RANKPCSF: A DISEASE MODULE IDENTIFICATION METHOD BY INTEGRATING NETWORK PROPAGATION WITH PRIZE-COLLECTING STEINER FOREST

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

ARDA ESKIN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF BIOINFORMATICS

SEPTEMBER 2022

rankPCSF: A Disease Module Identification Method by Integrating Network Propagation with Prize-Collecting Steiner Forest

submitted by **ARDA ESKIN** in partial fulfillment of the requirements for the degree of **Master of Science in Bioinformatics Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics**

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics**

Assist. Prof. Dr. Burçak Otlu
Supervisor, **Health Informatics, METU**

Assoc. Prof. Dr. Nurcan Tunçbağ
Co-supervisor, **Chemical and Biological Engineering, KOÇ University**

Examining Committee Members:

Assoc. Prof. Dr. Yeşim Aydın Son
Health Informatics, METU

Assist. Prof. Dr. Burçak Otlu
Health Informatics, METU

Assoc. Prof. Dr. Nurcan Tunçbağ
Chemical and Biological Engineering, KOÇ University

Assist. Prof. Dr. Aybar Can Acar
Health Informatics, METU

Assist. Prof. Dr. Tuğba Süzek
Computer Engineering, Muğla Sıtkı Koçman University

Date: 02.09.2022

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Arda Eskin

Signature :

ABSTRACT

RANKPCSF: A DISEASE MODULE IDENTIFICATION METHOD BY INTEGRATING NETWORK PROPAGATION WITH PRIZE-COLLECTING STEINER FOREST

Eskin, Arda

M.S., Department of Bioinformatics

Supervisor: Assist. Prof. Dr. Burçak Otlu

Co-Supervisor: Assoc. Prof. Dr. Nurcan Tunçbağ

September 2022, 31 pages

Identification of disease modules may lead to a better understanding on the progression of diseases, finding more accurate biomarkers, and drug targets. In this study, we develop a hybrid method – RANKPCSF- combining the strength of Steiner tree and diffusion approaches and release a new network reconstruction approach. RANKPCSF is capable to integrate multi-omic data (including phosphoproteomic and transcriptomic) with a reference interactome to unveil the optimal disease-associated network. We have compared RANKPCSF's performance on predicting known cancer genes and drug targets with NetCore, Hierarchical HotNet and DOMINO. On average, RANKPCSF was able to capture cancer driver genes and cancer drug targets with higher precision and identified more cancer genes from the set of genes that were deemed not significant in the pan-cancer experiment. Next, we compared the functional relevancy of the resulting networks from RANKPCSF and other methods. NetCore and RANKPCSF gave the highest functional relevances based on empirical p-values ($\approx$ 0.001). Finally, we have applied RANKPCSF to reconstruct ischemic heart disease (IHD)-associated network by using the transcriptomic data from 46 patients as an independent validation case study. We observed modules related to oxidative stress, extracellular matrix organization, and immune response, which were relevant to ischemic heart disease pathology. Overall, RANKPCSF captures known cancer genes and drug targets more accurately and performed better on functional relevance test compared to other selected algorithms. We believe that RANKPCSF – as a hybrid method - can be used for different tasks including functionally relevant subnetworks, and these subnetworks can be used to generate disease-associated hypotheses.

Keywords: Bioinformatics, Systems Biology, Network Medicine, Disease Module

ÖZ

RANKPCSF: AĞ ÜZERİNDE YAYILIM VE ÖDÜL TOPLAYAN STEİNER ORMANI KULLANARAK HASTALIK MODÜLÜ TESPİT METODU

Eskin, Arda

Yüksek Lisans, Biyoenformatik Bölümü

Tez Yöneticisi: Dr. Öğr. Üyesi. Burçak Otlu

Ortak Tez Yöneticisi: Doç. Dr. Nurcan Tunçbağ

Eylül 2022, 31 sayfa

Hastalık modüllerinin tanımlanması, hastalıkların ilerlemesinin daha iyi anlaşılmasına, daha doğru biyobelirteçlerin ve ilaç hedeflerinin bulunmasına yol açabilir. Bu çalışmada, Steiner ağacının gücünü ve difüzyon yaklaşımlarını birleştiren hibrit bir yöntem olan RANKPCSF geliştirdik ve yeni bir ağ yeniden yapılandırma yaklaşımı yayınladık. RANKPCSF, hastalıkla ilişkili optimal ağı ortaya çıkarmak için multi-omik verileri (fosfoproteomik ve transkriptomik dahil) bir referans interaktom ile entegre etme yeteneğine sahiptir. RANKPCSF'in bilinen kanser genlerini ve ilaç hedeflerini tahmin etme performansını NetCore, Hierarchical HotNet ve DOMINO ile karşılaştırdık. Ortalama olarak, RANKPCSF kanser sürücü genlerini ve kanser ilacı hedeflerini daha yüksek hassasiyetle yakalayabildi ve pan-kanser deneyinde önemli olmadığı düşünülen gen setinden daha fazla kanser geni tanımladı. Daha sonra, RANKPCSF ve diğer yöntemlerden elde edilen ağların işlevsel uygunluğunu karşılaştırdık. NetCore ve RANKPCSF, ampirik p değerlerine ($\approx$ 0,001) dayalı olarak en yüksek işlevsel alaka düzeyini verdi. Son olarak, bağımsız bir doğrulama vaka çalışması olarak 46 hastadan alınan transkriptomik verileri kullanarak iskemik kalp hastalığı (IHD) ile ilişkili ağı yeniden yapılandırmak için RANKPCSF'i uyguladık. İskemik kalp hastalığı patolojisi ile ilgili olan oksidatif stres, hücre dışı matris organizasyonu ve bağışıklık tepkisi ile ilgili modülleri gözlemledik. Genel olarak, RANKPCSF bilinen kanser genlerini ve ilaç hedeflerini daha doğru bir şekilde tespit etti ve seçilen diğer algoritmalara kıyasla fonksiyonel uygunluk testinde daha iyi performans gösterdi. RANKPCSF'in - bir hibrit yöntem olarak - işlevsel olarak ilgili alt ağlar dahil olmak üzere farklı görevler için kullanılabileceğine ve bu alt ağların hastalıkla ilişkili hipotezler oluşturmak için kullanılabileceğine inanıyoruz.

Anahtar Kelimeler: Biyoenformatik, Sistem Biyolojisi, Ağ Tıbbı, Hastalık Modülü

To My Family

ACKNOWLEDGMENTS

First person that I want to thank is my advisor Assoc. Prof. Dr. Nurcan Tunçbağ. This work wouldn't be possible without her guidance. I got introduced to a new field and learned many things with her as a mentor. I sincerely thank Dr. Tunçbağ for her patience and continuous support during my masters. I want to thank my other advisor Dr. Burçak Otlu, with her support I was able to progress in my studies.

I would also like to thank my friends, Oğuz Ulaş Yaman from Chemical Engineering, Ahmet Samet Özdilek from Bioinformatics, Ali Akman from Chemical Engineering, Oğuz Kahveci and Batuhan Güvercin for their long term companionship and support. I would like to express my gratitude to Afşin Peker, Deniz Berk Karaman, Umut Geçmişli, Egemen Sert, Tunahan Kanbak, Ogün Tarakçı and Ekin Dursun whose endless support was essential for my progress.

Last but not least, I would like to express my deepest gratitude to my parents Nevin Eskin and Atilla Eskin for their endless support, love and belief. My parents always supported me even in my hardest times, and nothing would be possible without them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AMI	Active Module Identification
APP	Amyloid Precursor Protein
CGC	Cancer Gene Census
COSMIC	Catalogue of Somatic Mutations in Cancer
DNA	Deoxyribonucleic Acid
FDA	Food and Drug Administration
FDR	False Discovery Rate
GDA	Gene-Disease Association
GEO	Gene Expression Omnibus Database
GO	Gene Ontology
GIT1	G Protein-Coupled Receptor Kinase Interactor 1
HIPPIE	Human Integrated Protein-Protein Interaction Reference
HTT	Huntingtin
IHD	Ischemic Heart Disease
KEGG	Kyoto Encyclopedia of Genes and Genomes
MDM2	Murine Double Minute 2
MutSig	Pan-cancer Somatic Mutation Significance
PCSF	Prize-Collecting Steiner Forest
PPI	Protein-Protein Interaction
PPR	Personalized PageRank
RNA	Ribonucleic Acid
TCGA	The Cancer Genome Atlas
TPM	Transcripts Per Million

CHAPTER 1

INTRODUCTION

Molecular mechanisms that occur inside a cell are predominantly carried out by proteins and their interactions. Proteins can interact with other proteins, small molecules, and nucleic acids to create a complex network of interactions that contribute to various functions. Therefore, Protein-protein interactions (PPIs) describe the fundamental molecular mechanisms of cellular life, and these interactions may be disturbed to become a diseased state [1]. These interactions contributing to various biological functions are estimated to have a size of 650,000 [2]. Only a fraction of the human PPIs are validated, and throughout the years with more experiments and new prediction strategies, curated human PPI networks are getting bigger.

PPIs can be in a normal state which is a state of homeostasis of an organism, and PPIs can be in a disturbed state (can be due to environmental stress, disease state...). Molecular interconnectivity implies that the impact of a specific abnormality that occurs for a molecule is not only confined to the activity of the molecule, but spreads along the connections of the network and affects other molecules that otherwise carry no abnormalities [3]. From this principle, a disease phenotype rarely happens by an abnormality in a single effector molecule, but a consequence of various pathobiological mechanisms inside the complex PPI network. Therefore, a disturbed state of a PPI can be explained as the consequence of a spread of information on the interaction network caused by the disturbed effector molecule. A better understanding on the disturbed state of a PPI can lead to the identification of more accurate biomarkers to restore the diseased state as well as more accurate disease classifications.

The rise of high-throughput omic data over the last decade, caused detailed human interactome data to be generated. Motivated by the new possibilities that came with the new data, the field of network medicine that uses interaction networks to learn about the molecular mechanism basis of complex diseases arose. Studies about networks have shown that either biological, technological or social networks do not have a random characteristic, but are organized under a core set of topological principles [3]. In the context of biological networks, understanding diseases using these principles can lead to answers on some fundamental properties of disease related proteins. Only a fraction of human proteins that are known to be associated with a disease. By utilizing PPI networks, questions related to the correlations between the locations of disease associated genes and their network topology can be answered.

Under the field of network medicine, active module identification (AMI) or disease module identification is one of the important problems that uses high-throughput data like gene expression and a PPI network, and aims to find subnetworks that explain the significant changes occurring under different conditions [4]. Two common steps that are used for this problem are: 1- Network ranking, where the vertices are ranked according to their initial weights from the high-throughput data and network topol-

ogy. 2- Subnetwork identification, where a subnetwork or subnetworks are identified in a PPI network using vertex scores and network topology [5]. The current paradigm of the network ranking step is to rerank the initial scores of the vertices coming from high-throughput data with a network propagation step. In disease genetics, disease related genes tend to cluster with each other in the interactome and some significant results coming from high-throughput data may result in false positives in the disease association (e.g. observing biological differences related to cell state in gene expression, driver and passenger genes may have similar mutation frequency in modest sized cohorts...) [6]. Propagation algorithms like random walk [7, 8] smoothen the initial weights coming from high-throughput data across the network, and prioritize nodes that are clustered with each other in the network. After reranking the nodes, the next step is to find a subnetwork using weight on the nodes.

In this work, a combination of network propagation step, and a subnetwork construction using Prize-Collecting Steiner Forest algorithm was utilized for active module identification. Prize-Collecting Steiner Forest algorithm previously applied for this problem with OmicsIntegrator [9] tool, uses a terminal node set from significant results in an high-throughput experiment to construct a subnetwork. Key issues of extending terminals, and false positive problem of high-throughput data were handled by a network propagation step. A common but costly step to reduce network propagation bias of high degree nodes by probability calculation using random degree preserving network permutations was handled by the degree penalization of Prize-Collecting Steiner Forest algorithm.

In Chapter 2, we present a detailed literature review by starting with protein-protein interaction and disease research by using protein-protein interaction networks. Then, we review subnetwork identification problem for diseases. We finish the chapter with a review on the algorithms for active / disease module identification algorithms.

In Chapter 3, we explain the steps that are applied in the algorithm, and what we did to validate the algorithm. We start by showing the overview of the algorithm, and we give a detailed explanation on how the algorithm operates. We give the sources of the data and what processing steps are applied on the data that is used for validation. We conclude the chapter by explaining how the validation is done and what tools were used in the validation.

In Chapter 4, we show the results for validation of the algorithm, and application on a test case. We apply the algorithm on pan-cancer mutation significance data to create a subnetwork. We compare the algorithm on other methods in their ability to find known cancer genes, cancer drug targets, and functional relevance. We also test the algorithm for functional relevance using 6 different gene expression datasets. We apply the algorithm to gene expression data of ischemic heart disease patients to perform a basic network analysis as a test case.

In Chapter 5, we give a discussion on our validation results and the results of the network analysis for ischemic heart disease patients. Potential future work for the algorithm, how its performance can be increased by incorporating multi-omic data, and challenges of the problem are also discussed.

CHAPTER 2

LITERATURE SURVEY

2.1 Human Protein-Protein Interaction Network & Diseases

After the Human Genome Project was completed, the emphasis on classifying the genes and proteins changed to surveying the networks of molecular interactions that occur between them. Deciphering these networks is critical because proteins do not function as isolate parts, instead they interact with each other, as well as DNA, RNA, and small molecules to create complex molecular machines. These modular machines involve both static and dynamic macromolecule assemblies, and can transmit and react to both extracellular and intracellular signals [10]. The focus on understanding networks of molecular mechanisms encoded by model species has shifted to understanding the mechanism networks of human diseases as a result of the increased knowledge of human protein interactome [11]. The advancements on understanding interaction networks of human diseases are divided into four categories: the investigation of the properties of human disease related genes on interaction networks; finding other genes that have an implication of disease using protein-protein interaction networks; the identification of disease associated subnetworks; and classifying case-control studies with a network-based approach [12].

An important property of protein networks is the observation of a few proteins that are highly connected (high degree nodes), often called as hub proteins, have an implication that these proteins must have unique biological roles [3]. Experimental evidence from the studies on model organisms have shown that these hub proteins come from essential genes [13], and these genes tend to evolve more slowly and are older than the genes related to non-hub proteins [14, 15, 16]. If the genes that encode hub proteins get deleted, phenotypic consequence is much larger than the deletion of other genes [17]. Effects related to hub proteins are still debated [18], due to the higher connections that these proteins have, removal of a hub protein from the interaction network has an effect on many more genes than the removal of a non-hub protein. Hypothesis of considering hub proteins as disease associated proteins in humans is a result of assuming hub proteins have higher importance in the protein network. However, disease associated genes are not always essential genes. Mutations that are observed in essential genes in early development can not grow in the population, as mutations in these genes have a cause on embryonic lethality; in contrast, most of the individuals that past their reproductive age, can tolerate disease associated mutations for longer. This example may suggest that there are more non-essential genes related to diseases than the essential genes [19, 20].

2.2 Disease-Related Subnetwork Identification

In addition to finding network properties of disease proteins and predicting disease associated proteins, a protein interaction network can also be used to identify subnetworks that are related to the disease. Identification of disease subnetworks are important, contrary to individual disease proteins, subnetworks can provide concrete hypotheses on signaling mechanisms, interactions of molecular complexes, and other molecular mechanisms that have an effect on disease outcome. In one study [21], a subnetwork was constructed around Huntingtin (HTT), which is a protein that is associated with Huntington disease and its mutated gene is a known cause of the disease. When all of the proteins that are directly connected to HTT were tested for their ability to increase HTT aggregation, the screening identified a new enhancer, GIT1, and it is validated for its role in disease progression.

In another study [22], a subnetwork of endotoxin inflammatory response was created from curated literature information on gene expression, where genes that are responsive to endotoxin were profiled. The subnetwork that is generated from endotoxin responsive genes have enabled the identification of new modules related to endotoxin response and transcriptional regulation mechanisms of leukocytes, inhibition of energy production in mitochondria and protein synthesis machinery were re-prioritized. In more general terms, interpreting the activity of gene expressions as hot spots and mapping this information on an interaction network has potential application in disease research.

Common to all of the studies above are mapping gene level information on interaction networks can lead to the identification of new genes that are associated to a disease and generating disease subnetworks offers causal information on the disease with mechanistic hypotheses. The protein interactions from disease subnetworks suggest metabolic pathways, signaling cascades, or interactions of molecular complexes that have an effect or cause on the phenotype of the disease. Essentially, these subnetworks provide explanations about the genetic and environmental factors that influence a disease in the context of a small sized discrete modules [12].

As the studies continue to provide answers on network properties of disease associated genes, three phenomena emerge that need to be distinguished [3]. In the context of an interaction network, the topological module is defined as a module consisting of a group of nodes that are within a locally dense region, such that these nodes have a tendency to have more interactions with the other nodes at the same local region compared to the nodes that are outside of this region. Topological modules are often identified using a graph clustering method that usually does not consider the functions of the nodes [23]. A functional module is a module having a group of nodes that are functionally the same and are in the same local neighbourhood. In the context of biological networks, function of a node represents the role of a gene in detectable phenotypes. The final phenomena is a disease module, which is a module that consists of nodes that contribute to a function or functions and when they are disrupted, a particular phenotype related to the disease emerges.

In the biological literature on interaction networks, one assumption is that these 3 phenomena are interrelated: nodes that are in a topological module tend to have similar molecular functions, it means that they can also form a functional module, and essentially an abnormality of a functional module is a cause of disease [24], which indicates that a topological module is also relevant to a disease module. Moreover, there are several special characteristics of a disease module that are significant [3]. A disease module is likely to be overlapped with a functional or topological module, but they may not be identical. A disease module is generated uniquely for a disease, thus each disease has its own module. Furthermore, a molecule can be associated with several diseases, and it can be observed in the modules of different diseases. Considering these characteristics can help to refine the module identification, which is a critical step of network medicine.

Being in a state of disease is a combinatorial problem, different disturbed parts and defects can cause a similar disease phenotype in the context of alterations in the disease module [3]. Some of the combinatorial mechanisms are well explained for cancer [25], but the usefulness of generating a disease module extends beyond polygenic diseases and its also important for some monogenic diseases. As an example, sickle cell disease, which is a Mendelian disorder, is caused by a mutation in the beta chain of haemoglobin [26]. However, there is not a single disease phenotype observed after a single point mutation in haemoglobin: sickle cell disease can cause osteonecrosis, stroke, anemia, acute chest syndrome, and painful crises. Observation of multiple pathophenotypes suggest that the disease module consists of various disease modifying genes that are related to different biological phenomena like transcriptional, post-translational, and epigenetic. Therefore, it is important to identify a disease module for the disease phenotype of interest, which can be critical for further studies on decoding the disease mechanism, identification of biomarkers, and drug development [3].

2.3 Algorithms for Disease Network Identification

Algorithms that seek for subnetworks that are over-represented by nodes with high activity scores are called active module or disease module identification algorithms. Modules reported by these algorithms are often called as active modules [27, 28]. AMI algorithms rely on a list of nodes that are scored, called activity scores, and a graph object to identify modules. After the analysis of omics profile and the calculation of activity scores, subnetworks detected by an AMI algorithm capture context-specific molecular mechanisms that are related to a phenotype or cellular state [28]. AMI methods can utilize different metrics for scoring, cost functions, and constraints. Activity scores can be calculated as binary or continuous, there can be penalization in the cost function to include low activity nodes, and the number of non-active nodes in the subnetwork can be arranged with constraints [29]. While the metrics may vary from one algorithm to another, the activity scores are always calculated from the data. Many AMI algorithms relies on combinatorial optimization and has been proven to be NP-hard [27], there are many heuristics suggested for solving the AMI problem [28].

A prominent class of AMI algorithms are network propagation based algorithms. Network propagation offers a refined re-ranking of the activity scores by considering all paths between the nodes, and gives higher scores to the nodes that are in a common neighbourhood. Network propagation based algorithms can overcome some of the difficulties that algorithms based on shortest path give. In detail, false positives that are predicted by a single path can have a low rank after network propagation, and potential causal genes that are missed, even if they are well connected in the graph to high activity nodes, can be promoted after re-ranking with network propagation [6]. In network propagation, ini-

tial activity scores are diffused through the paths between the nodes until a certain number of steps is reached. As an alternative to halt the process, network propagation can have a reset parameter that represents a probability of returning to initial score rather than continue to propagate. This probability causes the process to reach a steady state, and scores of distance nodes relative to the source will exponentially decrease based on its distance [6]. Although the process can deprecate distance hub nodes due to scores being diffused by the neighbours of the hub, local hubs can be prioritized due to the higher chance that these hubs can be neighbours with high scoring nodes. Additional penalization can be applied to prevent hub nodes being favored. As an example, a common penalization for hubs in network propagation based methods is assigning p values to the scores of every node by using distribution of scores for a particular node under network propagation of randomized scores [30]. This penalization has a down-weighting effect on the hubs. Network propagation based AMI algorithms often have an additional step for module identification. Main purpose of network propagation is re-ranking the nodes, and the scores after re-ranking can be used in a module discovery step to obtain relevant modules. For example, connected components of high scoring nodes can be extracted as modules. Various different module discovery steps are applied in different methods. One approach used by HotNet is to calculate a similarity matrix that records the propagation results, and clustering on the similarity matrix to extract modules [5].

CHAPTER 3

MATERIALS AND METHODS

In this chapter, we explain the materials and methods that were used for subnetwork identification and validation of the algorithm.

3.1 Overview of the Algorithm

Active module or disease module methods map the result of an omic data to a PPI network, and a subnetwork is generated that is a representative mechanism for the changes observed in the omic data. We have performed 2 steps to obtain a subnetwork from an omic data and a PPI network. In the first step, a weighted gene list (seeds) were used in a propagation step with a PPI network. Propagation was done using Personalized PageRank [31] with a core normalized adjacency matrix. In the next step, top k genes after PageRank that are either in the seed list or transcription factors were used as terminals in a network construction algorithm. Prize-collecting Steiner Forest (PCSF) algorithm that is implemented by OmicsIntegrator2 [9] was used as a network construction algorithm. Multiple optimal networks were generated from PCSF by running the algorithm with different parameters. Top nodes were selected by number of occurrences in the networks, and the final subnetwork is created by extracting the connected component graph from the PPI network. The flowchart of the algorithm is indicated in Figure 3.1. The algorithm was run using pan-cancer mutation data and 6 different gene expression data and 2 PPI networks for validation. Then, the algorithm was run using gene expression data from ischemic heart disease patients as a test case.

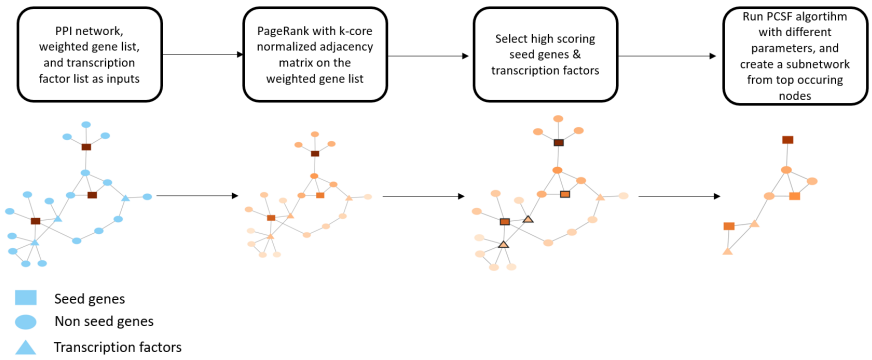


Figure 3.1: Flowchart of the algorithm.

3.2 Subnetwork Identification Algorithm

3.2.1 Propagation

In order to generate terminal nodes to construct a subnetwork, initial gene scores were propagated through the protein-protein interaction (PPI) network using Personalized PageRank (PPR) algorithm [31]. PPR ranking of the nodes represents the probability of a random walker in a graph to terminate its walk starting from initial nodes. Therefore, it is used as an importance metric between initial genes and all of the genes in the PPI network.

PPR scores for a damping parameter a, normalized initial gene scores vector s and normalized adjacency matrix M were calculated by solving following equation iteratively until maximum steps of iteration or convergence was reached:

$$PPR(a,s)_{i+1} = a * PPR(a,s)_i * M + (1-a) * s \tag{1}$$

The classical degree based normalization on adjacency matrix is:

$$M_{deg} = A * D^{-1} \tag{2}$$

A is an adjacency matrix such that, if (i,j) are neighbors A(i,j) = 1, otherwise 0, and D is a diagonal matrix with the node degrees.

We have also applied a k-core based normalization of the adjacency matrix. K-core number of a node is the largest value of k of a k-core graph containing that node. A K-core graph is a maximal subgraph that contains nodes having a degree of k or more. K-core graphs are found with an algorithm called k-shell decomposition [32]. First all nodes with degree 1 and their edges are removed, then remaining nodes having a degree of 1 are also removed until none such nodes remain. All of the nodes that are removed in this step have a k = 1. For k=2, the same procedure is applied with degree 2, and the algorithm ends when all of the nodes are removed. K-core numbers can be a representative of densely connected regions of a graph. Nodes receive more scores from densely connected neighbors when k-core normalized adjacency matrix is used.

K-core normalized adjacency matrix can be obtained by solving following equation:

$$M_{core} = D^T * A \tag{3}$$

D is a diagonal matrix with k-core numbers. Each column is divided by its sum after calculating equation 3 to obtain the core normalized adjacency matrix.

3.2.2 Terminal Selection & Subnetwork Construction

Before constructing the subnetworks, terminals from ranked genes were filtered using 2 constraints. First constraint is selecting nodes that are either in seed list or in transcription factor list. Second

constraint is selecting top k nodes having the highest ranks as terminals. Extracted terminals with 2 constraints were used in the prize-collecting Steiner forest algorithm (PCSF) with a PPI network.

First step in the PCSF algorithm is to assign prizes on the nodes. A weight was assigned to the initial gene scores of $p(v)$, and a penalization factor based on node degree was used to prevent protential bias introduced by hub nodes.

$$p'(v) = \beta * p(v) - \mu * degree(v) \tag{4}$$

PCSF algorithm uses a network of a node set V and edge set E, $G(V, E)$. Network can be a directed, partially directed, or an undirected network having $p'(v)$ prizes for each node $v \in V$, and edge costs $c(e) > 0$ for each edge $e \in E$. PCSF algorithm finds a subnetwork with vertices V_F, and edges E_F using the assigned costs and prizes by minimizing an objective function:

$$f'(G'(V_F, E_F)) = \sum_{v \notin V_F} p'(v) + \sum_{e \in E_F} c(e) + \omega * \kappa \tag{5}$$

Where κ is the number of trees in the forest, and is a weight to control the number of trees in the output. PCSF algorithm introduces artificial nodes v_0 that are connected a subset of nodes N in the initial network G, and is a uniform edge cost between the dummy node and nodes in N. Final subnetwork $G'(V_F, E_F)$, is an acyclic connected graph that is rooted at v_0.

PCSF algorithm was run using different combinations of β, μ, κ parameters. After calculating multiple subnetworks using different PCSF parameters, top nodes were extracted using the number of occurrences in the PCSF networks. Connected component subgraph was generated from the top nodes to be presented as the final result.

3.3 Datasets

3.3.1 Protein-protein Interaction Network Data

2 different PPI networks were used in the validation of algorithms and analyses. First PPI network is HIPPIE v2.2 [33]. HIPPIE retrieves molecular interactions from different sources and assigns stringent confidence scores to the interactions. After removing the self loops and selecting the largest connected component, the remaining network has 398902 edges, and 17886 nodes. Second PPI network that is used is ConsensusPathDB (release 35) [34]. ConsensusPathDB is a molecular interaction database that gathers interactions from 31 different public sources. The database has a web interface that allows users to search for interactions of molecules and pathways, and to perform analyses like enrichment and shortest interaction paths. All the binary interactions having a confidence score above 0.90 were selected. Self loops were removed, and the largest connected component of the network was extracted, the remaining network has 138400 edges, and 10426 nodes.

3.3.2 High-throughput Omic Datasets

The pan-cancer somatic mutation significance data (MutSig) [35] includes 4742 samples from 21 different tumor types, 12 of them were reported in The Cancer Genome Atlas and 14 of them from other projects at the Broad Institute. Duplicated samples and mutations were removed from the dataset. MutSigCV , MutSigCL and MutSigFN significance calculations were applied to the mutations. Both MutSig values were tested with a permutation test for non-silent coding mutations, and p values were combined into a single p value. P values were corrected into FDR Q values using the Benjamini and Hochberg correction method [36]. A total of 1489 genes had a Q below 1, 1294 of these genes were in the HIPPIE nodes, and 895 of these genes were in the ConsensusPathDB nodes. Remaining genes having Q < 1, and found in the networks were used as seeds in the analyses, and FDR Q values were converted into $-log_{10}(Q)$.

6 different gene expression datasets were collected having GEO accessions of GSE64233 [37], GSE109064 [38], GSE108693 [39], GSE106847 [40], GSE123689 [41], and GSE101788 [42]. Between test and control groups, differential expression analysis using edgeR was applied for RNA-seq and for microarray datasets Student's t-test was applied. P values were corrected into q values by Benjamini and Hochberg FDR correction method. Top 500 differentially expressed genes that are found in the PPI network were used in the analysis as seeds for each of the datasets. Q values were converted into $-log_{10}(Q)$ scores for analysis.

RNA-seq experiment dataset for ischemic heart disease patients (GEO accession GSE173594) [43] were collected. Based on SYNTAX Score 2 [44], which is a clinical parameter that shows complexity of ischemic heart disease, patients were separated into 2 groups. Low severity group has SYNTAX Score 2 < 20, and High severity group has SYNTAX Score 2 > 39. Differential expression analysis was performed between low severity and high severity groups using iDEP [45], a web application that performs differential expression analysis from read counts. A total of 321 genes having differential expression FDR < 0.1, and found in the PPI network were selected as seeds in the analysis. FDR Q values were converted into $-log_{10}(Q)$ scores for analysis.

Table 3.1: Gene sets that were used in validation.

Source	Gene set size	Coverage in HIPPIE	Coverage in ConsensusPathDB
MutSig	1489	1294	895
GSE64233	13478	10905	-
GSE109064	28103	14407	-
GSE108693	59958	16723	-
GSE106847	22347	16327	-
GSE123689	33033	16571	-
GSE101788	26970	13833	-
GSE173594	27937	14638	-
CGC	729	707	619
CancerDrugsDB	1015	976	835

3.3.3 Datasetst for Validation

Cancer Gene Census (CGC) genes, which are curated list of genes which contain mutations that have been causally implicated in cancer, and these genes were cataloged in COSMIC [46]. Genes having a type 1 or type 2 evidence were extracted, and 707 genes that were found in HIPPIE, 619 genes that were found in ConsensusPathDB were used as a validation set. In addition to known cancer genes, a list of FDA approved cancer drugs and their targets were collected from CancerDrugs_DB [47]. 976 of the cancer drug targets were found in HIPPIE, and 835 of the cancer drug targets were found in ConsensusPathDB. Cancer drug targets that were found in the networks were used in validation.

3.4 Validation of the Algorithm

3.4.1 Validation on Subparts of the Algorithm

MutSig data was used with HIPPIE and ConsensusPathDB to create a subnetwork. PageRank damping factor was selected as 0.7, and top 200 genes were selected as terminals to be able to extend the size of the network further. The PCSF algorithm was run using unique combinations of (1,3,5) for each 3 parameters, and a total of 27 subnetworks were created by PCSF for each run of the algorithm. PageRank only runs were done with the same inputs and same damping factor. PCSF only run was done by selecting the top 200 from seed list with the same PCSF parameter combinations. Non-active genes in the validation were defined as genes in the subnetwork that have a MutSig Q value > 0.05. DIAMOnD [48] algorithm that is used to test non-active genes was run with the same seed input and PPI networks, permutation number was fixed to 66 for HIPPIE, 60 for ConsensusPathDB to give the same number of non-active genes with our algorithm.

3.4.2 Comparison of Network Methods

Using the same inputs, the algorithm was run with damping factor 0.7, combinations of (1,3,5) as PCSF parameters and different top k sizes for terminal selection (100, 200, 300). Number of occurrences in the network parameter was selected to have a subnetwork size around 125. NetCore [49] algorithm was run using the same inputs, and the number of permutations were selected as 100. Hierarchical Hotnet [50] was run using the same inputs, the number of permutations were selected as 100, and lower size bound parameter was selected as 10, which was suggested for larger graphs. DOMINO [29] algorithm was run from their web application. Compared to the other algorithms, DOMINO requires a gene list without weights. When the same input is applied without weights, DOMINO's result becomes incomparable in size. As suggested, we have used only significant genes (MutSig Q < 0.05) as input to the DOMINO.

3.4.3 Functional Relevance Based Validation

We have used DIGEST [51] to do a functional relevance test. From the web application, we have selected subnetwork validation using genes, and we have selected the similarity measure as Jaccard Index, number of runs as 1000, and replace as 100. We have uploaded the HIPPIE and ConsensusPath

networks that are used in the algorithm to DIGEST to test functional relevance. 6 gene expression datasets were used in addition to MutSig data. Top 500 genes that are differentially expressed were selected as seeds, algorithm was run using a damping factor of 0.5, combination of (1,3,5) as PCSF parameters, and top 100 genes were selected as terminals. For each of the dataset, differential expression results were randomized 5 times under uniform distribution to create a random group for comparison. The algorithm was executed with the same parameters for the random group to obtain random subnetworks.

3.5 Network Analysis of Ischemic Heart Disease Patients

Differentially expressed genes between high severity and low severity patients having FDR $Q < 0.1$ were used as seeds in the algorithm. Expressed genes were identified as genes having a normalized expression value > 0.5 TPM over 90% of the samples. Largest connected component having a node set of expressed genes was selected from the HIPPIE network, and the filtered network was used in the algorithm. Damping factor for PageRank was selected as 0.5, PCSF algorithm was executed with a combination of (1,3,5) as parameters, and top 100 nodes as terminals after propagation. Nodes of the subnetwork were searched in the DisGeNET [52] disease-gene association database. Nodes that are related to cardiovascular diseases, and having a GDA score > 0.3 (1 or more curated source) were extracted. Clustering was applied on the subnetwork using Louvain clustering [53] which is implemented by the python-louvain module. The subnetwork was visualized , and eigenvector centralities were calculated using Cytoscape [54]. Functional enrichment results were obtained from ShinyGO 0.76 [55], which is a web service to perform enrichment analysis. FDR cutoff was selected as 0.1, minimum pathway size was selected as 15, and maximum pathway size was selected as 500.

CHAPTER 4

RESULTS

Identification of disease modules is a key to create concrete hypotheses for explaining the outcome of disease, and for more accurate biomarkers to cure the disease. In this work, we integrated data from an high-throughput experiment and a PPI network to create a subnetwork that represents a significant outcome in the data. To validate the algorithm, pan-cancer gwas data from TCGA and 6 different gene expression data were used as high-throughput datasets. The algorithm was executed on HIPPIE and ConsensusPathDB PPI networks. We tested the algorithm for its ability to identify cancer associated genes, FDA approved cancer drug targets, and to construct a functionally relevant subnetwork by comparing it to random subnetworks. Presented algorithm was compared to multiple other methods that were presented for the identification of disease modules. Furthermore, we used the algorithm on gene expression data for ischemic heart disease patients by comparing two groups showing high ischemia and low ischemia to create a network model for explaining functional outcomes.

4.1 Tests on Subparts of the Algorithm

The algorithm consists of a network propagation step, and a network construction step. To be able to select a better normalization step on the propagation, and a better terminal selection procedure the parts of the algorithm were compared with each other and pan-cancer mutation significance baseline. As hypothesized, with increasing number of top k nodes, PCSF algorithm loses its ability to extend the signal quickly as shown in Figure 4.1. We observed that doing a core normalization on PageRank rather than classic degree based normalization performed better on identifying cancer associated genes for both networks. Running the full algorithm compared to the baseline and its subparts performed better on identifying cancer associated genes. For the terminal selection part, we observed that limiting top nodes to only seed genes and transcription factors after reranking the nodes by PageRank resulted with more precision than directly selecting top nodes from PageRank as terminals. The difference being more apparent on HIPPIE network, which is a larger network, may suggest that PageRank without additional guidance introduces more false positives on non-seeds in the top terminals. We have tested the algorithm's ability to identify cancer associated genes on non-active genes, which are the genes that resulted as not significant in pan-cancer mutation significance. Subnetworks are created to be less than 150 nodes in size by adjusting the number of occurrence parameter, and the biggest subnetwork is selected. DIAMOnD algorithm which is a widely used algorithm for this purpose was selected as a reference. DIAMOnD uses a 'number of permutations' input that determines the size of the result, which is exactly the size of non-active genes, and the input was arranged to give the same amount of

non-active genes with the proposed algorithm for better comparison. As shown in Figure 4.1 (b), our algorithm found more cancer associated non-active genes, compared to DIAMOnD and PCSF.

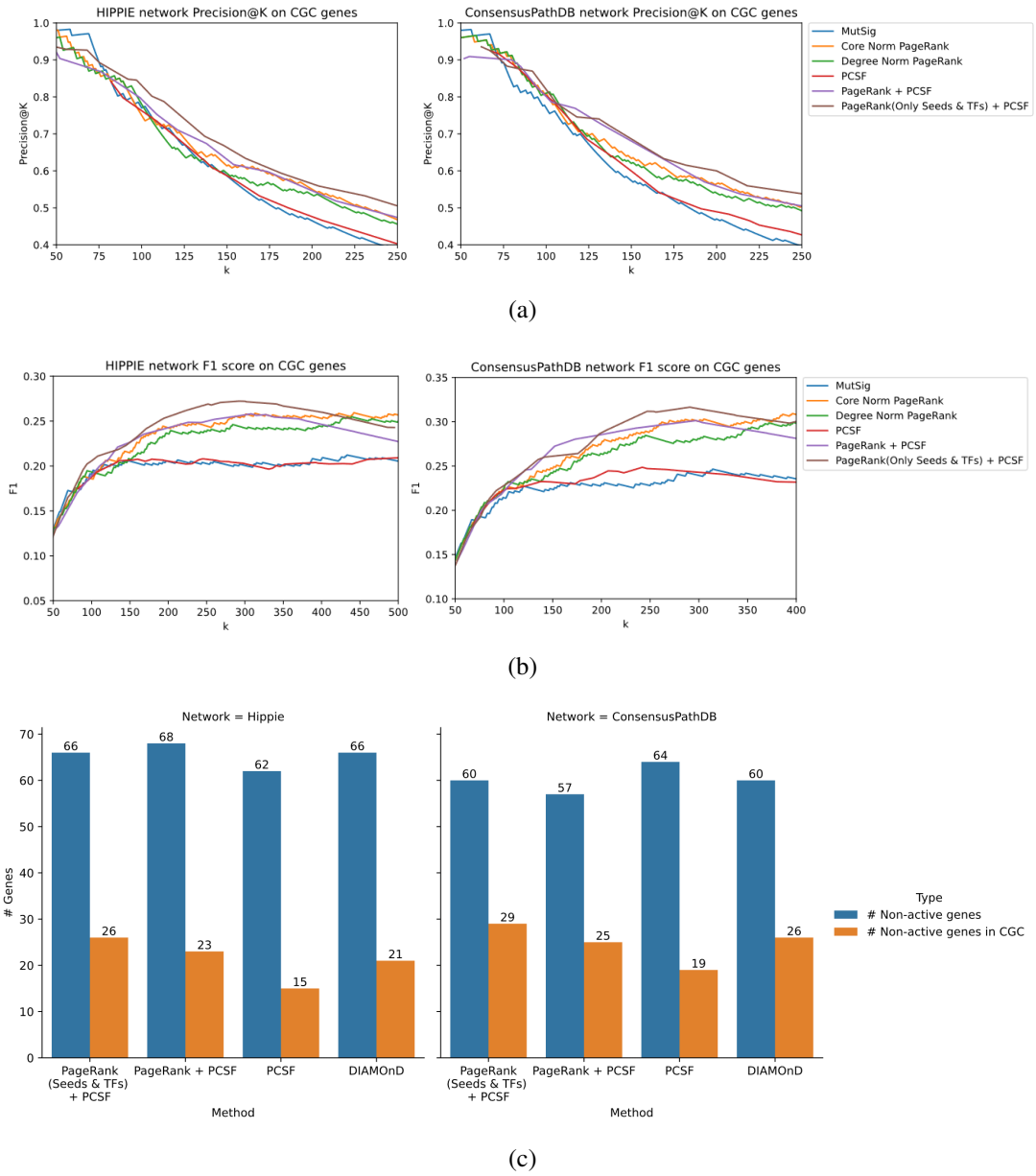


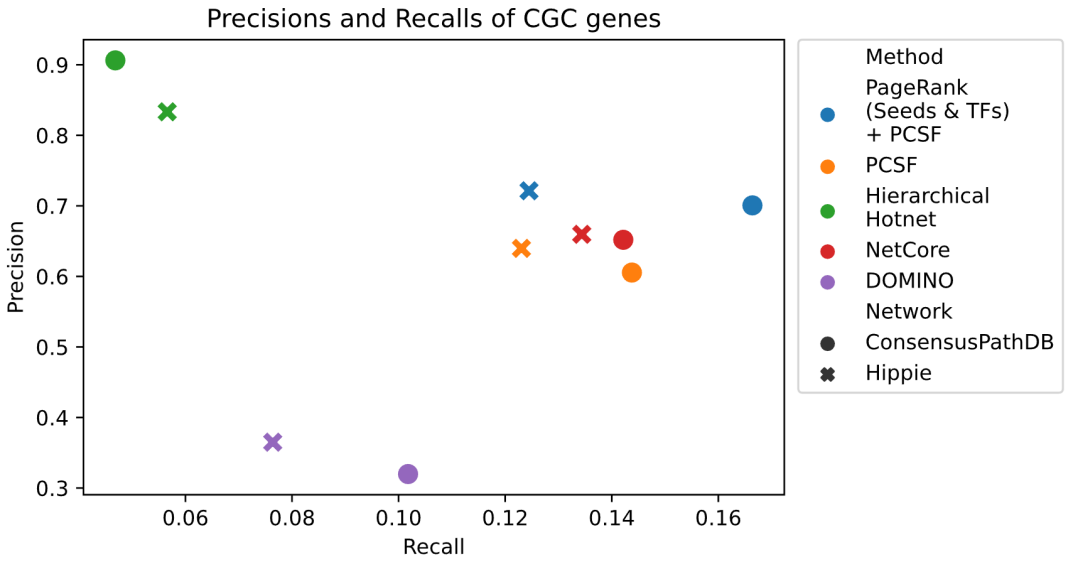
Figure 4.1: (a) Precision@K on known cancer genes for subparts of the algorithm.(b) F1@K on known cancer genes for subparts of the algorithm. (c) Number of non-active genes in the subnetwork & number non-active genes in the subnetwork that are known cancer genes

4.2 Comparison of Network Methods

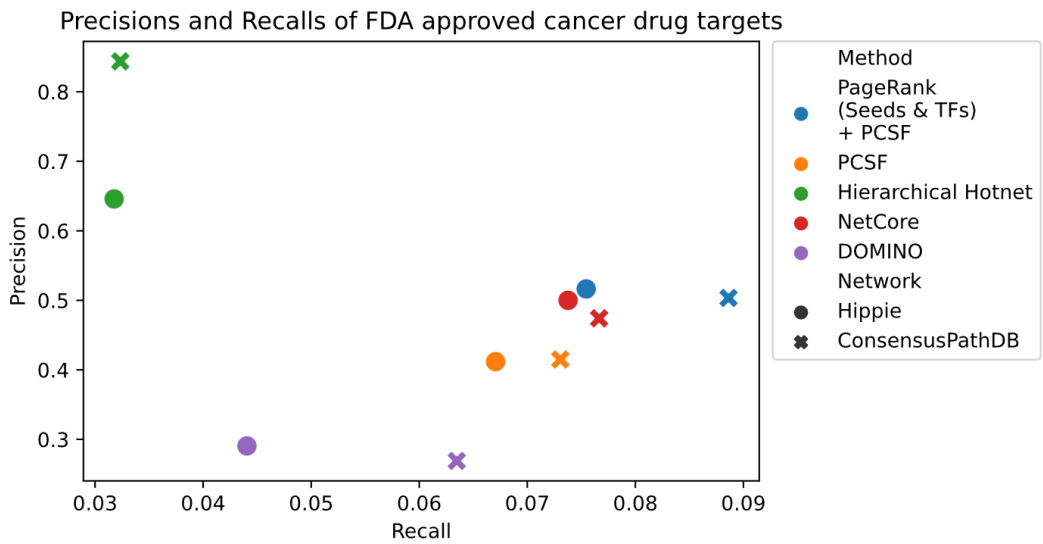
We compared PCSF, Hierarchical Hotnet, NetCore, DOMINO, and the proposed method using MutSig scores with two interaction networks: HIPPIE and ConsensusPathDB. The methods' ability to identify cancer associated genes that have type 1 and type 2 evidence in COSMIC were tested. Also, due to the incomplete nature of known cancer genes, the methods' ability to identify targets of FDA approved cancer drugs were analyzed. As shown in Figure 4.2, Hierarchical Hotnet has the highest precision in both networks, however the size of the subnetwork was very small and most of the nodes consist of active genes. On average, our method has higher precision and recall of identifying known cancer genes and targets of FDA approved cancer drugs compared to NetCore, DOMINO, and PCSF methods. Our method has higher average accuracy of finding known cancer genes in non-active genes on the 2 interaction networks compared to the other methods, number of non-active genes were illustrated in Figure 4.3 (a). While PCSF, NetCore, and our method have comparable numbers of non-active genes in the subnetworks, DOMINO has a significantly higher number of non-active genes in its modules. In the modules that were reported by DOMINO, it was observed that many non-active genes were nodes that are connected to a hub creating a star. Most of these non-active genes that were connected to a hub had low degrees, thus causing a significantly lower median degree compared to other methods as shown in Figure 4.3 (b). Hierarchical Hotnet and NetCore has degree penalization at their propagation step by doing a computationally expensive permutation based probability calculation using random degree preserving networks, our method penalizes the degree on PCSF step. Resulting subnetworks have comparable node degrees between these methods. Subnetwork that is created using our algorithm with HIPPIE network was visualized using Cytoscape in Figure 4. Colors were assigned to active genes, non-active genes, and non-active genes that are in known cancer genes. Subnetwork have a size of 121, around 30 % of the nodes are non-active genes which have $Q > 0.05$ in the MutSig data.

4.3 Functional Relevance Based Analysis

We have used DIGEST, a web service that performs in silico validation of subnetworks for functional relevance. DIGEST uses a random background subnetwork model and performs a permutation based test by applying enrichment analysis on known functional databases and reports an empirical p-value. We compared PCSF, Hierarchical Hotnet, NetCore, DOMINO, and the proposed method using the subnetworks that are reported for cancer from HIPPIE with DIGEST, results were shown in Figure 4.5. For the functional databases of GO biological processes and KEGG, all of the methods resulted with significant functional relevance. PCSF, NetCore, and the proposed method have resulted in significance for the 3 functional databases. NetCore, and the proposed method have the highest average functional relevances compared to other methods. Additionally, we have constructed subnetworks for 6 different gene expression data using our method with the HIPPIE PPI network. For each of the gene expression data, activity scores were randomly assigned under a uniform distribution 5 times and the method was applied on the random data to create 30 random subnetworks. We applied DIGEST to the original and random subnetworks, results were illustrated in Figure 4.6. For the 3 functional databases, median empirical p values were above the significance level for the subnetworks that were created by original activity scores, whereas median empirical p values for random subnetworks were below the significance level.

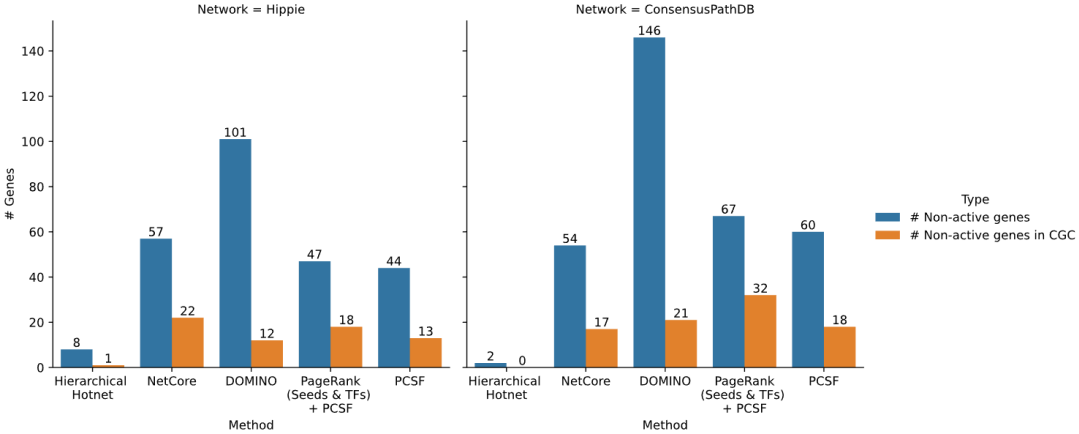


(a)

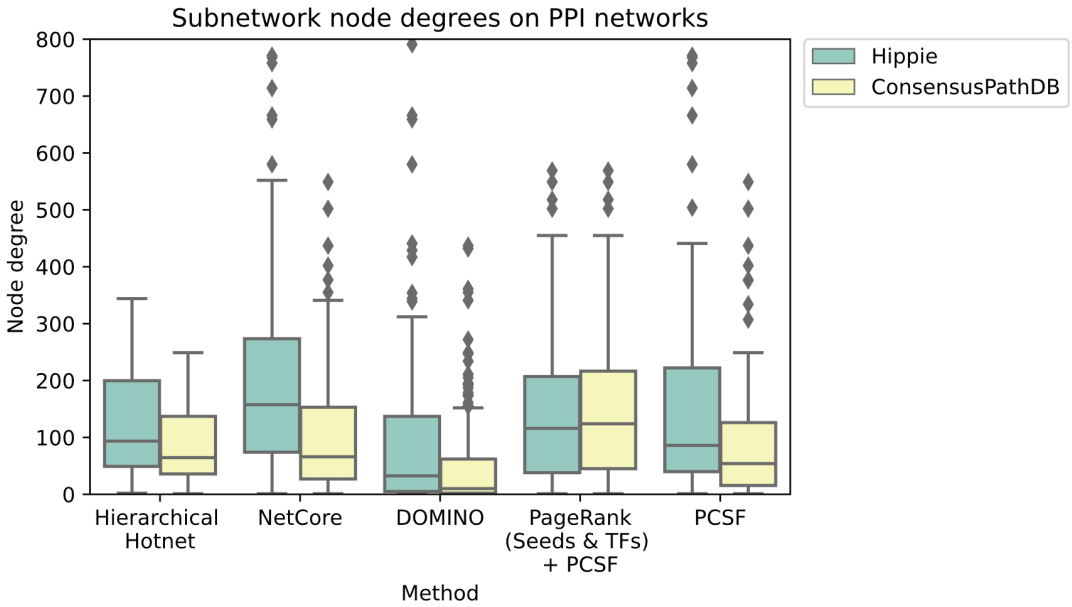


(b)

Figure 4.2: (a), Precision and recalls on known cancer genes using nodes of the subnetworks that were obtained from different methods. (b), Precision and recalls on known cancer drug targets using nodes of the subnetworks that were obtained from different methods



(a)



(b)

Figure 4.3: (a), Number of non-active genes in the subnetworks of the methods & number non-active genes in the subnetwork that are known cancer genes. (b), Degree distributions of the nodes in the subnetworks

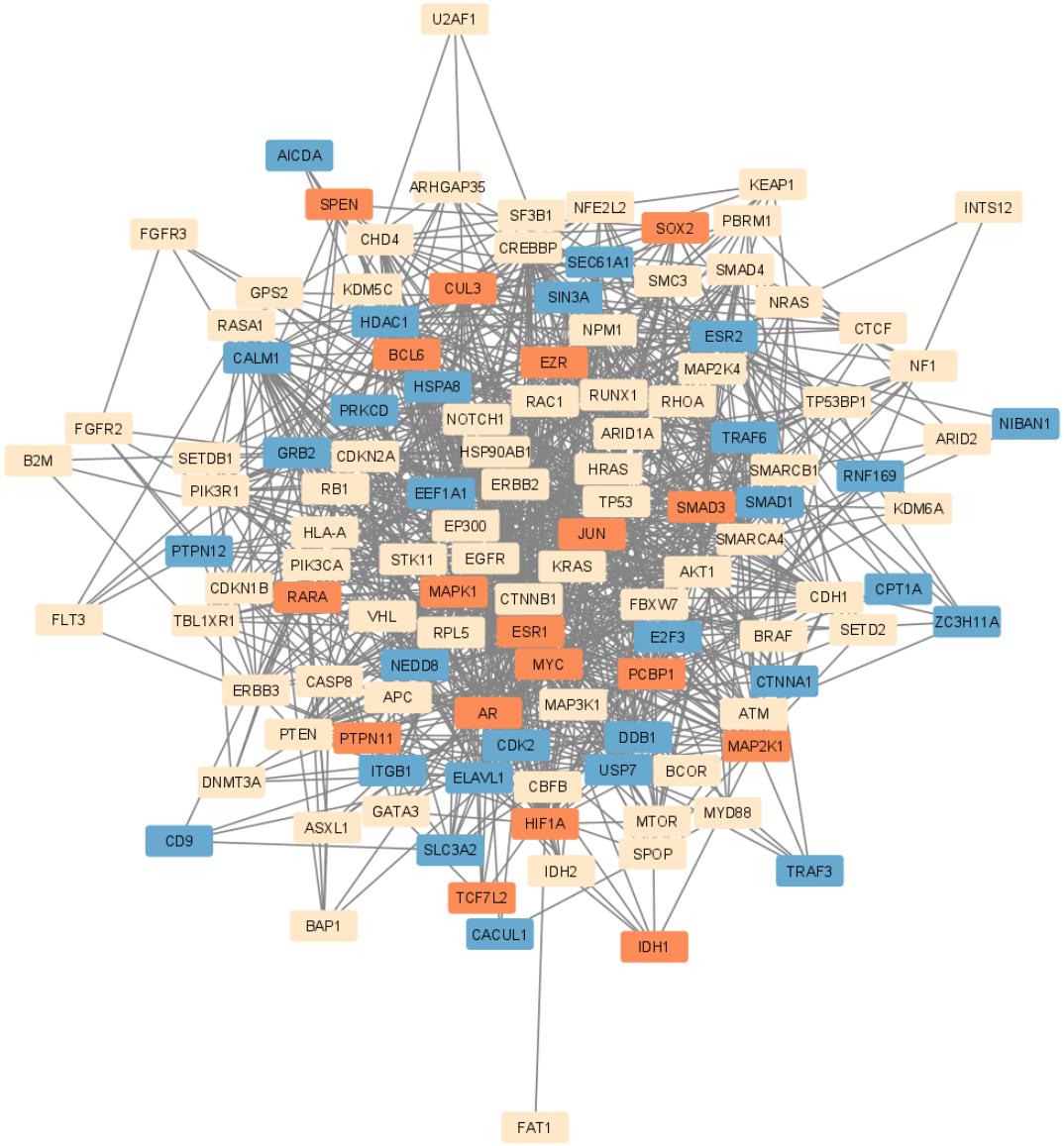


Figure 4.4: Subnetwork that is created using MutSig data, and HIPPIE network. Blue nodes are non-active genes, orange nodes are non-active genes that are known cancer genes, and the other nodes are active genes.

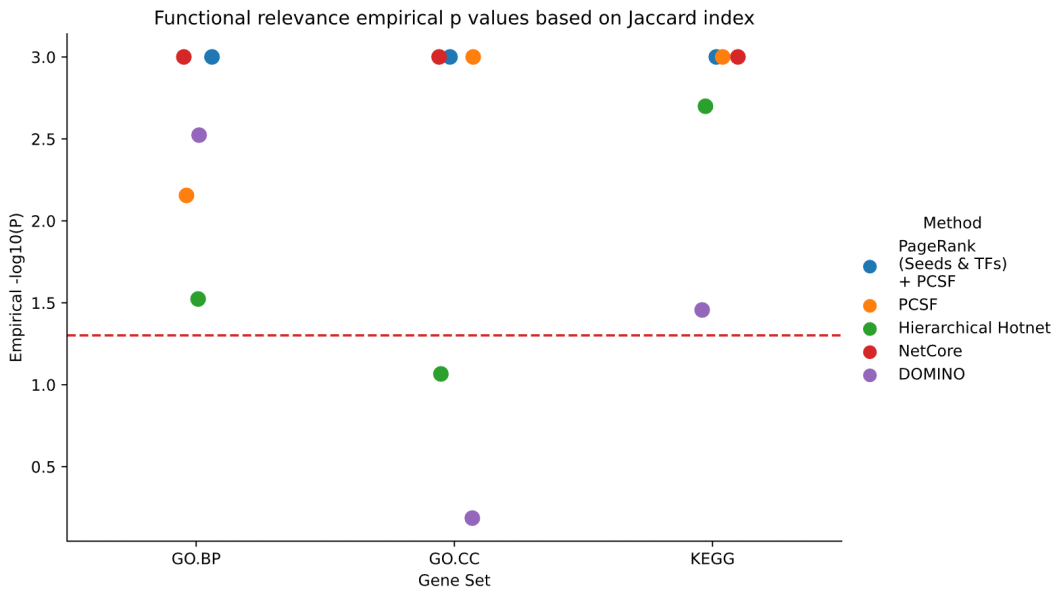


Figure 4.5: Functional relevance comparison of the methods

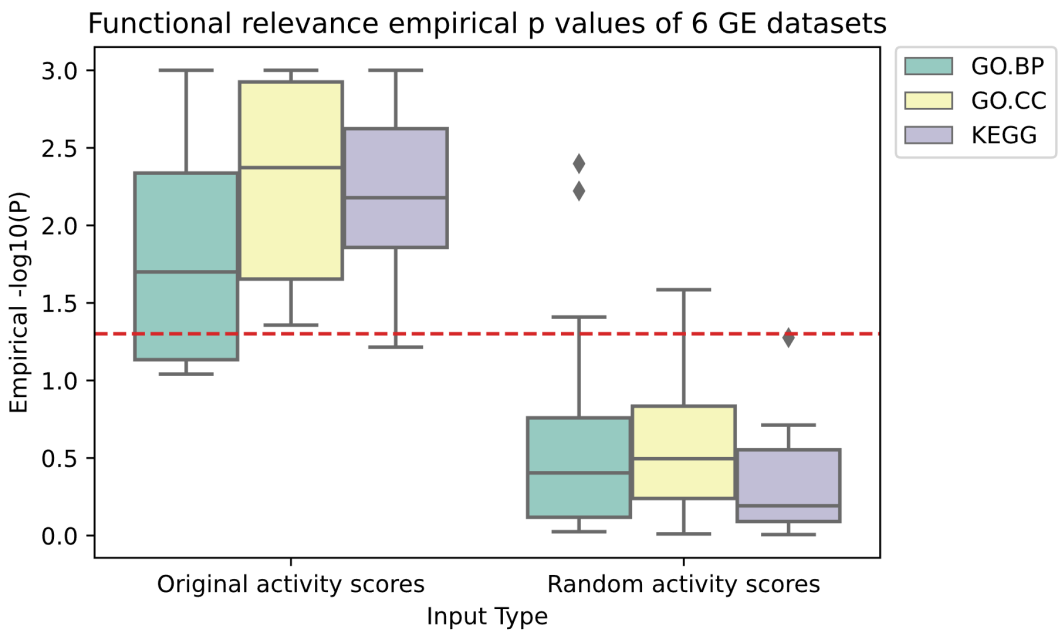


Figure 4.6: Functional relevance results of 6 gene expression subnetworks compared to random subnetworks.

4.4 Application of Propagated PCSF Approach to the Transcriptomic Data from Ischemic Heart Disease Patients

Cardiovascular diseases are the global leading cause of death, and ischemic heart disease also known as coronary artery disease is the most frequently observed cardiovascular disease. We have used differential expression analysis results between low & high severity ischemic heart disease groups for network analysis by applying the proposed method with HIPPIE network. Expressed genes were identified by checking whether over 90% of patients having > 0.5 TPM normalized expression value. Genes that were not labeled as expressed were removed from the HIPPIE network. Genes that have < 0.1 FDR in differential expression results were used as seeds, and top 100 genes were selected as terminals for PCSF after the network propagation step. Figure 4.7 shows the network of size 121 that was created using Cytoscape. Node colors represent the log fold change in the differential expression results, high values are upregulated genes in the high severity ischemic heart disease group. Node shapes depict whether the node is a steiner node or not. Genes in the subnetwork were searched in the DisGeNET database for entries in cardiovascular diseases. 19 genes in the subnetwork have cardiovascular disease related entries in the DisGeNET having GDA scores > 0.3 (1 or more curated sources). Differential expression results of these 19 genes were illustrated in Table 4.1. 7 of these genes were added to the subnetwork as steiner nodes. APP (Amyloid precursor protein) is one of cardiovascular disease related genes found in the network that is added as a Steiner node. APP has the highest eigenvector centrality in the network around 0.41, representing the amount of influence on the network. Amyloid-beta, which is a protein that is produced through the proteolytic processing of APP, has known to be a central protein for the development of dementia in elderly [56], and newer studies suggest that amyloid-beta may function as a link among cardiovascular diseases, alzheimer disease, and aging [57]. It is hypothesized that amyloid-beta has a pathophysiological role on cardiovascular disease, causing early vascular alterations and atherosclerosis that have a role in symptomatic coronary artery disease. Murine double minute 2 (MDM2) having the second highest eigenvector centrality (0.37), is a gene that functions as a E3 ubiquitin ligase that degrades p53, and is considered as a potential cancer drug target [58]. Since anti-cancer therapies cause cardiovascular complications, a research on MDM2 shows that upregulation of MDM2 may facilitate atherosclerosis, and MDM2 to be a central hub for stress response in the cardiovascular system is highly plausible [59].

Table 4.1: Genes in the subnetwork that are associated with cardiovascular diseases.

Gene name	Steiner node	logFC	FDR
ALDH6A1	Yes	-	-
REEP3	No	0.347	0.004
CYP1B1	No	1.023	0.006
DSP	No	-0.545	0.033
COL4A2	Yes	-	-
LOX	No	0.744	0.021
FCGR3A	No	0.860	0.038
LDHA	No	0.483	0.030
CALU	Yes	-	-
AR	No	-0.381	0.034
NRAS	No	0.434	0.013

Continued on next page

Table 4.1 – continued from previous page

Gene name	Steiner node	logFC	FDR
VAV3	No	0.459	0.028
SNTA1	Yes	-	-
APP	Yes	-	-
CRK	Yes	-	-
GCLM	No	0.390	0.033
PAPPA	No	0.470	0.016
TNC	No	1.836	0.002
ITGB1	Yes	-	-

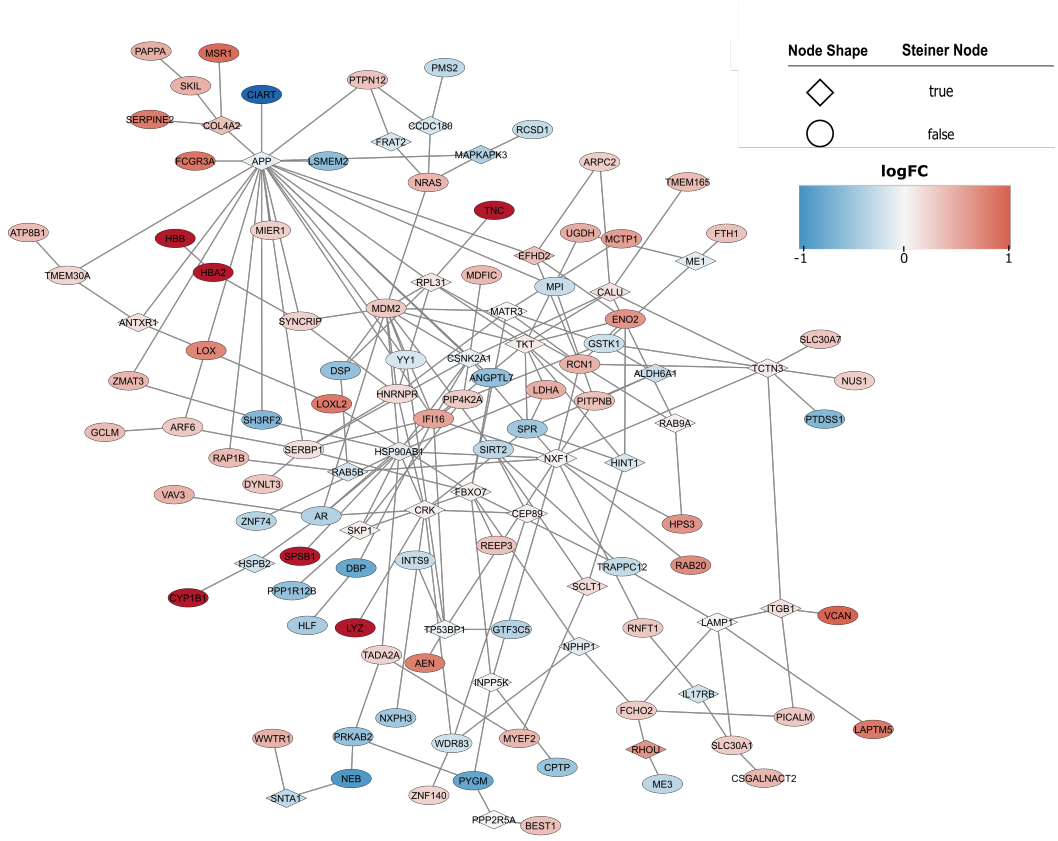


Figure 4.7: Subnetwork of high severity vs. low severity ischemic heart disease patients. Node shapes represent whether the node is a steiner node or not, node colors represent logFC values of differential expression.

Subnetwork that is generated for IHD patients, were separated into 11 modules using Louvain clustering. 11 modules are visualized in Figure 4.8, colors were assigned to the nodes of the modules by calculating average eigenvector centrality for each of the modules. Top 3 modules (cluster 4, 6, and 7) having the highest average eigenvector centrality, were used in functional enrichment analysis that is

performed by ShinyGO. Functional enrichment results for GO biological processes were illustrated in Figure 4.9. Most significant biological process for cluster 6, which has the highest average eigenvector centrality, is response to oxidative stress. Overproduced reactive oxygen species as a result of oxidative stress is known to be a cause of atherosclerosis, and abnormal production of these reactive oxygen species can lead to activation of matrix metalloproteinases, resulting with a rupture in atherosclerotic plaque [60]. Matrix metalloproteinases affected by reactive oxygen species, cause a remodeling in extracellular matrix which is another essential ischemic heart disease pathophysiology [61]. In cluster 7, the top enriched function is extracellular matrix organization, which is also relevant to ischemic heart disease. We have observed immune response related biological processes in the enrichment results of cluster 4. Fc receptor mediated stimulatory signaling pathway resulted as the most significantly enriched biological process in cluster 4. In preclinical models for cardiovascular diseases, activation of Fc-gamma receptors have resulted with the promotion of atherosclerosis [62].

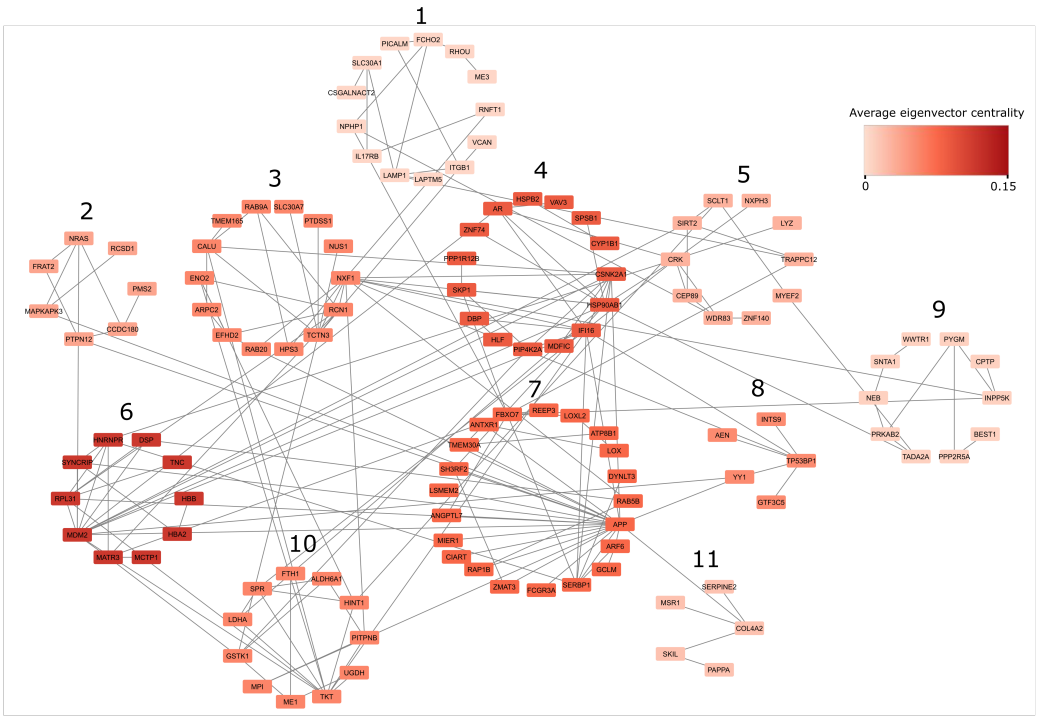
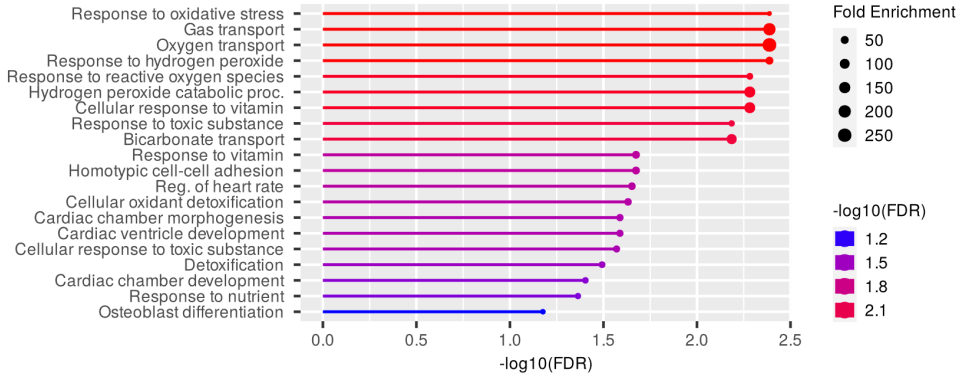
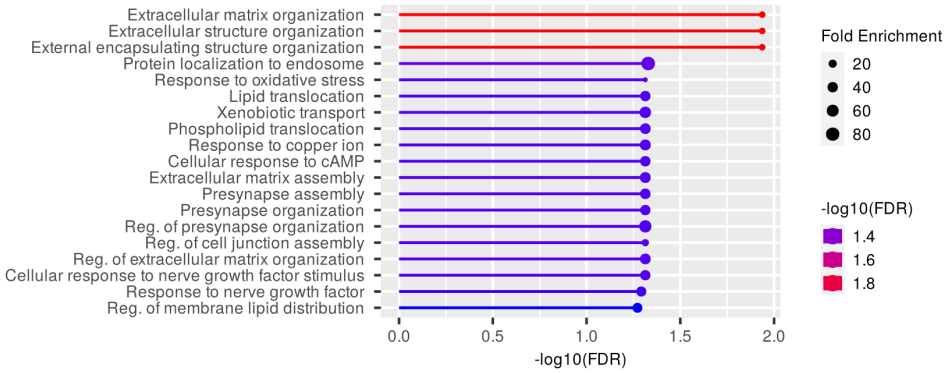


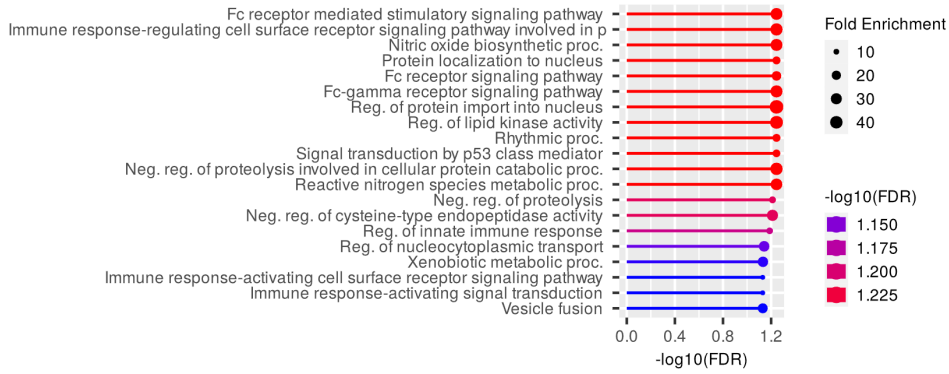
Figure 4.8: Cluster community representation of high severity vs. low severity ischemic heart disease subnetwork. Communities were mapped to a color by calculating average eigenvector centrality.



(a)



(b)



(c)

Figure 4.9: (a), (b), (c) Significant biological processes reported by ShinyGO (FDR < 0.1) for modules 6, 7 and 4 respectively.

CHAPTER 5

DISCUSSION AND CONCLUSION

In the field of network medicine, active / disease module identification methods have been applied increasingly in the last decade to guide biomedical research on finding hypotheses for disease progression, biomarkers that are related to a disease, and potential disease associated genes. In this study, we have presented an algorithm that utilizes network propagation and prize-collecting steiner forest algorithm to create a subnetwork that is relevant for a gene list that is supplied to the algorithm with a PPI network. Firstly, we have compared the full algorithm to its subparts to analyze whether the full algorithm gave more accuracy or not. Then, we compared the algorithm to other methods that were implemented for the same problem. We also performed another step of validation using 6 gene expression datasets and their randomized versions on a functional relevance test to examine the outcome. We concluded our analysis by applying the algorithm on ischemic heart disease patients to do a basic network analysis.

In the comparison of the algorithm to its subparts, overall precision on known cancer genes was higher for the full algorithm. Using a network propagation approach before PCSF added more true positives to the subnetwork than using only PCSF. Also, the increase in precision of using PCSF after PageRank suggests that PCSF can eliminate some false positives that come up from PageRank and can add false negatives from PageRank as steiner nodes to the subnetwork. Focusing on more densely connected regions with core normalized PageRank performed better than using classic degree based normalization. Full algorithm found more known cancer genes in non-active genes compared to unfiltered PageRank with PCSF. Using PageRank without constraint added more false positives from non-active genes to top scoring nodes. In the comparison of the algorithm to the other methods, our algorithm performed better on average in finding known cancer genes and known cancer drug targets. Our algorithm also added known cancer genes that were non-active in the data with higher average precision. DOMINO, which is a very robust algorithm, added nodes that are relatively low degree and had the lowest median degree compared to the other methods. The fact that it had the lowest median degree may have an effect on the low precision of finding known cancer genes. Degree distribution of our algorithm was comparable to the NetCore, and Hierarchical Hotnet, which are also propagation based algorithms that utilizes a permutation test as degree penalization step. In the functional relevance test, most of the methods ended up with significant results. Only NetCore and our algorithm had the highest significance in the functional relevance test in 3 functional datasets. Subnetworks that were created for 6 gene expression datasets had significant median functional relevances, and subnetworks created with randomized activity scores resulted as not significant in functional relevance. The functional relevance test suggests that our algorithm can create a functionally distinct subnetwork compared to other subnetworks from the random background model.

The algorithm was executed for low severity and high severity ischemic heart disease patients' gene expression data and HIPPIE network that was filtered by using only expressed genes. 19 of the genes in the subnetwork were found as related to cardiovascular diseases and nearly half of them were added as steiner nodes. Louvain clustering was applied on the graph and top 3 clusters having the highest average eigenvector centralities were used in functional enrichment analysis. Functional enrichment results have shown that our program can identify mechanisms that are highly relevant to ischemic heart disease. Functionally relevant mechanisms in the subnetwork can be utilized as another layer of information to generate hypotheses.

In summary, the proposed algorithm performed better compared to other methods in different tests, and the algorithm was able to create a subnetwork which had functionally relevant mechanisms related to the analyzed disease. We believe that our algorithm can be applied for disease research or biological research to obtain an insight on mechanisms that are relevant to the omic data. A key challenge in the active / disease module identification problem is that PPI networks are incomplete and biased. We believe our algorithm addresses the degree bias of the PPI networks with degree penalization, and more qualitative specialized networks may come up in future that can increase the performance of the algorithm. Moreover, we will work on the algorithm further to add an ability of incorporating multi-omic data to create a subnetwork. We believe utilizing multi-omic data to create a subnetwork will increase the performance considerably.

REFERENCES

[1] U. Kuzmanov and A. Emili, "Protein-protein interaction networks: probing disease mechanisms using model systems.," *Genome medicine*, vol. 5, no. 4, p. 37, 2013.

[2] M. P. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, and C. Wiuf, "Estimating the size of the human interactome," *Proceedings of the National Academy of Sciences*, vol. 105, no. 19, pp. 6959–6964, 2008.

[3] A. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[4] D. Li, Z. Pan, G. Hu, Z. Zhu, and S. He, "Active module identification in intracellular networks using a memetic algorithm with a new binary decoding scheme," *BMC Genomics*, vol. 18, Mar. 2017.

[5] F. Vandin, E. Upfal, and B. J. Raphael, "Algorithms for detecting significantly mutated pathways in cancer," *Journal of Computational Biology*, vol. 18, no. 3, pp. 507–522, 2011.

[6] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan, "Network propagation: a universal amplifier of genetic associations," *Nature Reviews Genetics*, vol. 18, no. 9, pp. 551–562, 2017.

[7] L. Lovász, "Random walks on graphs," *Combinatorics, Paul erdos is eighty*, vol. 2, no. 1-46, p. 4, 1993.

[8] H. Tong, C. Faloutsos, and J.-Y. Pan, "Random walk with restart: fast solutions and applications," *Knowledge and Information Systems*, vol. 14, no. 3, pp. 327–346, 2008.

[9] N. Tuncbag, S. J. Gosline, A. Kedaigle, A. R. Soltis, A. Gitter, and E. Fraenkel, "Network-based interpretation of diverse high-throughput datasets through the omics integrator software package," *PLoS computational biology*, vol. 12, no. 4, p. e1004879, 2016.

[10] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature*, vol. 402, no. 6761, pp. C47–C52, 1999.

[11] M. G. Kann, "Protein interactions and disease: computational approaches to uncover the etiology of diseases," *Briefings in bioinformatics*, vol. 8, no. 5, pp. 333–346, 2007.

[12] T. Ideker and R. Sharan, "Protein networks in disease," *Genome research*, vol. 18, no. 4, pp. 644–652, 2008.

[13] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[14] R. Saeed and C. M. Deane, "Protein protein interactions, evolutionary rate, abundance and age," *BMC bioinformatics*, vol. 7, no. 1, pp. 1–13, 2006.

[15] H. B. Fraser, A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman, "Evolutionary rate in the protein interaction network," *Science*, vol. 296, no. 5568, pp. 750–752, 2002.

[16] E. Eisenberg and E. Y. Levanon, "Preferential attachment in the protein network evolution," *Physical review letters*, vol. 91, no. 13, p. 138701, 2003.

[17] H. Yu, P. Braun, M. A. Yıldırım, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, *et al.*, "High-quality binary protein interaction map of the yeast interactome network," *Science*, vol. 322, no. 5898, pp. 104–110, 2008.

[18] I. K. Jordan, Y. I. Wolf, and E. V. Koonin, "No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly," *BMC evolutionary biology*, vol. 3, no. 1, pp. 1–8, 2003.

[19] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8685–8690, 2007.

[20] I. Feldman, A. Rzhetsky, and D. Vitkup, "Network properties of genes harboring inherited disease mutations," *Proceedings of the National Academy of Sciences*, vol. 105, no. 11, pp. 4323–4328, 2008.

[21] H. Goehler, M. Lalowski, U. Stelzl, S. Waelter, M. Stroedicke, U. Worm, A. Droege, K. S. Lindenberg, M. Knoblich, C. Haenig, *et al.*, "A protein interaction network links git1, an enhancer of huntingtin aggregation, to huntington's disease," *Molecular cell*, vol. 15, no. 6, pp. 853–865, 2004.

[22] S. E. Calvano, W. Xiao, D. R. Richards, R. M. Felciano, H. V. Baker, R. J. Cho, R. O. Chen, B. H. Brownstein, J. P. Cobb, S. K. Tschoeke, *et al.*, "A network-based analysis of systemic inflammation in humans," *Nature*, vol. 437, no. 7061, pp. 1032–1037, 2005.

[23] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[24] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[25] L. D. Wood, D. W. Parsons, S. Jones, J. Lin, T. Sjoblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, J. Ptak, *et al.*, "The genomic landscapes of human breast and colorectal cancers," *Science*, vol. 318, no. 5853, pp. 1108–1113, 2007.

[26] A. Ashley-Koch, Q. Yang, and R. S. Olney, "Sickle hemoglobin (hb s) allele and sickle cell disease: a huge review," *American journal of epidemiology*, vol. 151, no. 9, pp. 839–845, 2000.

[27] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, "Discovering regulatory and signalling circuits in molecular interaction networks," *Bioinformatics*, vol. 18, no. suppl_1, pp. S233–S240, 2002.

[28] K. Mitra, A.-R. Carvunis, S. K. Ramesh, and T. Ideker, "Integrative approaches for finding modular structure in biological networks," *Nature Reviews Genetics*, vol. 14, no. 10, pp. 719–732, 2013.

[29] H. Levi, R. Elkon, and R. Shamir, "Domino: a network-based active module identification algorithm with reduced rate of false calls," *Molecular systems biology*, vol. 17, no. 1, p. e9593, 2021.

[30] A. Mazza, K. Klockmeier, E. Wanker, and R. Sharan, "An integer programming framework for inferring disease complexes from network data," *Bioinformatics*, vol. 32, no. 12, pp. i271–i277, 2016.

[31] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," tech. rep., Stanford InfoLab, 1999.

[32] A. Garas, F. Schweitzer, and S. Havlin, "A k-shell decomposition method for weighted networks," *New Journal of Physics*, vol. 14, no. 8, p. 083030, 2012.

[33] M. H. Schaefer, J.-F. Fontaine, A. Vinayagam, P. Porras, E. E. Wanker, and M. A. Andrade-Navarro, "Hippie: Integrating protein interaction networks with experiment based quality scores," *PloS one*, vol. 7, no. 2, p. e31826, 2012.

[34] A. Kamburov, C. Wierling, H. Lehrach, and R. Herwig, "Consensuspathdb—a database for integrating human functional interaction networks," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D623–D628, 2009.

[35] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, *et al.*, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, no. 7457, pp. 214–218, 2013.

[36] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[37] S. F. Schmidt, B. D. Larsen, A. Loft, R. Nielsen, J. G. S. Madsen, and S. Mandrup, "Acute tnf-induced repression of cell identity genes is mediated by nfκb-directed redistribution of cofactors from super-enhancers," *Genome research*, vol. 25, no. 9, pp. 1281–1294, 2015.

[38] T. Ito, Y. V. Teo, S. A. Evans, N. Neretti, and J. M. Sedivy, "Regulation of cellular senescence by polycomb chromatin modifiers through distinct dna damage-and histone methylation-dependent pathways," *Cell reports*, vol. 22, no. 13, pp. 3480–3492, 2018.

[39] V. Miano, G. Ferrero, V. Rosti, E. Manitta, J. Elhasnaoui, G. Basile, and M. De Bortoli, "Luminal lncrnas regulation by erα-controlled enhancers in a ligand-independent manner in breast cancer cells," *International journal of molecular sciences*, vol. 19, no. 2, p. 593, 2018.

[40] H. Kroeger, N. Grimsey, R. Paxman, W.-C. Chiang, L. Plate, Y. Jones, P. X. Shaw, J. Trejo, S. H. Tsang, E. Powers, *et al.*, "The unfolded protein response regulator atf6 promotes mesodermal differentiation," *Science signaling*, vol. 11, no. 517, p. eaan5785, 2018.

[41] K. E. Connelly, T. M. Weaver, A. Alpsoy, B. X. Gu, C. A. Musselman, and E. C. Dykhuizen, "Engagement of dna and h3k27me3 by the cbx8 chromodomain drives chromatin association," *Nucleic acids research*, vol. 47, no. 5, pp. 2289–2305, 2019.

[42] J. A. Pulikkan, M. Hegde, H. M. Ahmad, H. Belaghzal, A. Illendula, J. Yu, K. O'Hagan, J. Ou, C. Muller-Tidow, S. A. Wolfe, *et al.*, "Cbfβ-smmhc inhibition triggers apoptosis by disrupting myc chromatin dynamics in acute myeloid leukemia," *Cell*, vol. 174, no. 1, pp. 172–186, 2018.

[43] S. Mulari, A. Eskin, M. Lampinen, A. Nummi, T. Nieminen, K. Teittinen, T. Ojala, M. Kankainen, A. Vento, J. Laurikka, *et al.*, "Ischemic heart disease selectively modifies the right atrial appendage transcriptome," *Frontiers in cardiovascular medicine*, vol. 8, 2021.

[44] V. Farooq, D. Van Klaveren, E. W. Steyerberg, E. Meliga, Y. Vergouwe, A. Chieffo, A. P. Kappetein, A. Colombo, D. R. Holmes Jr, M. Mack, *et al.*, "Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: development and validation of syntax score ii," *The Lancet*, vol. 381, no. 9867, pp. 639–650, 2013.

[45] S. X. Ge, "idep: An integrated web application for differential expression and pathway analysis," *bioRxiv*, p. 148411, 2017.

[46] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes, "The cosmic cancer gene census: describing genetic dysfunction across all human cancers," *Nature Reviews Cancer*, vol. 18, no. 11, pp. 696–705, 2018.

[47] P. Pantziarka, R. Capistrano I, A. De Potter, L. Vandeborne, and G. Bouche, "An open access database of licensed cancer drugs," *Frontiers in pharmacology*, vol. 12, p. 627574, 2021.

[48] C. O. Wilke, "Bringing molecules back into molecular evolution," *PLoS Computational Biology*, vol. 8, no. 6, p. e1002572, 2012.

[49] G. Barel and R. Herwig, "Netcore: a network propagation approach using node coreness," *Nucleic Acids Research*, vol. 48, no. 17, pp. e98–e98, 2020.

[50] M. A. Reyna, M. D. Leiserson, and B. J. Raphael, "Hierarchical hotnet: identifying hierarchies of altered subnetworks," *Bioinformatics*, vol. 34, no. 17, pp. i972–i980, 2018.

[51] K. Adamowicz, A. Maier, J. Baumbach, and D. B. Blumenthal, "Online in silico validation of disease and gene sets, clusterings or subnetworks with digest," *Briefings in Bioinformatics*, vol. 23, no. 4, p. bbac247, 2022.

[52] J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. I. Furlong, "The disgenet knowledge platform for disease genomics: 2019 update," *Nucleic acids research*, vol. 48, no. D1, pp. D845–D855, 2020.

[53] L. GUILLAUME, "Fast unfolding of communities in large networks," *Journal Statistical Mechanics: Theory and Experiment*, vol. 10, p. P1008, 2008.

[54] S. Paul, M. Andrew, and O. Owen, "Baliga nitin s, wang jonathan t, ramage daniel, amin nada, schwikowski benno, ideker trey. cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[55] S. X. Ge, D. Jung, and R. Yao, "Shinygo: a graphical gene-set enrichment tool for animals and plants," *Bioinformatics*, vol. 36, no. 8, pp. 2628–2629, 2020.

[56] E. H. Koo and S. L. Squazzo, "Evidence that production and release of amyloid beta-protein involves the endocytic pathway," *Journal of Biological Chemistry*, vol. 269, no. 26, pp. 17386–17389, 1994.

[57] D. A. Stakos, K. Stamatelopoulos, D. Bampatsias, M. Sachse, E. Zormpas, N. I. Vlachogiannis, S. Tual-Chalot, and K. Stellos, "The alzheimer's disease amyloid-beta hypothesis in cardiovascular aging and disease: Jacc focus seminar," *Journal of the American College of Cardiology*, vol. 75, no. 8, pp. 952–967, 2020.

[58] A. J. Levine and M. Oren, "The first 30 years of p53: growing ever more complex," *Nature reviews cancer*, vol. 9, no. 10, pp. 749–758, 2009.

[59] B. Lam and E. Roudier, "Considering the role of murine double minute 2 in the cardiovascular system?," *Frontiers in cell and developmental biology*, vol. 7, p. 320, 2019.

[60] D. Moris, M. Spartalis, E. Spartalis, G.-S. Karachaliou, G. I. Karaolanis, G. Tsourouflis, D. I. Tsilimigras, E. Tzatzaki, and S. Theocharis, "The role of reactive oxygen species in the pathophysiology of cardiovascular diseases and the clinical significance of myocardial redox," *Annals of translational medicine*, vol. 5, no. 16, 2017.

[61] G. L. Gallagher, C. J. Jackson, and S. N. Hunyor, "Myocardial extracellular matrix remodeling in ischemic heart failure," *Frontiers in Bioscience-Landmark*, vol. 12, no. 4, pp. 1410–1419, 2007.

[62] K. Tanigaki, N. Sundgren, A. Khera, W. Vongpatanasin, C. Mineo, and P. W. Shaul, "Fcγ receptors and ligands and cardiovascular disease," *Circulation research*, vol. 116, no. 2, pp. 368–384, 2015.