomposition, Ieee-Acm Transactions on Computational Biology and Bioinformatics 8(6) (2011) 1592-1603.

[33] J.F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G.F. Berriz, F.D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D.S. Goldberg, L.V. Zhang, S.L. Wong, G. Franklin, S.M. Li, J.S. Albala, J.H. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R.S. Sikorski, J. Vandenhaute, H.Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M.E. Cusick, D.E. Hill, F.P. Roth, M. Vidal, Towards a proteome-scale map of the human protein-protein interaction network, Nature 437(7062) (2005) 1173-1178.

[34] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F.H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, E.E. Wanker, A human protein-protein interaction network: A resource for annotating the proteome, Cell 122(6) (2005) 957-968.

[35] D.S. Huang, H.J. Yu, Normalized Feature Vectors: A Novel Alignment-Free Sequence Comparison Method Based on the Numbers of Adjacent Amino Acids, Ieee-Acm Transactions on Computational Biology and Bioinformatics 10(2) (2013) 457-467.

[36] H. Paik, H.S. Heo, H.J. Ban, S.B. Cho, Unraveling human protein interaction networks underlying co-occurrences of diseases and pathological conditions, Journal of Translational Medicine 12 (2014).

[37] S. Park, J.S. Yang, Y.E. Shin, J. Park, S.K. Jang, S. Kim, Protein localization as a principal feature of the etiology and comorbidity of genetic diseases, Molecular Systems Biology 7 (2011).

[38] S. Park, J.S. Yang, J. Kim, Y.E. Shin, J. Hwang, J. Park, S.K. Jang, S. Kim, Evolutionary history of human disease genes reveals phenotypic connections and comorbidity among genetic diseases, Scientific Reports 2 (2012).

[39] F. He, G.H. Zhu, Y.Y. Wang, X.M. Zhao, D.S. Huang, PCID: A Novel Approach for Predicting Disease Comorbidity by Integrating Multi-Scale Data, Ieee-Acm Transactions on Computational Biology and Bioinformatics 14(3) (2017) 678-686.

[40] P. Akram, L. Liao, Cancer Specific Non-Synonymous Single Nucleotide Polymorphism Prediction in the Context of Haplotype and Protein Interacting Sites, Journal of Biomedical Science and Engineering 10(05) (2017) 28.

[41] P. Akram, L. Liao, Prediction of missing common genes for disease pairs using network based module separation, 2016 IEEE 6th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), 2016, pp. 1-1.

[42] P. Akram, L. Liao, Prediction of missing common genes for disease pairs using network based module separation on incomplete human interactome, BMC genomics 18(10) (2017) 902.

[43] L. Breiman, Random forests, Machine learning 45(1) (2001) 5-32.

[44] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20(3) (1995) 273-297.

[45] A. Hamosh, A.F. Scott, J.S. Amberger, C.A. Bocchini, V.A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, Nucleic Acids Research 33(suppl_1) (2005) D514-D517.

[46] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L.J. Jensen, C.v. Mering, The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored, Nucleic Acids Research 39(suppl_1) (2011) D561-D568.

[47] A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, Nucleic Acids Research 28(1) (2000) 45-48.

[48] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, L. Serrano, The FoldX web server: an online force field, Nucleic Acids Research 33(suppl_2) (2005) W382-W388.

[49] L.R. Brunham, R.R. Singaraja, T.D. Pape, A. Kejariwal, P.D. Thomas, M.R. Hayden, Accurate prediction of the functional significance of single nucleotide polymorphisms and mutations in the ABCA1 gene, PLoS genetics 1(6) (2005) e83.

[50] P.D. Thomas, A. Kejariwal, N. Guo, H. Mi, M.J. Campbell, A. Muruganujan, B. Lazareva-Ulitsky, Applications for protein sequence–function evolution data: mRNA/protein expression analysis and coding SNP scoring tools, Nucleic Acids Research 34(suppl_2) (2006) W645-W650.

[51] G.A. Thorisson, A.V. Smith, L. Krishnan, L.D. Stein, The international HapMap project web site, Genome research 15(11) (2005) 1592-1593.

[52] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, The human genome browser at UCSC, Genome research 12(6) (2002) 996-1006.

[53] C. The Genomes Project, An integrated map of genetic variation from 1,092 human genomes, Nature 491 (2012) 56.

[54] M.A. Yıldırım, K.-I. Goh, M.E. Cusick, A.-L. Barabási, M. Vidal, Drug—target network, Nature biotechnology 25(10) (2007) 1119.

[55] J. Park, D.S. Lee, N.A. Christakis, A.L. Barabási, The impact of cellular networks on disease comorbidity, Molecular systems biology 5(1) (2009) 262.

[56] N.J. Perkins, E.F. Schisterman, The Inconsistency of "Optimal" Cutpoints Obtained using Two Criteria based on the Receiver Operating Characteristic Curve, American Journal of Epidemiology 163(7) (2006) 670-675.

[57] N.D. Wong, Epidemiological studies of CHD and the evolution of preventive cardiology, Nature Reviews Cardiology 11 (2014) 276.

[58] Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013, The Lancet 385(9963) (2015) 117-171.

[59] F.G.R. Fowkes, D. Rudan, I. Rudan, V. Aboyans, J.O. Denenberg, M.M. McDermott, P.E. Norman, U.K.A. Sampson, L.J. Williams, G.A. Mensah, M.H. Criqui, Comparison of global estimates of prevalence and risk factors for peripheral artery disease in 2000 and 2010: a systematic review and analysis, The Lancet 382(9901) (2013) 1329-1340.

[60] M. PrabhuDas, D. Bowdish, K. Drickamer, M. Febbraio, J. Herz, L. Kobzik, M. Krieger, J. Loike, T.K. Means, S.K. Moestrup, S. Post, T. Sawamura, S. Silverstein, X.-Y. Wang, J. El Khoury, Standardizing Scavenger Receptor Nomenclature, The Journal of Immunology 192(5) (2014) 1997.

[61] T.F. Cox, M.A.A. Cox, Multidimensional scaling, CRC press2000.

[62] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290(5500) (2000) 2319-+.

[63] G.E.a.S.T.R. Hinton, Stochastic Neighbor Embedding, (2003) 857--864.

[64] L.v.d. Maaten, G. Hinton, Visualizing data using t-SNE, Journal of machine learning research 9(Nov) (2008) 2579-2605.

[65] C.V. Cannistraci, G. Alanis-Lobato, T. Ravasi, Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding, Bioinformatics 29(13) (2013) 199-209.

[66] C.V. Cannistraci, T. Ravasi, F.M. Montevecchi, T. Ideker, M. Alessio, Nonlinear dimension reduction and clustering by Minimum Curvilinearity unfold neuropathic pain and tissue embryological classes, Bioinformatics 26(18) (2010) i531-i539.

[67] O. Kuchaiev, M. Rašajski, D.J. Higham, N. Pržulj, Geometric de-noising of protein-protein interaction networks, PLoS computational biology 5(8) (2009) e1000454.

[68] D. Croft, A.F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M.R. Kamdar, The Reactome pathway knowledgebase, Nucleic acids research 42(D1) (2013) D472-D477.

[69] A. Fabregat, K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann, R. Haw, B. Jassal, S. Jupe, F. Korninger, S. McKay, The reactome pathway knowledgebase, Nucleic acids research 44(D1) (2015) D481-D487.

[70] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143(1) (1982) 29-36.

[71] K. Suzuki, T. Akama, A. Kawashima, A. Yoshihara, R.R. Yotsu, N. Ishii, Current status of leprosy: epidemiology, basic science and clinical perspectives, The Journal of dermatology 39(2) (2012) 121-129.

[72] B.T. Hennessy, E.O. Hanrahan, P.A. Daly, Non-Hodgkin lymphoma: an update, The lancet oncology 5(6) (2004) 341-353.

[73] T. Hershey, J.W. Mink, Using functional neuroimaging to study the brain's response to deep brain stimulation, Neurology 66(8) (2006) 1142-1143.

[74] A.L. Golob, M.B. Laya, Osteoporosis: screening, prevention, and management, Medical Clinics 99(3) (2015) 587-606.

[75] R.S. Fisher, C. Acevedo, A. Arzimanoglou, A. Bogacz, J.H. Cross, C.E. Elger, J. Engel, L. Forsgren, J.A. French, M. Glynn, ILAE official report: a practical clinical definition of epilepsy, Epilepsia 55(4) (2014) 475-482.

[76] J.A. Schwartzbaum, J.L. Fisher, K.D. Aldape, M. Wrensch, Epidemiology and molecular pathology of glioma, Nature Reviews Neurology 2(9) (2006) 494.

[77] S.G. Berntsson, B. Malmer, M.L. Bondy, M. Qu, A. Smits, Tumor-associated epilepsy and glioma: are there common genetic pathways?, Acta oncologica 48(7) (2009) 955-963.

[78] C. Roger and L Li, Transductive Learning with EM Algorithm to Classify Proteins Based on Phylogenetic Profiles, Int. J. Data Mining and Bioinformatics (2007) 1:337-351

Appendix A

LIST OF PUBLICATIONS, PRESENTATIONS AND CONSENT TO REPRINT

Published:

1. P. Akram, L. Liao, Cancer Specific Non-Synonymous Single Nucleotide Polymorphism Prediction in the Context of Haplotype and Protein Interacting Sites, Journal of Biomedical Science and Engineering 10(05) (2017) 28.

2. P. Akram, L. Liao, Prediction of missing common genes for disease pairs using network based module separation, 2016 IEEE 6th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), 2016, pp. 1-1.

3. P. Akram, L. Liao, Prediction of missing common genes for disease pairs using network based module separation on incomplete human interactome, BMC genomics 18(10) (2017) 902.

.

Accepted:

- P. Akram and L. Liao. Predicting Comorbid Diseases with Geometric Space Embedding of Human Interactome. (ISBRA 2018)

Presentation:

- **Cancer Specific Non-Synonymous Single Nucleotide Polymorphism *Prediction in the Context of Haplotype and Interacting Sites* (Authors:** Pakeeza Akram and Li Liao) https://www.ijcai-boom.org/2016-proceeding.html

Consent to Reprint:

Open access articles

The open access articles published in BioMed Central's journals are made available under the Creative Commons Attribution (CC-BY) license, which means they are accessible online without any restrictions and can be re-used in any way, subject only to proper attribution (which, in an academic context, usually means citation).

The re-use rights enshrined in our license agreement include the right for anyone to produce printed copies themselves, without formal permission or payment of permission fees. As a courtesy, however, anyone wishing to reproduce large quantities of an open access article (250+) should inform the copyright holder and we suggest a contribution in support of open access publication.

Open Access Statement

All articles from **Journal of Biomedical Science and Engineering (JBiSE)** have "free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself." (From the BOAI Definition of Open Access)

Please find further information about Open Access at SCIRP on

- SCIRP's Open Access Page
- About SCIRP ("What is Open Access?", " How Open is SCIRP on the 'Open Access Spectrum'?")