I, Cheng Zhu, hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Computer Science & Engineering.

It is entitled: Efficient network based approaches for pattern recognition and knowledge discovery from large and heterogeneous datasets

Student’s name: Cheng Zhu

This work and its defense approved by:

Committee chair: Kenneth Berman, PhD
Committee member: Anil Jegga, DVM, MRes
Committee member: Fred Annexstein, PhD
Committee member: Anca Ralescu, PhD
Committee member: Marepalli Rao, PhD
Efficient network based approaches for pattern recognition and knowledge discovery from large and heterogeneous datasets

By

Cheng Zhu

B. E., Huazhong University of Science and Technology, China, 2003
M. S., University of Cincinnati, USA, 2009

A dissertation submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science
at the University of Cincinnati
February 2013

Advisors & Committee Chairs: Dr. Kenneth Berman
Dr. Anil Jegga
Abstract

With rapid technological advances, the potential for transformational science and engineering for all scientific domains is enormous. Discovering useful and meaningful patterns and knowledge extraction from large, diverse, distributed and heterogeneous datasets however continues to pose a formidable challenge. Thus, there is an urgent need for more efficient and robust computational approaches to effectively manage, use, and exploit these heterogeneous data sources. This in turn can accelerate the progress of scientific discovery and innovation; gain new insights in a timely manner; lead to new fields of inquiry hitherto impossible. In this dissertation, we tackle this challenge by developing and applying novel and efficient network-based computational approaches. To demonstrate the utility of our algorithms, we use several large and heterogeneous datasets from biomedical domain, focusing specifically on rare or orphan diseases (OD) as an application. Our research has three facets:

First, we conduct a global network analysis of orphan diseases (OD) and demonstrate the utility of topological analyses in deducing the underlying biology for rare diseases and their causal genes. Specifically, starting with a bipartite network of known OD and OD-causing mutant genes, using the human protein interactome, functional enrichment and literature co-citation, we constructed and topologically analyzed several networks. Our analyzed results revealed that a majority of orphan disease-causing mutant genes are essential, in contrast to common disease-causing mutant genes, which are predominantly nonessential.
In the second facet, we designed a novel algorithm based on vertex similarity to identify and rank novel orphan disease candidate genes. We tested and validated this algorithm using leave one out cross-validation approach on known orphan disease gene sets. We also compared its performance with previously reported similar approaches and found that its performance was comparable to the current state-of-art approaches.

Finally, we designed and developed a novel drug repositioning candidate discovery framework that combines both information theory and network analyses-based approaches. Integrating fourteen heterogeneous gene-gene networks, this framework quantifies similarities between disease causal genes and drug target genes based on topological similarity (vertex similarity score) and mutual information score. By extracting the related drug and disease information from the top ranked gene pairs or gene clusters, we discovered several drug repositioning candidates for both common and orphan diseases.
Acknowledgement

I would like to express my acknowledgement to advisors Dr. Anil Jegga and Dr. Kenneth Berman for their supervision and support throughout my graduate study. This dissertation would not have been possible without them. I am also grateful to my other dissertation committee members, Dr. Anca Ralescu, Dr. Fred Annexstein, and Dr. Marepalli Rao, for their guidance and helpful feedbacks.

I am thankful to the Department of Computer Science, University of Cincinnati, and the Division of Biomedical Informatics, Cincinnati Children’s Hospital Medical Center for supporting my research. I am also thankful to the faculty Dr. Zhongming Zhao, Dr. Bruce Aronow, Dr. Qing-An Zeng, and Dr. Zhang Bin for their assistance and support on research projects and publications.

Many thanks to my friends and colleagues: Dr. Minlu Zhang, Dr. Zhen Hu, Chao Wu, Divya Sardana, Akash Kushwaha, and many others from University of Cincinnati for collaborations and stimulating discussions.

And last but not least, I would like to thank Qin Li and my family for their constant support, encouragement, and love.
Publications arising from this thesis

Journal Articles:

Zhu C, Wu C, Aronow B, and Jegga AG. Heterogeneous network analyses to identify drug repositioning candidates for orphan diseases (in preparation)


Book Chapters:


Abstracts:

## Contents

Chapter 1: Introduction .....................................................................................................................1  
1.1 Motivation .....................................................................................................................1  
1.2 Contribution of this thesis .............................................................................................2  

Chapter 2: Background and Related Work ........................................................................................4  
2.1 Network Approaches for Biomedical Systems .................................................................4  
2.2 Network Analysis Concepts ............................................................................................5  
  2.2.1 Measurements in Network ........................................................................................5  
  2.2.2 Sub-network, Modules and Communities ...................................................................9  
2.3 Disease Gene Prioritization and Prediction Through Computational Approaches in Networks ..................................................................................................................................................12  
2.4 Drug Repositioning ..........................................................................................................15  
  2.4.1 Drug Repositioning Principles and Strategies ..........................................................16  
  2.4.2 Network and Computational Approaches for Drug Repositioning ..........................17  

Chapter 3: The Orphan Disease Network Analysis .........................................................................20  
3.1 Introduction .......................................................................................................................20  
3.2 Method ................................................................................................................................22  
  3.2.1 Resources of Data and Analysis .............................................................................22  
  3.2.2 Measures of Topological Importance, Connected Components and Modularity ...23  
3.3 Results ..................................................................................................................................24  
  3.3.1 Constructing Networks of Orphan Diseases, Mutant Causal Genes and Protein Interactions ........................................................................................................................................................24  
  3.3.2 Orphan Disease-causing Mutant Gene Interactome (ODMGI) ...................................28  
  3.3.3 Orphan Causing Mutant Genes Encode Proteins Show Tendency to Be Essential 33  
  3.3.4 Orphan Disease Networks based on Shared Functional Terms .................................35  
  3.3.5 Orphan Disease Networks with Shared Literature ....................................................37  
3.4 Discussion and Conclusion ...............................................................................................40  

Chapter 4: Analyzing the Orphan Disease Interactome to Find Novel Drug Targets and Treatments ....................................................................................................................................................45  
4.1 Introduction .......................................................................................................................45  
4.2 Methods ................................................................................................................................48  
  4.2.1 The Vertex Similarity Ranking Algorithm ..................................................................48  
  4.2.2 Prioritization Performance Evaluation Methods .......................................................52  
  4.2.2.1 Data Resource .....................................................................................................52  
  4.2.2.2 Comparing Ranking Performance Comparison with Other Algorithms ..............52  
  4.2.2.3 Collecting OD Candidate Genes as Test Set for Identifying and Ranking ..........56  
4.3 Results ..................................................................................................................................57  
  4.3.1 Leave One Out Cross Validation Results ...................................................................57  
  4.3.2 Using VS Based Approach to Identify and Rank Novel OD Candidate Genes ..........60  
4.4 Discussion and Conclusion ...............................................................................................63  

Chapter 5: Drug Repositioning Through an Integrative Framework Across Multiple Heterogeneous Networks ........................................................................................................................................65  
5.1 Background .......................................................................................................................65  
5.2 Material and Methods .......................................................................................................68  
  5.2.1 Gene Pairs and Biological Networks .......................................................................68
5.2.2 Mutual Information (MI) Between Genes .............................................................. 71
5.2.3 Vertex Similarity (VS) Between Genes .............................................................. 72
5.2.4 Filtering Gene Pairs ............................................................................................ 73
5.2.5 Combining MI and VS to Construct Different Weighted Networks ................. 73
5.2.6 Clustering Different Weighted Networks and Meta Cluster Associating ............ 74
5.2.7 Recovery Rate and Literature Search for Validations ........................................ 77

5.3 Results and Analysis ............................................................................................. 78
  5.3.1 Known Indications Dataset ................................................................................ 78
  5.3.2 OD-All-Drugs Dataset ....................................................................................... 80

5.4 Conclusions .......................................................................................................... 84

References ..................................................................................................................... 87
Chapter 1: Introduction

1.1 Motivation

In the past decade, technological advances in the biomedical and genomic fields have generated a huge amount of “omics” data, posing a great challenge to mine them and extract knowledge to cater both existing and unmet medical needs. Developing and applying efficient computational approaches on data and pattern mining thus continues to be a hot field of research in today’s computer science and biomedical informatics studies.

The various types of the “omics” data, such as protein interactions, gene expressions, gene to disease relations and disease to drug relations, etc., can be viewed from a network perspective, where a node represents a biological entity and an edge its annotations and/or associations. Because networks in biomedical domain have been found to be comparable to communication and social networks [1] through commonalities such as scale-freeness and small-world properties, the algorithms used for social and Web networks should be equally applicable to biomedical networks. The network approaches have been proven useful in discovering meaningful patterns in previous studies because a network of biomolecules can provide a systems-level view of the corresponding system. Retrieving useful patterns and knowledge from these systems are important as patterns in the networks may uncover the underlying biomedical characteristics and provide possible solutions for medical treatment.
Therefore, in this thesis, we aim to develop and apply efficient computational methods on pattern and knowledge discovery through networks analysis.

We designed, developed and applied network-based approaches on a relatively understudied biomedical domain, namely, rare or orphan diseases. There have been relatively few efforts focusing on a large scale global analyses of orphan diseases. The lack of scientific knowledge and quality information on several of these diseases often results in a delay in diagnosis. By focusing on multiple orphan diseases that are related via biomedical networks (e.g., shared causal genes, pathways or processes), we aim to exploit the “guilt-by-association” principle to analyze orphan diseases, identify and rank causal genes, and identify drug repositioning candidates.

1.2 Contribution of this thesis

In this thesis, we propose novel network-based computational approaches to analyze large, heterogeneous biomedical networks, focusing on orphan/rare diseases. Although we apply our approaches for the analysis of datasets from biomedical domain, they can be readily adapted to other domains (social network, mobile network, etc.).

In Chapter 2, we review related literature and present, in brief, several network concepts. In chapter 3, we present methods and results from a global analysis of all orphan diseases. Briefly, starting with a bi-partite network of all orphan diseases and
orphan disease-causing mutant genes, we build disease-disease (shared causal gene),
gene-gene (shared orphan disease), and protein interaction networks and analyze them
specifically to understand the biological underpinnings of orphan. In chapter 4, we
propose a novel and efficient framework to address a relatively unmet need in
biomedical domain. Specifically, extending on our previous work [2], we design and
apply a novel algorithm to the problem of discovery and ranking of potential orphan
disease-causing genes. We validated our approach through cross-validation studies
using known orphan disease-causing genes and also compare it with other existing
methods in the field. Finally, we present a novel integrated heterogeneous network
mining framework to identify drug repositioning candidates for orphan diseases in
Chapter 5. Briefly, similarity scores between approved drug target genes and disease
causal genes calculated based on network similarity score and mutual information.
The network similarity score is calculated by vertex similarity in several
heterogeneous networks while the mutual information score is calculated based on the
frequencies of the genes and pairs in these networks. By extracting the related drug
and disease information from the top ranked gene pairs or gene clusters, potential
drug repositioning candidates are discovered.
Chapter 2: Background and Related Work

2.1 Network Approaches for Biomedical Systems

Several recent studies on biological systems have approached them as complex networks comprising bioentities (genes, biomolecules, etc.) as nodes and their functional interactions as edges [3, 4]. The biomedical system and data can be processed and constructed to networks such as genes and protein interactions between them, signal pathways and functional annotations between molecules, gene to disease relationships, and drug to disease relationships [5-10]. The concepts from computer science, graph theory and network are widely and successfully adapted to understand the biomedical systems. For example, network analyses-based approaches have been successful in protein function predictions, drugs and target discovery and novel biomarker identification, to name a few [11-13]. Motivated by these research findings, we applied similar approaches to one of the relatively understudied fields in biomedical and health sciences – orphan or rare diseases. Using available knowledge about these diseases as a starting point, we construct the network of orphan diseases (OD) and their mutant genes (see Figure 2.1) and analyze them in great detail using network analyses-based approaches. Each of the ODs or OD-causing mutant genes (ODMG) is regarded as a node/vertex in these networks. The edges on the other hand represent the biological associations and/or interactions between nodes. For instance, an edge may represent shared genes between two ODs or shared ODs between two genes. They can also be physical interactions such as protein-protein interactions or
common functionalities (e.g., shared biological processes, pathways or phenotypes).

2.2. Network Analysis Concepts

2.2.1 Measurements in Network

Several network measurements are typically used to characterize the properties of a network ranging from node degree to centrality measures to their modular or community structure. These network measurements have been extensively employed to analyze the properties of large-scale networks in biomedical systems to extract biologically meaningful patterns and knowledge. In the following sections, we briefly describe some of the relevant network features.
**Degree**

Node degree indicates the number of direct connections (edges) that a node has. When a network is directed, it has two respective measures of degree, namely in-degree and out-degree. In-degree represents the number of edges that directed to a node while out-degree as the name indicates represents the number of outgoing connections or edges from a node.

**Centrality**

Centrality in general is used a measure to calculate the “importance” of a node in a network. Three common centrality measures apart from the node degree that are commonly used are: betweenness centrality, closeness centrality and eigenvector centrality.

A node’s betweenness centrality, also referred to as the shortest path betweenness, is the ratio of the number of the shortest paths that pass through a node to the number of the shortest path between any pair of nodes in the network [14].

Closeness centrality is a measurement on how close a node is connected to all other nodes of the network [15, 16]. A node’s closeness is defined as the inverse of sum of the node’s distances to all other nodes in the network. Closeness can be defined as a measurement of how fast the information originated from a node can spread to all other nodes sequentially.
Eigenvector centrality is a measurement of the influence of a node in a network. It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node than connections to low-scoring nodes. Google's PageRank[17] is a variant of the Eigenvector centrality measure.

**Shortest path length**

The shortest path length, sometimes also called geodesic distance, is the smallest number of edges or “hops” that are required to pass from one node to another.

**Clustering coefficient**

Clustering coefficient is a measurement of degree to nodes that tend to cluster together in a network [18]. Clustering coefficient indicates how densely the neighbors of a node are connected. The average clustering coefficient for nodes in a network represents the tendency to form clusters. It is also an indicator of the modularity in a network.

**Density**

The density measures the extent of the contacts between pairs of nodes of the network [19]. The density is computed as the proportion of contacts that could possibly occur in the network compared with those that are actually observed in the network.

Table 2.1 summarizes several commonly used network measurements along with equations to calculate.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree(with In-degree and Out-degree)</td>
<td>*Undirected networks: $D =$ the number of</td>
</tr>
</tbody>
</table>
edges connected to a node.

*Directed networks: $D_{in} =$ the number of directed-in edges; $D_{out} =$ the number of directed-out edges of a node.

*Average degree $<D> = 2E/N$, $E$ is the total number of edges, $N$ is the total number of nodes, and $<>$ denotes average.

<table>
<thead>
<tr>
<th>Centrality Measure</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Betweenness centrality</strong></td>
<td>$B_{ci} = \frac{\sum #SP_{through\ node\ i}}{#SP}$</td>
</tr>
<tr>
<td><strong>Closeness centrality</strong></td>
<td>$C_{ci} = \frac{1}{\sum_{j=1}^{N} d(n_i, n_j)}$</td>
</tr>
<tr>
<td><strong>Eigenvector centrality</strong></td>
<td>$E_i = \frac{1}{\lambda} \sum_{t \in M(i)} E_t = \frac{1}{\lambda} \sum a_{i,t} E_t$</td>
</tr>
<tr>
<td><strong>Shortest path length</strong></td>
<td>$L_{ij} =$ the smallest number of edges between nodes $i$ and $j$.</td>
</tr>
<tr>
<td><strong>Clustering coefficient</strong></td>
<td>$Cc = \frac{2n}{E(E-1)}$</td>
</tr>
<tr>
<td><strong>Density for undirected network</strong></td>
<td>$D = \frac{2E}{N(N-1)}$</td>
</tr>
<tr>
<td><strong>Density for directed network</strong></td>
<td>$D = \frac{E}{N(N-1)}$</td>
</tr>
</tbody>
</table>
Using statistics of network measurements, influential nodes in the network can be easily identified. For example, nodes with high degree and high global betweenness centrality are defined as hubs and bottlenecks respectively. Figure 2.2 shows the examples of hubs and bottlenecks, which are the most commonly studied important nodes in biomedical networks.

![Hubs and bottlenecks in the network](image)

**Figure 2.2 Hubs and bottlenecks in the network**

### 2.2.2 Sub-network, Modules and Communities

Sub-network, or connected component, or loosely connected network, is a way to interconnect the components in a system or network so that those components, depend on each other to the least extent. A loosely connected network can be easily decomposed into definable modules.
A module, on the other hand, is a group of nodes that are more densely connected to each other than to the nodes outside the group in a network[20]. In biomedical domain, networks are composed of modules[3]. A functional module contains a group of molecules and functional associations to perform a specific function[21]. In a biomedical network, functional modules are believed to comparable to topologically densely connected modular structures that are relatively isolated from other parts of the network[5].

A community, similarly, is a densely connected sub-network within a large network, such as a close-knit group of genes in a protein interaction network or a group of close friends in a social network. Figure 2.3 demonstrates the structure of loosely connected networks, modules and communities in a network.

Figure 2.3 Demonstrations of loosely connected networks, modules and communities
Finding modules or clusters in a network is a primary task and a widely used technique in many fields, including machine learning, information retrieval, pattern recognition and bioinformatics. The extraction and analysis of clusters in large-scale biomedical networks have been extensively studied in protein interaction networks, functional enrichment networks, and regulatory networks [20, 22, 23].

Network clustering approaches can be generally divided to several categories, such as: connectivity based clustering (hierarchical clustering), centroid-based clustering, density-based clustering and centrality based clustering.

Connectivity based clustering is also known as hierarchical clustering. The basic idea is that objects are more related to nearby objects than to objects that are farther away. These algorithms connect "objects" to form "clusters" based on the pair-wise distance [24-26].

A typical centroid-based clustering algorithm is K-means, where the number of clusters in the network is fixed to $k$. K-means clustering tries to find the $k$ cluster centers and allocate the objects to their nearest cluster center, such that the squared distances of each objects to their cluster center are minimized [27].

A density-based clustering approach tries to find higher densely connected components than the remainder of the network. Sparsely distributed objects in the
areas are usually considered to be noise or border points[28]. The most popular density based clustering method is DBSCAN[29].

Centrality-based clustering methods start by removing an edge that has the highest centrality measure (e.g. the edge betweenness centrality), the edge removal procedure will be repeated iteratively until a network has been partitioned to several clusters after a fraction of edge have been removed [5, 30].

Besides these main categories, other clustering approaches have also been developed and applied to network research, for example, the Louvain method for community detection[31]. This method employs a greedy optimization method which attempts to iteratively optimize the "modularity" of a partition from the network, the procedure stops until a maximum of modularity is attained and a hierarchy of communities is generated.

For analyzing biomedical networks, different network clustering approaches may have different objectives, such as detecting signaling pathways, functional units, or other defined sub-network structures.

### 2.3 Disease Gene Prioritization and Prediction Through Computational Approaches in Networks

The identification of disease causal genes is a primary challenge in genomic and
translational biomedical research. Traditional approaches such as positional cloning are not effective for complex disease-associated gene discovery[32]. A recent developed trend is to use computational approaches that integrate multiple sources of features of human diseases and the genome (e.g., gene functions, protein interactions, etc.), and use them as references to screen the whole genome for the most potential novel disease candidate genes. Several computational approaches for ranking disease candidate genes have been proposed based on the assumption that one disease can be caused by a group of functionally related genes. The functional similarities of candidate genes to known disease causal genes are measured in such approaches.

Among these approaches, network-based ones have demonstrated a good utility. One common research hypothesis of network-based methods is that a disease’s causal genes are likely to locate to each other closer in a biological network [33, 34]. In other words, the main concept is based on a "guilt-by-association" hypothesis, that is, the probability of the neighbor of a disease gene to cause either the same or a similar disease is higher[9, 33, 35, 36]. When one or more genes are already implicated with one disease, it is probably that their directly associated neighboring genes also have close relationships with the disease.

The prioritization for disease candidate genes of some network-based methods was based on the fact that whether they directly interact with known disease causal genes [37, 38]. Other methods took the shortest-path distance between candidate genes and
known disease causal genes into consideration if there are no direct connections between them [36, 39]. Xu et al.[6] built a K-nearest neighbors-based classifier using all known disease causal genes from the OMIM (Online Mendelian Inheritance in Man) database, the author concluded that their approach reveals that the disease causal genes from OMIM are characterized by a larger degree, and show a tendency to interact with other disease genes with more common neighbors, and have quick communication with each other. In another study, Chen et al.[37] proposed a graph connectedness-based scoring system to obtain other Alzheimer’s disease causal genes according to a list of already known target genes of Alzheimer.

On the other hand, different methods may propose distinct scoring frameworks for candidate gene prioritization. Recently, Kohler et al. [40] used a social and web network algorithm Random Walk(RW) to rank and identify disease candidate genes. Vanunu and Sharan [41] proposed a propagation-based approach that integrates information of known disease causal genes and confidence scores of protein interactions to prioritize the candidate genes. Lage et al. [15] introduced a Bayesian predictor which combines the protein interactome and phenotype to infer putative protein complexes that are likely to associate with a disease. Wu et al. [42] developed the CIPHER method using a regression model of phenotype similarity and gene closeness to score the candidate genes. Other network-based approaches, including page rank with priors[43], network partition[44], and network clustering [45], were also introduced to rank disease candidate genes. In all, it becomes a central dogma for
disease candidate gene discovery methods that close genes in a network will cause similar diseases. 

Although different approaches are used to solve disease gene prediction and prioritization problem, the key idea remains the same – all of the approaches involve a set of candidate genes as test set, along with a set of known disease causal genes as training set, and the analysis are based on the physical or functional network connectivity between them. “De novo” approaches that do not depend on prior knowledge of disease genes are yet to be developed [34]. In this thesis, we also proposed a novel computational approach for orphan disease causal gene ranking and predictions, see chapter 4 for more details.

2.4 Drug Repositioning

Drug discovery and development in general are long-term, complicated and expensive procedures, whereas the success rate is extremely low. To overcome or by-pass this productivity gap, more and more companies are resorting to “Drug Repositioning" or "Drug Repurposing", or simply identifying and developing new treatments for existing or abandoned pharmacotherapies[46]. In other words, drug repositioning is the procedure to explore new therapeutic indications for existing drugs[47], which is an alternative and efficient strategy to boost the discovery of disease therapeutics. The hypothesis is that by reshuffling what are already known about diseases and approved drugs in novel and interesting ways, the potential new indications that may
lead to better therapies could be discovered[48]. Such approaches can significantly reduce the risks and failures that accompanied with drug development. Repositioned drugs can enter clinical phases with a shorter time and at a lower cost than novel compounds because the starting points are usually approved compounds with known bioavailability and safety profiles, proven formulations and manufacturing routes, and well-characterized pharmacology[49]. It is therefore not surprising that in recent years, of the new drugs that reach their first markets, repositioned drugs have taken up to a percentage of ~30%!

2.4.1 Drug Repositioning Principles and Strategies

Conceptually, drug repositioning is motivated by two fundamental scientific principles. First, drugs by nature are “promiscuous”, i.e., a single drug can bind multiple targets and/or impact several pathways. Second, drug targets relevant to a specific disease or pathway can also play critical roles in other related or unrelated diseases or pathways [48, 50]. Some commonly used drug repositioning strategies are listed as follows:

**Knowledge base-based screening**

The fundamental approach in this strategy is that various accumulated heterogeneous data sets, such as pharmacological, biomedical, genomic, and chemical data, will be integrated and mined by novel computational and analytical algorithms. The “virtual screening” is performed to discover hidden or non-explicit relations between a drug, target gene, and disease.
**Pharmacopeia scanning**

This unbiased strategy re-screens the existing compounds against a variety of targets by a semi-blind approach to identify possible therapeutic benefits or side-effects, the recent advanced screening technologies are adopted.

**Screening for specific phenotypes**

The strategy starts from a point of an interested phenotype and the existing drugs that are screened to discover drugs that may generate an unexpected (either as on- and/or off-target effects of compounds), but yet desired phenotypic result.

### 2.4.2 Network and Computational Approaches for Drug Repositioning

There are experimental and computational approaches for drug repositioning. The experimental approaches for drug repositioning have three categories, namely direct-binding partners of existing drugs, cell-based screen approaches and gene expression analysis. However, experimental approaches encountered some great obstacles such as obtaining physical collections of approved drugs[47]. Additionally, experimental approaches for drug repositioning are usually complex, expensive, and time consuming with extremely low success rate.

Computational approaches, on the other hand, have been employed as efficient, fast and economical alternations. Many computational approaches have been published in recent years. Some methods compare the similarity between drugs[51], proteins[52],
or side effect phenotypes[53]. The hypothesis of these methods is based on that drugs with similar chemical structures or side effects are also likely to have similar targets.

Many approaches identified the drug targets from network analysis. For example, in a protein-protein interaction (PPI) network, topologically important proteins, such as hubs and bottlenecks, tend to be essential and may serve as potential drug targets [30, 54]. Florez et al. recently predicted a PPI network for the pathogenic trypanosomatid Leishmania major[55] and identified 142 hub and bottleneck proteins that have no orthologs in humans as potential drug targets. Raman et al. proposed a drug target identification pipeline, namely targetTB, to predict and refine drug targets for this tuberculosis bacterium by combining important proteins/genes from both the interactome and the reactome of Mycobacterium tuberculosis[56]. Potential drug targets can be inferred from similar known drug targets. This requires the knowledge of known drug-target relationships, as well as the measures of drug similarity and target similarity. By applying a two-directional Z-transformation on a score matrix that characterizes the docking strength between chemicals and targets, potential drug targets were identified by Yang et al. for Alzheimer’s disease[57]. Recently, Zhao and Li combined the measures of drug therapeutic (or phenotypic) similarity and drug chemical structural similarity, and predicted links between drugs and target proteins by a regression model called drug CIPHER, based on known drug-target relationships collected from FDA and DrugBank [58, 59].
Computational methods have also been used to analyze the existing experimental data in public databases to explore new target-disease associations. A novel glioblastoma target protein with an approved drug has been identified through network analysis in one recent study[60]. Furthermore, mining methods in literatures with network analysis can search for associations that already exist but have yet to be linked.

The most frequently used resources for computational approaches are data sets of known drug relationships from public databases. Such databases which focus on approved drugs include DrugBank, the Therapeutic Target Database, and Kyoto Encyclopedia of Genes and Genomes (KEGG) Drug and Matador [59, 61, 62]. Overall, computational methods are important supplementary approaches to experimental studies. In this thesis, we mainly focus on using network-based computational approaches for drug repositioning (see chapter 5).
Chapter 3: The Orphan Disease Network Analysis

3.1 Introduction

In the United States, a rare or orphan disease (OD) is any disease fewer than 200,000 inhabitants. There are about 8000 known ODs of which a majority of them are of genetic origin. A large number of ODs typically affect pediatric population, and are often life-threatening or chronically debilitating [63, 64]. About one third of children with ODs die before the age of five. While ODs exist in all classes of disease, based on prevalence, they vary from exceptionally rare to more prevalent. Given the low prevalence rate and large number of ODs, it is practically difficult to develop a specific public health policy for each of the ODs. Thus, using a global approach rather than per-OD approach for the OD analyses and orphan drug research is preferred [65].

Prior to the US Orphan Drug Act (ODA) in 1983, there were only about ten FDA-approved drugs for rare diseases. Following the ODA, there are more than 300 orphan drugs approved [66]. However, most of these orphan drugs are designed to cure rare cancers or metabolic diseases, while there are very few for ODs of other classes. Additionally, the prices of such approved drugs are very expensive, posing a great burden on patients or health insurers [67]. Interestingly, an estimation reveals that nearly 90% of the OD patients have been prescribed at least one off-label drug [68]. However, patients will undertake a safety risk since the ad hoc manner of off-label regulation and usage of drugs constitute experimentations on humans.
without informed consent [69].

Previous research literatures showed that many human diseases tend to inter-related or associated because of their perturbation on same gene. Disease networks, disease gene networks and drug target networks [9, 10, 33] are increasingly adopted as a favorable complements to networks where the edges are protein or gene interactions. However, the quality of these networks is constrained not only by the information for their creation but also by the number of known disease-causing genes [70]. One way to overcome this is using feature based functional linkages other than genes alone. For instance, in Linghu’s study [71], functional linkages were used to explore associations between genes that involved in different disorders. The relationship identification may not be from the same associated genes but sets of genes that are functionally related. It is also reported by a recent study that ODs with a high volume of research literatures have a more likelihood to derive a therapeutic product than those with a low volume of research literatures [72].

Most of the research efforts on ODs focus on either a single OD or a small set of related ODs. To overcome this issue, it is important to conduct a global analysis on a spectrum of ODs with known OD-causing mutant genes (ODMGs). Elucidating the mechanisms and inter-connectivity of the ODs based on their shared genes and functional features can identify and investigate relationships between them and thus are helpful for orphan drug development. It is also important for the understanding of normal biological
pathways and common diseases. In this study, starting with ODs and their known causal genes [65, 72, 73], we built bipartite networks of the human orphan diseasome, namely, the gene networks that are based on protein interactions, functional linkages, shared ODs and literatures, to investigate the ODs and causal genes

### 3.2 Method

#### 3.2.1 Resources of Data and Analysis

We used the UniProt Knowledgebase [73] query interface to extract the ODs and their causal gene information from the Orphanet [65] and the current curated list of all known disorder-gene associations from OMIM databases [72]. The human protein interactome used in our study was integrated from multiple sources [74-79], where both redundant interactions and self-loops were removed (Figure 3.1). We defined essential genes (n=2,481) as previously described[9], a list of human orthologs of mouse genes which resulted in lethal phenotype in embryonic and postnatal stages upon knockout was retrieved (Mouse Genome database [80]). The list of ubiquitously expressed human genes was integrated from the works of Ramskold et al. [81]and Tu et al. [82]. The mitochondrial genes, an inventory of mammalian mitochondrial genes, were downloaded from the MitoCarta database [83]. We used the ToppGene Suite to identify the enriched features including BP, CC, MP, and pathways [84]. We constructed feature-based orphan disease networks (ODNs) with the shared enriched feature as an edge. Similarly, the literature-based ODNs were constructed by using the literatures that cited two ODs as edges in the OMIM records.
3.2.2 Measures of Topological Importance, Connected Components and Modularity

We define the nodes that are in the highest 20% of the degree distribution as hubs (ODs or ODMGs that have the top 20% number of direct connected neighbors), while we also define bottlenecks as the nodes that are in the top 20% in terms of betweenness[30]. Betweenness is a measurement that calculates the total number of the shortest paths in a network that passing through a certain node or an edge. The betweenness, along with the degree, is used to measure the relevance of the location of nodes in a network[5]. The degree and betweenness centrality values are assessed by TopNet-like Yale Network Analyzer (tYNA) [85]. Three well-known centrality measures, namely betweenness centrality, closeness centrality, and eigenvector centrality (available in Gephi package [86]) were adopted to investigate the ODN and orphan disease-causing mutant gene network (ODMGN). In brief, eigenvector centrality measures a node’s importance according to this node’s connections in the network, and closeness centrality is derived as the average distance of one given node to all other nodes in the network. The subnetwork or a connected component in our study is defined as a part of the network of which nodes are only reachable from nodes in the same network. We use Gephi[86] to determine the number of subnetworks and their respective sizes for all the networks generated in this study. The community or modularity, on the other hand, was used to represent the tightness of coupling among a specific group of nodes in comparison to other nodes in the entire network. The clustering algorithm[87] integrated in Gephi was used to generate
the modules in each of the networks in the study.

3.3 Results

3.3.1 Constructing Networks of Orphan Diseases, Mutant Causal Genes and Protein Interactions

Our analysis starts with 1,772 ODs that have at least one implicated gene mutation (2124 OD causal genes or ODMGs). The relationship comprises a bipartite network by connecting genes and ODs if a known mutation in that gene is implicated as a causal mutation for the OD. As a result, there are 3,437 edges (representing genes mutations between OD and ODMG) for 1,772 ODs and 2,124 ODMGs. Among our analyzed 1,772 ODs, there are 1,223 (~69%) ODs have only one known casual gene, while the rest (549) have more than one causal gene, of which 39 ODs have 10 or more known causal genes. On the other hand, 1,393 of the 2,124 ODMGs are implicated in only one OD while the rest 731 genes are causative for 2 or more ODs. For example, gene LMNA is of 17 OD’s causal gene, while the OD nonsyndromic genetic deafness has the highest number of causal genes(43).

In the global bipartite network, the average degree of each OD is 1.94 (number of ODMG per OD), and it is 1.62 (number of ODs per ODMG) for each ODMG. There are a total number of 786 connected components or sub-networks in the network, with the sizes ranging from 734 genes and 530 ODs to just one gene and one OD. Most
(602 out of 786, or ~77%) of these sub-networks are with only one OD with one gene. There are 1,254 communities(tightly connected subnetworks) or modules are detected by Louvain’s modularity algorithm[87] (Louvain modularity=0.81).

From this OD to gene bi-partite network, two types of networks were built and analyzed. Namely they are: 1) Orphan Disease Network (ODN), 2) Orphan Disease Mutant Gene Network (ODMGN), When generating ODN, two ODs are connected if they share one or more mutant genes, and it is similar for ODMGN as a reverse.

In the next stage, we further selected a subset of all ODs with 4 or more causal genes and connected them based on shared enriched features (e.g., biological processes, cellular components, pathways, or mammalian phenotypes) to perform a functional enrichment analysis(Figure 3.1). We connect two ODs based on their shared feature even if they do not share a gene. Finally, with the literature records that citing the ODs, we also built a literature-based OD network to investigate and compare it with the traditional gene-based OD networks. Figure 3.2 shows the gene-based OD network (ODN) which contains 1,170 nodes and 2,259 edges.
Figure 3.1 Workflow of Generating the ODNs based on the Shared Genes or Enriched Functional Features (A) The details of the ODN, ODMGN, and ODMGI that are generated with the orphan disease and OD-causing mutant gene bipartite network and the human protein interactome. (B) Outlines of the method and results of the functionally connected ODs

In this network, each node represents an OD, and each edge represents the shared ODMGs between the OD pair. There are 184 connected communities extracted from the ODN. The largest connected community (or sub-network) has 530 nodes and 1,396 edges. The OD-based mutant gene network (ODMGN) on the other hand contains 1,521 nodes and 6,855 edges (Figure 3.3). In contrast to ODN, each node in ODMGN represents a gene, while an edge between a gene pair represents at least one OD shared by them. In the case of the ODMGN, there are 183 connected communities extracted and the largest connected community contains 734 nodes and 4817 edges.
To investigate the significance of connectivity in the ODN and ODMGN, we performed a random shuffle on the connections between ODs and ODMGs in the bipartite graph, with the number of links for OD or ODMG remain unchanged. From the randomized ODN and ODMGN, the average sizes of the largest connected community are 954±18 and 1,305±19, which are dramatically larger than those of actual ODN (530) and ODMGN (734) with p-values <1.0e-05 (One-sided Student’s t-test), respectively. The results indicate that the decomposition of ODs and ODMGs deviates significantly from a random distribution. This is consistent with a previous study on the common disease network[9], which attributed the important pathophysiological relatedness with clustering algorithms between different diseases and disease genes.
Figure 3.2. Network of OD with Edges are Shared Genes (A) The loosely connected 184 components (subnetworks) of the ODN. (B) There are 76 modules within one of the largest subnetworks of the ODN. The coupling tightness of among a specific module of nodes compared to others in the entire network is indicated by modularity.

Figure 3.3 A global network view of ODMGs, where each node is an ODMG and each edge denotes at least one shared OD. The network contains 1,521 nodes and 6,855 edges. Different colors represent different connected components

3.3.2 Orphan Disease-causing Mutant Gene Interactome (ODMGI)

Previous studies show that genes do not tend to be essential and to encode hub
proteins[9]. In order to check whether ODMGs are similar or different from common disease-causing mutant genes, we next construct an OD-causing mutant gene interactome network with an assembled human protein interactome. The human protein interactome adopted in our study includes protein to protein interactions (PPI) from large-scale yeast two-hybrid experiments[74, 75], computational predictions[76] and literature curation [77-79]. The redundant interactions and self-loops were removed. As a result, there are 12,260 proteins and 70,576 interactions in the assembled PPI network. Of the 2,124 ODMGs, 1,811 encode proteins are part of human PPI network. And 1,488 of them interact with another protein encoded by an ODMG. There are 559 overlapped interactions between the ODMGN and the orphan disease-causing mutant gene interactome (ODMGI). Additionally, the network has 3,662 interactions with 1,488 proteins of ODMGs. The number is much more than the expected number of 1,539 interactions. The expected number is calculated as follows: We divide the number of all PPIs in the PPI network (70,576) by the number of all possible PPIs between all protein pairs (75,147,670), and then multiply it by all possible PPIs between ODMG pairs (1,638,955). The 559 protein interactions (representing 590 ODMGs for 266 ODs) not only interact physically but also share an OD. We organized them as 145 connected clusters of size 3 or larger (at least two interacting ODMGs and an OD) of proteins associated with the same or a related disorder. Our findings and conclusions drawn from the ODMGI analyses are demonstrated in the following three sections:
**OD-Causing Mutant Genes Tend to Have High Degree or Serve as Bridges between ODMG Modules**

We noticed that proteins encoded by ODMGs in the human PPI network tend to have degrees which are higher than-average, and higher betweenness centrality as well when compared to all other proteins in the network. On one hand, the percentage of ODMGs that are hubs in the PPI network is about 28% (507 out of 1,811). This is higher than the 20% cutoff definition for all hubs. On the other hand, the average degree is 15.40 for proteins encoded by 1,811 ODMGs in the PPI network, it is also significantly higher than that of other proteins with an average degree value of 10.84 in the network. (p value < 1.0 e-5; Wilcoxon rank sum test). Similarly, both the bottleneck percentage and the average betweenness values for ODMGs are higher than those of all proteins in the network. Previous studies[9] (based on all diseases in the OMIM database) reported that there is only a weak correlation between hubs and disease genes, and thus a majority of disease genes do not tend to be essential and to encode hub proteins, our discoveries, however, is in contrast to this.

It is naturally comes to the question that if the opposite is true. In other words, whether highly connected ODMGs in the human PPI network are responsible for multiple ODs. We found that ODMGs encoding protein hubs (or bottlenecks) in the PPI network tend to be implicated in more ODs than non-hubs (or non-bottlenecks). Of all the 1,811 ODMGs, the average number is 1.65 for their implicated ODs (the OD degree). The differences between the average OD degree of hubs (1.85) and
non-hubs (1.58) are also significant (p value= 0.0167, one-sided Wilcoxon rank sum test). Similar results were discovered through bottlenecks and non-bottlenecks’ comparisons (1.87 versus 1.56; p value=4.32e-4).

**Protein Products of ODMGs Tend to Be More Physically Interacting with Those of Other ODMGs**

The average number of interacting partners that are also known to be OD-causing mutant genes for all 1811 ODMGs in the PPI network is 4.04, and the average ratio between the ODMG-interacting degree and the PPI degree is 0.358. The number is much higher than the expected ratio of 0.148 (1,810 out of 12,259) (p < 1.0e-5; one-sample t test), suggesting that protein products of ODMGs tend to physically interact with other protein products of ODMGs. Although only PPIs might not be able to detect every novel OD protein, it is still promising with the relatively high fraction of other OD proteins localized within the immediate ODMG-protein interactome space. Indeed, it has been shown that the systematic use of PPI data can improve positional candidate gene prediction by 10-fold in previous studies[38].

**Hubs in the ODMGN Are Not Likely to Be Hubs/Bottlenecks in the Human PPI Network**

To address the question of whether an ODMG encoded protein is a hub both in the OD and PPI network, we next compared the ODMGN (edges represent shared ODs) with ODMGI (edges are protein to protein interactions). There are 1,302 ODMGs
from the 1,521 genes in the ODMGN have known protein interactions. Among these 1,302 ODMGs, on one hand, 375 are hubs (from the top 20% of nodes with the highest degree values) in the human interactome, whereas the remaining 927 nodes are non-hubs. On the other hand, 388 are bottlenecks (from the top 20% of nodes with highest betweenness centrality values), whereas the remaining 914 are non-bottlenecks. All the degree and betweenness centrality values were calculated by tYNA [85]. We found that hubs in the ODMGN do not tend to be hubs or bottlenecks in the human PPI network (or ODMGI). The average degree of ODMGs that are hubs (8.28) in the PPI network is not significantly different from the degree that are non-hubs (9.00) (p=0.404; one-sided Wilcoxon rank sum test). Moreover, the average ODMG degree that are bottlenecks in the PPI network (8.48) is not significantly different from that of non-bottlenecks (8.92) (p=0.544; one-sided Wilcoxon rank sum test). There were 220 ODMGs for which the encoded proteins do not have any known protein interactions. However, the average degree (10.34) of these ODMGs in the ODMGN is higher compared to ODMGs in the PPI network (8.79), although not statistically significant (p=0.173). The result indicates the importance of the ODMG hubs in ODMGN is irrespective to their status in the PPI network. However, it is still need to note that many conclusions about global measures (e.g., network topology) should be viewed with some skepticism due to the knowledge of the interactome remains incomplete[88].
3.3.3 Orphan Causing Mutant Genes Encode Proteins Show Tendency to Be Essential

According to our findings that OD genes tend to encode hub or bottleneck proteins in the PPI network, we assume most of them could be essential genes. To verify this, we performed a direct comparison with essential genes as described earlier[9]. The results showed that a percentage of 36% (765/2,124) of the ODMGs are essential genes whose ortholog gene knockout in mice is lethal, indicating a significant higher rate than the 22% (398/1,777) of essential genes in the disease network reported by Goh et al [9]. In addition, we have also observed that about 18% of the ODMGs (376) can cause premature deaths in mouse ortholog gene knockout models. Thus, combined them all, there are 907 genes (~43%) from the 2,124 ODMGs result in either premature death and/or lethality in mouse gene knockout models. Since in Goh et al.’s work[9], it comprised several ODs, we believe that it will be even more significant and specific to ODs. The reported 22% is probably due to the presence of some of the ODs and genes in their dataset.

To test whether the hypothesis is indeed true, we divided all ODMGs from the entire set of OMIM disease genes[72]. Thus two classes of disease genes have been generated: 2,124 ODMGs and 1,901 non-ODMGs (NODMG) or common disease genes. Although ODMGs, as defined earlier, are genes that when mutated caused an orphan disease, NODMGs are genes whose mutant forms are not associated with any orphan disease (based on current orphan disease and gene relationships in the
Orphanet database). Compared to NODMGs, ODMGs are significantly enriched for lethality, mitochondrion, and premature death as well (p < 1.0e-5; Fisher’s exact test) (Figure 3.4). A total of 765 (~36% of 2,124) of ODMGs are essential, while on the other side, only 10% (192/1901) of NODMGs are essential. When we extend to check which essential genes overlapped with the entire set of disease genes from OMIM Morbid Map (as in Goh et al.[9] but with updated disease and essential gene lists), there were 920 (24%) essential disease genes. It is interesting that the result is similar to the 22% reported by Goh et al. This confirms the original findings from Goh et al., s’ study, which was based on all disease genes, still stand good even there is increase in the database sizes of human disease genes (from 1776 to 3864) and the essential genes (from 1267 to 2481). It also strengthens our conclusion that the enrichment of essential genes is something specific to ODMGs because the essential ODMGs have higher percentage when compared to either NODGMs or all disease genes from OMIM. Moreover, these results suggest the robustness of our conclusions as well as previous conclusions, and we do not expect them to change significantly even if the resource databases are updated with additional genes and annotations.
Figure 3.4 A comparison of relationships between ODMG and NODMG with different categories of genes showed by Venn Diagrams (A) The overlap between OMIM disease genes and ODMG. (B) The overlap of ODMG and NODMG to essential genes, while C and D indicate the overlaps with mitochondrial genes and genes whose knockout in mouse causes premature death. The table in (E) shows that, compared to NODMGs, ODMGs have more essentiality, mitochondrial genes, and genes associated with premature death in the mouse knockouts.

3.3.4 Orphan Disease Networks based on Shared Functional Terms

In addition to ODMGI and ODMG networks, we also derived the functional term based orphan disease networks. In order to obtain a statistically significant and representative functional signature from the 1,772 ODs, we first extracted all those ODs with four or more mutant genes from our original data set. Starting with this filtered sub-bipartite network of 196 ODs and 1,087 genes (1,283 total nodes and 1,395 total edges), we generated OD-OD networks with the shared genes and shared
functional terms. The enriched functional terms ($p < 0.05$) for each of the 196 ODs were determined with the ToppFun web application [84]. The shared functional terms considered for the enrichment analysis included biological processes (BP) and cellular components (CC) from Gene Ontology, KEGG pathways, and mammalian phenotype (MP). We rebuilt the orphan disease networks using the enriched features for each of the orphan diseases. Unlike previous approaches, this time the edge between one pair of ODs represents an enriched shared functional term (BP, CC, Pathway, or MP) and not necessarily a shared gene. After generating these function-based OD networks, we compared them with the gene-based orphan disease networks to find the overlapping nodes and edges. Surprisingly, the gene-based OD network (153 OD nodes and 191 edges where an edge represents a shared ODMG) is largely different from various functional term-based OD networks, including a BP-based OD network (176 OD nodes and 2,244 edges; edges are shared BP terms), a CC-based OD network (153 OD nodes and 1,135 edges where edges are shared CC terms), an MP based OD network (155 OD nodes and 745 edges where edges are shared MP terms), and a pathway-based OD network (159 OD nodes and 511 edges where edges are shared pathways). Although the node agreement between the gene-based ODN and functional term-based ODNs was higher and corresponding Jaccard indices ranged from 0.647 to 0.732, the edge agreement was much lower, and Jaccard indices ranged from 0.0592 to 0.162 ($p < 1.0e-5$ compared with $p$ for random expectations, one sample t test; we assessed random expectations by calculating the overlap between the gene-based network and randomized function-based networks with shuffled edges and
unchanged node degrees). To address the effect of data incompleteness, we added up to 20% random edges into the gene-based and term-based networks to approximate uncovered associations and compared the overlap of edges with what would be expected as a result of chance, and the results are consistent.

### 3.3.5 Orphan Disease Networks with Shared Literature

To investigate the effectiveness of literature-based networks other than traditional gene-based approaches in identifying OD-OD relationships, we re-constructed the orphan disease network where the edges are shared publications instead of a shared gene. Mining the relationships in literatures may cause potential false positives. To avoid this, we used the corresponding OMIM records of ODs that summarize results of publications on gene-disease relationships. Specifically, we used the cited literature (the links to PubMed records for the references cited in an OMIM entry) in the OMIM records. There is a corresponding OMIM record for 1,461 ODs (derived from Orphanet). Of the 1,475 mapped OMIM records, 1,370 can be found at least one cited article (indicated by presence of at least one PubMed ID). We used this subset of 1,370 ODs to compare the gene-based OD network with the literature-based OD network. The gene-based OD network contained 811 nodes (ODs) and 1,277 edges, indicating common ODMGs shared by a pair of ODs. The literature-based OD network contained 747 OD nodes and 927 edges, representing shared literature (PubMed IDs) for a pair of ODs.
To evaluate the significance of connectivity, we randomly shuffled the connections in the bipartite graph between ODs and PubMed IDs, with the number of connections per OD or PubMed ID unchanged. The average size of the largest module in the randomized ODN is 823±12, which is dramatically larger than the one in the PubMed-based ODN (432 nodes, with a p < 1.0 e-5, one sample Student’s t test). The result suggests pathophysiological modularity of ODs differs from a random distribution which is similar in the case of ODN and ODMGNs.

Although a large number of the common nodes are found between the gene-based and literature-based networks (517 ODs, 0.5 by Jaccard index), there are fewer overlapping edges (255 overlapping edges, 0.13 by Jaccard index; p-value < 1.0e-5 compared with random expectations, one-sample t-test; random expectations were derived by calculating the overlap between the gene-based network and randomized PubMed-based networks with shuffled edges with node degree unchanged). In addition, among the 517 common ODs, less than 25% of the hubs (31 out of 166 and 143 hubs in the gene-based and the PubMed-based networks, respectively) are conserved. To tackle the effect of data incompleteness, we randomly introduced 20% edges into the gene-based and PubMed-based networks to cover the possible connections and compared the overlap of edges with what would have been expected, and the results shows a consistence. These results demonstrated that the wirings of these two networks are largely different, suggesting that many ODs might still be related even there is no shared mutant genes between them. It was also observed that
the measures of topological importance have a significant difference between the two networks with little overlap. For instance, comparing the top 100 ranked OD nodes (ranks are based on three centrality measures—betweenness centrality, closeness centrality, and eigenvector centrality) in these two networks, it shows very little overlap. Moreover, the literature-based OD network can discover additional relationships for those diseases with no shared known disease genes but having potential functional links between their corresponding disease gene sets. Among the 927 potentially related OD pairs with literature support, 255 (~28%) pairs also have shared known disease genes and are identified by both methods. However, a large number (672 edges; ~72%) have no shared known disease genes, and their associations are identified solely on the basis of literature-connectivity (Figure 3.5).

**Literature-based ODs**

![Orphan Disease Network Based on Shared Literature](image)

**Figure 3.5** Orphan Disease Network Based on Shared Literature (A) Each node represent an
OD (577 ODs in all) and each edge (672 edges in all) represent a shared published article. Although there are no shared OD-causing mutant genes between these diseases, they are still connected by a series of shared publications. (B) Some clusters of the literature-connected orphan diseases.

3.4 Discussion and Conclusion

Although recent research progress provides opportunities to accelerate the understanding of the basis for more orphan diseases, and brought the development of innovative medical approaches, there is still rare efforts have successfully addressed scientific questions across a spectrum of orphan diseases. Therefore, if we intend to make more than baby steps in orphan disease research, it is critical to find common genes, targets and pathways between them. Building networks that with orphan diseases and their underlie biological processes and pathways can boost the identification of the functional units that respond to genetic perturbations and potentially affect disease risk or therapeutic response, moving the field to a favorable direction systematically. We believe that the relationship between orphan diseases and their genetic mechanisms can be better understood by the decomposition of orphan disease networks. Studies on biological networks can reveal common pathways or processes for multiple orphan diseases which are biologically related. A comprehensive understanding on such molecular basis could raise opportunities for interventions which can benefit an array of related orphan diseases. This capability could lead a way to discover of single therapies which can benefit multiple disorders and also for more possible common diseases.
It was observed that there is a weak correlation between hubs in PPI networks and disease genes from previous studies which focusing on all human diseases. The conclusion was based on the discovery that most of disease genes are nonessential with no tendency to encode hub proteins[9]. However, our study shows a contrast result: genes whose mutations cause orphan diseases tend to encode hubs in the PPI network and also tend to be essential. We consider it is reasonable because: First, the rarity of orphan diseases in a population can be explained by an evolutionary argument, the mutations in hubs affect many proteins and may not be compatible with survival and hence are less likely to be existing in a population. Second, most of the ODs have severity and lethality, and it is possibly because of the mutations in hubs that may have wider repercussions compared to those in non-hubs. Third, protein hubs are the possible cause of the complex phenotypic nature of ODs impacting multiple physiological systems, with their involvement in heterogeneous cellular processes through their multiple interacting proteins.

Biological networks are found to be modular, and their modularization provides deep insight into living systems and human diseases[6]. The discovery of high connectivity among different ODs or ODMGNs can be used not only to infer the common mechanism and targeted pathways, but also help us to find possible drug repositioning or drug repurposing candidates (i.e., to find novel indications for already approved drugs), especially when one or more orphan diseases in the same module has an approved drug.
As most of the previous studies are gene centric where the elucidating of disease relationships are based on disease genes, the identified disease relationships are restricted to known causing genes[89]. To tackle this, three recent studies[71, 89, 90]suggested functional linkage maps where diseases were connected by specific functional annotations. However, each of these three approaches used an incomprehensive set of annotations such as gene expression, biological processes, PPIs, or pathways. Our results on the functional term based networks suggest that it is largely different from the wiring of the gene-based OD network to the wiring of functional term based OD networks, and the associations between the ODs cannot be fully captured by the gene-based OD networks alone. Hence, by taking functional connectivity between causative genes that involved in different orphan diseases into consideration, it can assist us to reveal relations between orphan diseases. Such relations can be helpful to generate novel hypotheses on the molecular mechanisms of diseases, and provide guidance on the development of relevant therapy[71].

A literature-based discovery method[91] has demonstrated its effectiveness in disease gene identification. The relationship between orphan diseases based on literature co-citations can provide immediate helpful insight on generating novel hypotheses in therapeutic strategies. In this study, we identified ~670 OD to OD relationships that can only be connected by shared publications but not by shared genes. We used only cited literature in OD records in order to limit the number of false positives hits
brought by literature-mining. However, potential limitations may still be with this approach. For example, other than relating some ODs’ mechanistically or functionally, literature may just list some of the ODs in a context.

Besides leading to new insights of the biological underpinnings on a spectrum of ODs, we believe that the global network analysis on orphan diseases will facilitate the new and innovative research development on these rare conditions which have been hitherto understudied. The global analysis on various ODs can assist in analyzing comorbidities and the underlying molecular basis, as well as establishing potential networking approaches. The functional feature-based OD networks not only partially address the limitations of the gene-based connectivity networks of diseases, but also have direct implications to drug discovery process. Physical protein-protein interaction based ODMGI networks can be used to construct lists of genes potentially enriched for new candidate ODMGs. We have also used several sources of biological data to construct functional term-based networks of ODs that have advantages over gene-based disease networks. These functional term based networks of ODs can provide a common framework for integrating various data types into a common predictive network. Moreover, the shared functional terms between ODs can be analyzed to predict specific OD genetic modifiers or drug targets. Previous studies have already successfully used the integration of various interactome and functional relationship networks to predict cancer and other types of susceptible disease candidate genes[92, 93]. ODN provides an important tool in orphan drug discovery as
it represents a genome-wide roadmap for future studies on orphan dis easome and druggome. As such, it can be used to assess interactions between ODs and the ODMGs through the orphan dis easome web site that we made available online. It offers global view and a quick visual reference of the genetic connections between orphan diseases and mutant genes. For instance, if we overlay the networks of ODs and/or genes with orphan drugs or common disease drugs, it can be used as a platform for discovering potential drug repositioning candidates.
Chapter 4: Prioritizing and Discovering Disease Casual Genes via a Vertex Similarity Based Network Approach

4.1 Introduction

Although recent biotechnological advances such as next-generation sequencing technologies have accelerated the disease causal gene discovery pipeline, the prioritization for a large list of candidate genes is still an important step for disease-gene discovery [94]. We [2], and other earlier studies [9, 41, 42, 69], have demonstrated that genes involved in phenotypically close diseases tend to share molecular signatures with similar expressions, and the participation in the same pathways or biological processes, complexes or protein interactions, literature co-citations. In chapter 3, we completed a global analysis of all ODs that are with at least one known associated mutant gene (data source from Orphanet[95] and OMIM [72]), our analysis results show that the relationship between ODs cannot be fully captured by the gene-based network alone. The integration of diverse types of data on biomedical and genomic domains can facilitate hypotheses synthesis on disease causing mutant genes. In addition, it can help us in addressing an important question: are there any candidate genes related to known causal genes for a disease? An effective method to tackle this question is to prioritize the candidate genes in a “guilt by association” way, where the prioritization of test set is based on their similarity to training or ‘seed’ set. Such ranking approach has become an important way to rank
candidate disease causal genes that are found in genome-wide association or linkage studies [96]. The genes shown to be linked to a specific disease, for example, can be prioritized based on their similarities to a reference set of known genes for that disease. Several computational approaches that perform this task automatically have been developed by our and others’ studies [84, 94, 97-106].

Network-based analyses have been increasingly used not only in identification of novel disease candidate genes but also in prioritization of disease candidate genes [35, 40-45, 107-109] especially when the genes are relatively less annotated. Generally, the network-based candidate gene ranking approaches can be categorized into two groups: parameter-based and parameter-free methods. The parameter-based methods, such as PageRank with Priors (PRP [43]), Random Walk (RW [40]) and PRIoritizatioN and Complex Elucidation (PRINCE [41]), as the names annotated, they usually require auxiliary parameters that need to be trained during the prioritizing process. The PRP for instance needs a parameter $\beta$ to determine the probability of returning to the initial node [43]. Similarly, the PRINCE algorithm takes a parameter to describe the relative importance of prior information [41]. Since selecting the optimal parameters could be a challenge for users, the parameter-based approaches are considered more user-unfriendly compared to parameter-free ones[109]. Additionally, most parameter-based approaches require the global network information of the entire network and thus they often need to perform extensive computation. For example, in PRP, all the nodes’ scores in the network are updated.
iteratively, the computation stops until convergence of the scores. However, when the
network size is large, the computation process typically becomes extremely slow and
inefficient. The parameter-free methods (e.g. Interconnectedness or ICN [109]), on the
other hand, measure closeness of each candidate gene to known disease genes by
taking into account local network information such as direct link and the shared
neighbors between two genes, and thus the computations tend to be relatively less
intensive. The ranking performance of parameter-free methods however is usually not
comparable to parameter-based ranking approaches. To address this, we developed a
novel network-based parameter-free framework for discovering and prioritizing
candidate rare disease genes. In our study, we specifically focus on two aspects: a)
improve the prioritizing performance compared to current parameter-free methods and
b) achieve a comparable performance to the parameter-based ones while the
computation load is less. We tested our approach in a leave-one-out cross-validation
settings on prioritizing genes for 172 ODs that have at least five known causal genes
(from Orphanet database [95]). We also compare the performance of our method with
other two approaches, one is from parameter-based and the other is from
parameter-free methods. To demonstrate the utility of our approach, we further
collected the immediate neighbor set of known OD genes as test set, and ranked them
as potential novel candidate genes. The immediate neighboring gene sets were
compiled using (a) protein interactions; (b) functional linkage network [110, 111]; and
(c) literature co-citations.
4.2 Methods

4.2.1 The Vertex Similarity Ranking Algorithm

Our approach is based on the hypothesis of “guilt-by-association principle”. We consider genes that have associations to one or more known disease genes (“seed genes”) tend to have higher probability to implicate in the same disease. Hence our aim is to discover such novel candidate genes that have “strong” associations to the seed genes. In our approach, we define that two nodes or vertices in the biological networks are considered similar if their immediate neighbors are themselves similar (common genes, pathways, biological process, etc.). According to this principle, we can build a self-consistent matrix formulation of functional similarity that can be evaluated iteratively using only knowledge of the adjacency matrix of the network (based on functional annotations of genes). To this effect, we use similarity between vertices (genes) to measure their strength of association in a network. Thus, two vertices are likely to be strongly related if they have a high similarity. In order to measure the similarities between the seed and the candidate genes from test set, we proposed a vertex similarity measurement in our algorithm. Similarity measurements, such as cosine similarity, have been successfully applied for comparing the similarity between documents that are described as vectors of keywords [112]. However, it is to the best of our knowledge that there have been no reports of using it as a measure to compute similarity between two genes in a biological network and use it for candidate disease gene prioritization.
In our approach, if two genes are directly connected, then the connection vectors of these two genes can be constructed based on the topology of them and their directly associated genes (Figure 4.1). When two genes are not directly connected, a shortest path between them is considered for this case. As demonstrated in Figure 1, the similarity score $\text{Sim}(A,B)$ between gene A and B is defined as:

$$
\text{Sim}(A, B) = \frac{\sum_{i=1}^{n} \omega_{A,i} \times \omega_{B,i}}{\sqrt{\left(\sum_{i=1}^{n} (\omega_{A,i})^2\right) \times \left(\sum_{i=1}^{n} (\omega_{B,i})^2\right)}}
$$

(1)

where $\omega_{A,i}$ indicates the edge weight of node A to node $i$, and $\omega_{A,i} = \omega_{B,i} = 1$ (protein interactions) is defined here; $n$ is the number of the nodes that involved in A and B’s connection, which includes A, B and all the neighboring nodes of A and B. Equation (1) calculates A and B’s similarity based on a cosine similarity manner, where the numerator is the dot product of vector A and vector B, and the denominator is the cross product of vector A and vector B. Since we define Equation (1) applies only when nodes A and B are directly connected, the value of $n$ can be calculated by

$$
n = \Gamma_{A} + \Gamma_{B} - \sigma_{\text{shared}}
$$

(2)

where $\Gamma_{A}$ and $\Gamma_{B}$ are the degree (number of edges a node has to connect other nodes) of nodes A and B respectively, and $\sigma_{\text{shared}} = |\Gamma_{A} \cap \Gamma_{B}|$ and represents how many neighbor nodes shared by both A and B.

When node A and node B are not adjacent (there is no direct connection between each
other), we try to find a shortest path between them. In this case, the similarity score between A and B will be calculated by

$$Sim^*(A, B) = \begin{cases} \prod_{k=1}^{K} Sim(C_k, C_{k+1}) & \text{if } K \leq r \\ 0 & \text{otherwise} \end{cases}$$

(3)

where $C_k$ represents the nodes on the shortest path of A and B, and $r$ is the control range that limits the maximum hops between each other (maximum $r$ hops). In other words if the length of the shortest path between A and B is more than $r$ hops or if there is no shortest path between them, $Sim(A, B)$ will be 0.

![Figure 4.1 Illustration of the connections between hypothetical genes A and B. Each node represents a gene and each edge represents either a physical interaction or functional association. $\omega$ is the weight of each connection which in case of protein interactions is 1.](image)

The prioritization of the candidate genes in the test set are based on the similarity scores derived from equation (1) and equation (3). If a disease $d$ has more than one known casual gene, all of its casual genes will be seed or training gene set $S_j$. And each candidate gene’s final score to this disease is the summation of the similarity
scores between the candidate gene and each of the seed genes from the training or seed set $S_d$. The final score of a candidate gene $i$ is calculated as:

$$score_i = \sum_{j \in S_d} Sim(i, j)$$

(4)

where $Sim(i,j)$ is the VS score between gene $i$ and $j$. All candidate genes are then sorted based on these final scores (Figure 4.2).

Algorithm 4.1 has summarized the VS algorithms for disease candidate gene ranking.

Algorithm 4.1 Vertex Similarity algorithm for disease candidate gene rankings

1: INPUT: A global network $N$, a list of seed nodes $S[0,m-1]$, a list of training nodes $T[0,n-1]
2: OUTPUT: A sorted list of candidate nodes $T_s[0,n-1]
3: for 0 \leq i \leq n-1 do
4: construct $T[i]$’s neighboring connection vector $V_i$
5: for 0 \leq j \leq m-1 do
6: construct $S[j]$’s neighboring connection vector $V_j$
7: If ($T[i]$ and $S[j]$ direct connected) then
8: $Sim(T[i], S[j]) = (V_i \cdot V_j) / (V_i \times V_j)$
9: else

Figure 4.2 Calculate all similarity scores for one candidate gene to all the disease causal genes for an given disease
Find the node set \( G[0, k-1] \) on shortest path of \( T[i] \) and \( S[j] \)

\[
\text{for } 0 \leq l \leq k-1 \text{ do }
\]

\[
\text{Sim}(G[l], G[l+1]) = (V_l \cdot V_{l+1}) / (V_l \times V_{l+1})
\]

\[
\text{Sim}(T[i], S[j]) = \text{Sim}(T[i], S[j]) \times \text{Sim}(G[l], G[l+1])
\]

\[
\text{end for}
\]

\[
\text{end if}
\]

\[
\text{Sim}(T[i]) = \text{Sim}(T[i]) + \text{Sim}(T[i], S[j])
\]

\[
\text{end for}
\]

Sort \( T[0, n-1] \) based on \( \text{Sim}(T[i]) \)

Return \( T_s[0,n-1] \)

---

### 4.2.2 Prioritization Performance Evaluation Methods

#### 4.2.2.1 Data Resource

The ODs and their causal gene information were downloaded from Orphanet [95]. We merged some of the OD subtypes of a single disease based on their given disorder names as described previously [2, 9]. From this, we selected 172 ODs that each has at least five known causal genes. The total number of genes across 172 selected diseases was 1,598, with 1,312 genes exist in the human protein interaction network. The human protein interactome data used in this study was integrated from several resources [74-79] with both redundant interactions and self-loops removed.

#### 4.2.2.2 Comparing Ranking Performance Comparison with Other Algorithms

In order to compare the ranking performance of our VS-based approach to other approaches in candidate disease gene ranking, two methods were selected for the comparison, namely the PageRank with priors (PRP) [43] and Interconnectedness
(ICN) [109], with one each from parameter-based and parameter-free methods. The framework was partially implemented using JUNG (Java Universal Network/Graph; jung.sourceforge.net) [113] as described earlier [43]. To investigate the VS-based approach performance and compare with two other methods, we employed a leave-one-out cross-validation procedure. In each cross-validation trial for an OD, we removed a single OD causal gene (“target gene”) from its known disease casual genes (training set), and added 99 random genes to this gene to form the test set. We then performed each of the 3 algorithms to derive their rank assignment to the “target gene” in the test set. The procedure will be repeated for all the genes in training set for an OD.

PageRank with Priors (PRP)

PageRank with Priors is an extended version of PageRank proposed in White and Smyth’s work [114]. It simulates a random walk in the network that starts from the root nodes with a jump back probability. The stationary probability of the walk on a node is calculated by an iterative equation, as shown in Equation (5).

\[
\pi(v)^{(i+1)} = (1 - \beta) \left( \sum_{u=1}^{d_{in}(v)} p(v | u) \pi^{(i)}(u) \right) + \beta p_v \tag{5}
\]

In this equation, \( p_v \) represents the “prior bias”. Initially for all the nodes in root set, \( p_v \) is set to \( 1/|R| \), while \( p_v=0 \) otherwise. In addition to \( p_v \), a “back probability” \( \beta \) is also defined where \( 0 \leq \beta \leq 1 \). The value of \( \beta \) determines the probability that it jumps back to the set of root nodes in R. For other parameters, \( d_{in}(v) \) is the in-degree of node \( v \) and \( p(v | u) \) is the probability of jumping from node \( u \) to node \( v \).
Equation (5) can be interpreted as a Markov chain process that a surfer randomly jumps “back” to the root set \( R \) with probability \( \beta \) at each time-step. It is similar in spirit to the use of weighted paths as follows: the method evaluates the probability of landing on a node in the modified Markov chain where a random graph walker starts in the set \( R \) (with appropriate prior probabilities) and executes a random walk that ends stochastically with probability \( \beta \) (at which point the process restarts). This process defines an (infinite) set of walks of variable length starting at the root set (in fact they will follow a geometric distribution with mean \( 1/\beta \)). The “rank” equation above estimates the relative probability of landing on any particular node during this set of walks. The computational complexity is the same as that of the standard PageRank algorithm.

**Interconnectedness (ICN)**

Interconnectedness (ICN) is a parameter-free method that proposed by Hsu et al.[109]. In this approach, candidate genes are ranked by deriving their closeness to known disease causal genes from a network. The closeness between genes in a network was calculated by considering two genes’ direct and indirect interaction connectors. Briefly, ICN determines that these genes are more likely to be involved in a same functional module if they have more shared interacting genes.

We performed a leave-one-out cross-validation using the 172 ODs and 1,312 OD causing genes that exist in PPI network. We used the human protein interaction
network as the global network to evaluate the prioritizing performance of VS and other two methods. The human protein interactome used in our study contains protein-protein interactions from large-scale yeast two-hybrid experiments [74, 75], computational predictions [76], and curation of the literature [77-79], with both redundant interactions and self-loops removed. The assembled PPI network consists of 11,765 proteins and 69,167 interactions. During each set of a validation trial, one seed gene (“target gene”) from one of the selected 172 ODs was picked out and mixed with 99 random genes from PPI network to form a test set of 100 candidate genes. The remaining seed genes of an OD were used as the training set. The test set genes were then prioritized using the three approaches: PRP (with back probabilities 0.3 and 0.05), ICN, and VS-based approach. During each run, the rank of the “target gene” was noted. We evaluated the performance of each algorithm in terms of the success rate versus rank cut-off (k). If the “target gene” is ranked among the top k in a particular validation run, it is considered as a ‘success’. The validation runs are repeated until all the seed genes have been used as the target gene and their ranks are obtained. The “success rate” is defined as the ratio of successful validation runs and the total validation runs for all the existing OD genes from 172 ODs. The same strategy was followed for all the three algorithms. In case of PRP which is a parameter-based method, we selected a back probability of 0.3 since we have shown previously that the performance of PRP in ranking candidate disease genes was best at \( p=0.3 \) [43].
4.2.2.3 Collecting OD Candidate Genes as Test Set for Identifying and Ranking

For collecting the novel OD candidate genes as test set, we used the neighbor genes that are directly connected to known OD genes as the test set. The candidate genes were compiled by immediate neighbors of 172 OD causal genes from three networks, namely, (a) protein interactions; (b) functional linkage network [110, 111]; and (c) literature co-citations. The protein interactome data as described earlier was compiled from several resources [74-79]. The functional linkage network-based candidate gene sets were derived from two resources: (i) HumanNet, a probabilistic functional gene network of Homo sapiens [111] and (ii) functional protein interaction network built upon expert-curated pathways [110]. For genes in test set that are based on literature co-citations, we compiled them using the OMIM database. Briefly, for the selected 172 ODs, we extracted their corresponding OMIM records, which summarize results the gene to disease relationships in publications. For the OD mapped OMIM mapped records, we first derived the cited literature (PubMed records for the references cited in an OMIM entry) in the OMIM records. With this OD-related PubMed records, we extracted the involved genes from the ‘gene2pubmed’ file from NCBI [115]. For each of the 172 OD2 with known causal genes, we gathered all neighboring genes (immediate neighbors) of causal genes from different networks and used them as a test set for the prioritization by VS-based approach in protein interactome network (Figure 4.3).
4.3 Results

4.3.1 Leave One Out Cross Validation Results

For the leave one out cross validation, 172 ODs that have 5 or more known casual genes (1,598 in total) were selected. Of the 1,598 genes, there were 1,312 existing in the protein interactome network. Figure 4.4 presented the results from the leave one out cross validation of the three approaches. As can be seen from the figure, when the rank cut-off k = 1, both VS-based (parameter-free) and PRP (with back probability set to 0.3; parameter-based) methods, performed equally the best with a success rate of \(~43\%\). In other words, in VS-based method, the selected target gene was ranked top1 in 568 out of 1,312 cases (43.3\%). On the other hand, the target gene was ranked top1 559 times out of a total 1,312 cases (42.6\%) using PRP (with back probability set to
Meanwhile, another parameter-free method ICN, achieved a lower success rate at 35.3% with the target gene ranked top1 by 463(out of 1,312) times. PRP (with another parameter setting, where the back probability was set to 0.05) achieved a lower success rate than VS at 37.2% (488/1312) and was better than ICN.

![Figure 4.4 Performance Comparison of parameter-based (PRP) and parameter-free (VS and ICN) methods in leave one out cross validation.](image)

When we increased the rank cut-off (k), VS-based approach performed similarly as PRP0.3. In addition, compared to another parameter-free method ICN, our VS-based approach achieved a significant better performance. It was also noted that the performance of VS was beyond PRP when the back probability was set to 0.05.

We believe that the improvement of the performance of VS over ICN is due to the “extended guilt by association” [116] principle from VS. For instance, if we consider an un-weighted network with all edge weights equal to 1 in Figure 4.5. Nodes A and B are not adjacent and have no shared common neighbors. The shortest path between

58
them is A-C-D-B. In this case, the score from ICN [109] would be 0 since there is not even a single shared node between A and B. However, by applying VS, we can calculate the similarity score of A and B by multiplying the similarity scores between all pairs of the adjacent nodes on the shortest path (Sim(A,B)=Sim(A,C)*Sim(C,D)*Sim(D,B)=0.276). Although, extensive analysis on disease gene connectivity have been not performed yet, from the examples we have analyzed, we have discovered that several causal genes of a specific disease are connected distantly (e.g. 3-hop away).

![Diagram of gene network](image.png)

**Figure 4.5** An example of distant connections between hypothetical genes A and B. Nodes A and B do not have an common connected neighbor and the shortest path connecting them is A-C-D-B

However, since the diameter of biological networks tend to be low [4], we believe the lower values of the steps/hops, the more preferable. Interestingly, a previous study reported examples on two real data applications where the number of hops (m) between disease causal genes were set to two, the result showed that when m = 2, it was more preferable compared to m = 1 [85]. Since it is usually noisy and incomplete for edge information between genes, we believe that our VS-based approach for novel
candidate disease gene ranking is desirable as it takes into account alternative measures of pair-wise interconnectedness and it is not only restricted to direct connections or having a shared neighbor node.

4.3.2 Using VS Based Approach to Identify and Rank Novel OD Candidate Genes

After the validation of our method, we next applied our algorithm on several ODs to identify and rank potential novel candidate genes for these ODs. We used the entire protein-protein interaction (PPI) network as our global network for the candidate gene ranking. The top-five ranked candidate genes of the ten selected ODs which have known protein interactions for all of their causal genes were analyzed. The test set genes were gathered from several different sources that consisted of protein interactions, functional linkages and literature co-citations. Briefly, for each causal gene of an OD, we collected its directly connected neighboring genes from the above mentioned resources (see Figure 4.3 and Methods for additional details). Table 4.1 shows the top-five ranked gene predictions for ten ODs.

To check whether our VS top-ranked genes were already reported to be associated with their query OD, we performed searches to online databases and scientific publications. Interestingly, we found indications for most of the top candidate genes that they were related to the respective OD. For instance, the top five predictions for cone rod dystrophy are *RDH5, EFEMP1, CRB1, USH1C* and *CABP4*. Among these
five genes which are all genes associated with visual perception (Gene Ontology), two of them (CRB1 and CABP4) are found to be implicated in eye photoreceptor cell development and differentiation [117-119]. For this particular example, we also used PRP (with back probability set to 0.3) and ICN to perform the candidate gene ranking. Another ranking method for candidate genes, which is functional annotation based (ToppGene [84]) was also used. We derived twenty top ranked genes from each of these four methods, and compared their overlaps. As a result, there were five genes (CRB1, EFEMP1, NPHP4, CNGB1 and GUCA1B) in common to all (Figure 4.6), which suggests they are most potential candidate causal genes for this OD.

Table 4.1: Examples of orphan diseases and VS-ranked top 5 candidate genes

<table>
<thead>
<tr>
<th>Orphan disease</th>
<th>No. of known causal genes</th>
<th>VS ranked top 5 candidate genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cone-rod dystrophy</td>
<td>20</td>
<td>CRB1, RHDS, USH1C, EFEMP1, CABP4</td>
</tr>
<tr>
<td>Severe combined immunodeficiency</td>
<td>17</td>
<td>CD35, JAK1, ZAP70, IL2RB, IL4</td>
</tr>
<tr>
<td>Fanconi anemia</td>
<td>15</td>
<td>HES1, SAMO3, CYP1A1, XRCC3, USF1</td>
</tr>
<tr>
<td>Zellweger syndrome</td>
<td>14</td>
<td>PED1, PHX5, ABCD2, ABCD1, ABCD3</td>
</tr>
<tr>
<td>Autosomal dominant Charcot-Marie-Tooth disease, type 2</td>
<td>12</td>
<td>STAT4, FAM1, MARCH6, ST8H, CRYGC</td>
</tr>
<tr>
<td>Gonadal dysgenesis</td>
<td>12</td>
<td>ZFY, ZFX, PCH1, SOX9, AMH</td>
</tr>
<tr>
<td>Hereditary nonpolyposis colon cancer</td>
<td>11</td>
<td>MRC1, USH3, CARO, TRIT1, ELO1</td>
</tr>
<tr>
<td>Papillary or follicular thyroid carcinoma</td>
<td>11</td>
<td>CCR22A, ZBTB83, MT1, AAM5, SETH4</td>
</tr>
<tr>
<td>Romano-Ward syndrome</td>
<td>11</td>
<td>RHE3, NMY1, KCNJ3, ALG10B, KCN19</td>
</tr>
<tr>
<td>MODY syndrome</td>
<td>10</td>
<td>GCKR, IDDM7, NAPA, STEGAL1, NSRL</td>
</tr>
</tbody>
</table>
Among other examples, *HES1* is the top1 ranked gene for Fanconi anemia. Studies showed that it is a novel interacting protein of the Fanconi anemia core complex, and when cells depleted of *HES1*, they exhibit a Fanconi anemia-like phenotype [120]. For gonadal dysgenesis, the two top-ranked genes are *ZFX* and *ZFY*. They are reported to function in sex differentiation, additionally, *Zfx* mutant mice are found to have fewer germ cells than wild-type mice [121]. Likewise, MODY syndrome is connected to kinetic alterations and regulation of glucokinase activity [122, 123] and in our ranking results, GCKR is the top ranked gene for MODY syndrome. Moreover, a recent study in the Japanese families proposes GCKR as a susceptibility gene for familial diabetes [124]. And here we provide the further support for the involvement
of the top-ranked genes of the investigated ODs. The result also suggests that the top scoring candidates that are previously not discovered to associate with these ODs could be potential candidates for further investigation.

4.4 Discussion and Conclusion

In this work, we proposed a vertex similarity method (VS), which is a parameter-free approach for prioritizing candidate disease genes. Unlike the parameter-based approaches which will update or train the parameters and data sets in each step, VS calculates the similarity between vertices. We show that the ranking performance of VS is better than another parameter-free method ICN through the leave-one-out-cross-validation experiments. We also show that it is comparable to parameter-based methods such as PRP, while our method requires less network information and computation load. We utilized the VS-based parameter-free ranking approach to prioritize OD candidate genes. And more importantly, the top ranked candidate genes from our method match the reports from known research literature, which suggest several novel causal relationships for further investigation.

There are some limitations for our approach: First, like most of the gene prioritization approaches with training set, the assumption is that the OD causal genes that we try to explore will be consistent with what is already known about an OD, whereas its genetic basis may not always be the case. Moreover, the method requires known disease casual gene information, which means our approach cannot be used to identify
novel candidate OD genes if such information misses. Furthermore, if an OD has known causal genes but not protein interactome data available then we cannot use VS for such cases. We may consider using other types of networks (coexpression or functional networks) as an alternative approach. Second, it is important to note that the ranking results are from the current protein interactome data we have, the accuracy of the result is based on the current data version. Third, in some cases, if a seed gene has only one known interaction with the other candidate gene, where the degree is 1 for both genes, then the VS score for that candidate gene will be ranked high.
Chapter 5: Drug Repositioning Through an Integrative Framework Across Multiple Heterogeneous Networks

5.1 Background

The time and cost of bringing a new drug to market are enormous. In addition, the cost of failure on developing a new drug is even higher. In order to save the *de novo* drug development cost and reduce the risk of failure, drug repositioning (that using existing drugs to new therapeutic indications of diseases) provides an efficient and economical route in today’s drug development research[46]. The main advantage of drug repositioning is that in *reusing* drugs that have previously passed clinical trials, the risk of failure in future late-stage clinical trials is relatively minimized with potentially faster drug approvals (see Figure 5.1 for some examples of successfully repositioned drugs.

Current *in silico* drug repositioning strategies are built around integration and mining of large data sets from heterogeneous but related biomedical domains [125]. However, identifying “useful” information from such large data silos is not trivial. Thus, there is a critical need for novel, efficient and robust computational approaches to effectively manage, integrate and mine such data to facilitate knowledge extraction and generate testable hypotheses for drug repositioning candidate discovery.
Network-based approaches provide a powerful option for drug discovery and repositioning[126]. Network-based approaches have been proven useful in biomedical domains too for discovering meaningful patterns since a network provides a systems-level view of the corresponding system. For instance, by integrating drug, gene, disease and other biological entity relationships and representing them as different types of networks, and applying appropriate network analysis, interesting drug and disease combinations for repositioning candidates may be discovered [90, 127-129]. In a recent study, Lee et al. [130] constructed a tripartite network by integrating known disease, drug, and protein interactions and by applying a shared neighborhood scoring (SNS) algorithm identified a potential novel indication for the a hypertension drug. In another study, Li et al. [131], constructed a Chinese herb network to measure the strongly connected herbs and herb pairs through a co-module analysis across herb-biomolecule-disease multilayer networks. Other computational approaches for identifying drug repositioning candidates include using connectivity map of gene expression information [132, 133], text mining [134], and side-effect similarity-based approaches [135].
Most of the existing works intend to find the repositioning candidates by comparing the similarity between drugs and diseases in a homogeneous or heterogeneous network or focus only on one few types of feature networks (e.g., expression-based or PPI-based). Here, we integrate as many as fourteen different features to derive the similarity between all possible gene pairs, namely the drug target genes and disease mutant genes. Specifically, we designed and developed a novel drug repositioning candidate discovery framework that combines both information theory and network analyses-based approaches. The similarity between a gene pair is computed by combining network similarity score and mutual information score between them. The network similarity score is calculated by vertex similarity in fourteen different feature-based gene-gene networks. The mutual information score on the other hand is calculated based on the frequency a gene-gene pair occur in the fourteen feature networks. We believe it is a more comprehensive and unbiased approach compared to existing approaches since our framework quantifies gene pairs not only from the network topology but also from their statistical property. The final pair similarity score contains more information since it is not only from a single network. By extracting the related drug and disease information from the top ranked gene pairs or clusters, potential drug repositioning candidates may be discovered.

In this study, we applied our approach to identify existing drugs to new indications on orphan diseases (OD) as an application. Orphan diseases are rare diseases affecting a
small percentage of the population. In the United States, an orphan disease is defined as any disease or condition that affects less than 200,000 persons or about 1 in 1,500 people. Despite the low prevalence, there are currently around 7,000 ODs affecting approximately 25 million patients in North America [2]. However, there are only about 325 approved drugs (Orphan Drugs) available to treat these diseases, which cover just ~5% of all known ODs [48]. The need and opportunity to discover therapeutics for ODs is thus critical. Because the conventional drug discovery process is both time consuming and very expensive, identifying new indications (orphan diseases in this case) for drugs already in the market is appealing.

5.2 Material and Methods

5.2.1 Gene Pairs and Biological Networks

In order to capture the potential relationships between new indications or diseases and approved drugs, we first generate all disease-gene and drug-target pairs. In other words, using known disease-gene and drug-gene relationships, we generate all possible pairwise combinations wherein one node is a disease-gene and another is drug-gene. We exclude those genes which are both disease-gene and drug-gene.

We used two data sets in our study. The first data set, which we called the OD-All-Drug set, includes all ODs that have at least one known causal gene, and all the approved drugs that have at least one known target gene. We downloaded the approved drugs and their target information from DrugBank [136]. There were 1,261
FDA-approved drugs that had a known target (total 1,302 target genes for 1261 drugs). The ODs and their causal gene information were downloaded from Orphanet [65], and this set contained 2,451 OD casual genes representing 1,690 ODs. The final matrix thus constituted ~3 million unique pairs (1,302 drug genes multiplied by 2,451 OD genes), with self-loops removed.

The second data set (“known indications set”) comprises 818 known therapeutic indications (approved disease-drug relationships). This data set consists of 157 diseases (with 552 disease casual genes) and 414 drugs (with 531 drug target genes)[137]. This data set thus yielded about 0.29 million disease-gene and drug-target pairs (513 drug targets multiplied by 552 disease genes, as described in previous section). The diseases in this data set include predominantly common and some ODs. The corresponding drugs and drug targets are downloaded from the DrugBank. Figure 5.1 summarized the two data sets we used in our study. All the subsequent analyses were performed on these two data sets with three goals:

1. Can we identify potential OD indications for existing approved drugs (data set 1)?
2. How many of the original indications can our framework discover (data set 2)?
3. Can we identify new indications for the approved drugs (data set 2)? This is same as (1) above however the discovered potentially novel indication includes predominantly common diseases and is not limited to ODs alone.
We selected 14 feature based gene networks for our investigating and quantifying the gene pairs (Table 5.1). The orphan disease mutant gene network (ODMGN) [2] was constructed from the bipartite network of ODs to their mutant genes, where the nodes are orphan disease mutant genes and the edges are their shared ODs. The structural interaction network (SIN) was downloaded from [138]. The human protein-protein interaction (PPI) network was compiled from several resources [74-79] with both redundant interactions and self-loops removed. The pathway_interactome network was downloaded from [110]. The other feature networks were downloaded from Gene2Fans [139]. Table 5.1 gives a summary of the properties of the 14 networks.
used in our study.

Table 5.1 14 feature based gene networks

<table>
<thead>
<tr>
<th>Feature based gene networks</th>
<th>Number of Nodes(genes)</th>
<th>Number of edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>ODMGN</td>
<td>1913</td>
<td>11087</td>
</tr>
<tr>
<td>CMAP</td>
<td>8924</td>
<td>61362</td>
</tr>
<tr>
<td>Domains</td>
<td>6746</td>
<td>46463</td>
</tr>
<tr>
<td>Drug_target</td>
<td>2121</td>
<td>16807</td>
</tr>
<tr>
<td>GeneRIFs</td>
<td>3777</td>
<td>27487</td>
</tr>
<tr>
<td>GeneSigDb</td>
<td>10536</td>
<td>65776</td>
</tr>
<tr>
<td>GO_BP</td>
<td>4287</td>
<td>29988</td>
</tr>
<tr>
<td>GO_MF</td>
<td>2944</td>
<td>23356</td>
</tr>
<tr>
<td>Mammalian_phenotype</td>
<td>7552</td>
<td>52637</td>
</tr>
<tr>
<td>Metabolities</td>
<td>3577</td>
<td>28617</td>
</tr>
<tr>
<td>OMIM_disease</td>
<td>1618</td>
<td>22643</td>
</tr>
<tr>
<td>SIN</td>
<td>2341</td>
<td>3133</td>
</tr>
<tr>
<td>PPI</td>
<td>11765</td>
<td>69167</td>
</tr>
<tr>
<td>Pathway_interactome</td>
<td>6045</td>
<td>88611</td>
</tr>
</tbody>
</table>

5.2.2 Mutual Information (MI) Between Genes

We propose $MI$ as a measurement of genes’ similarity indicated by the frequency with which a gene pair occurs among the 14 feature networks. MI is calculated using an entropy notion (Shannon's entropy [140]). This measure gives us a metric that is
indicative of how much "information" from the frequency profile of one gene can be
obtained to predict the behavior of the other gene. The higher the $MI$ score, the more
similar are the two genes.

The mutual information between a gene pair $(x, y)$ is quantified by:

$$MI(x, y) = P(x,y) \log \left( \frac{P(x,y)}{P(x)P(y)} \right)$$

(1)

Where $P(x,y)$ is calculated by the frequency of pair $(x, y)$ in all of the 14 networks
divided by the sum of the number of pairs in the 14 networks. $P(x)$ is calculated by
the frequency of node $x$ in all of the 14 networks (indicating how many times $x$ occurs
in all pairs) divided by the sum of total number of pairs in the 14 networks. $P(y)$ is
also calculated in similar manner.

5.2.3 Vertex Similarity (VS) Between Genes

In our previous studies [141], we have proposed a vertex similarity (VS) framework
to prioritize orphan disease casual genes. We proposed VS score to define the
similarity of two vertices based on the topology of the gene networks - higher the VS
score, higher the similarity between two genes. The VS score between a gene pair $(x, y)$
can be calculated by:

$$VS(x, y) = \frac{\sum_{i=1}^{n} \omega_{x,i} \times \omega_{y,i}}{\sqrt{\sum_{i=1}^{n} (\omega_{x,i})^2} \times \sqrt{\sum_{i=1}^{n} (\omega_{y,i})^2}}$$

(2)

where $\omega_{i,j}$ represents the edge weight of node $x$ to node $i$, and $n$ is the number of the
nodes which includes \( x, y \) and all the nodes that are directly associated with \( x \) and \( y \). In this work, we define VS score for gene pairs with either a direct or indirect connection (1 or 2 hops distance) only. In other words, we calculated the VS scores for gene pairs that are 1 or 2 hops away in the network. If the distance between a gene pair is larger than 2 hops, we consider the similarity score for that pair to be 0.

### 5.2.4 Filtering Gene Pairs

Since ~3 million gene pairs (0.29 million pairs for known indications data set) is a large number, we pruned it further by removing all such pairs that do not appear in any of the 14 networks. We reasoned that by doing this we will eliminate the possibility of pairs that have no MI. Additionally, we also reasoned that the presence of a gene pair in at least one of the 14 feature networks indicates some prior or known relationship. Surprisingly, after applying this condition, we were left with a much smaller data set comprising 29k pairs and 4k pairs from the OD-all-drugs and known indication data sets respectively.

### 5.2.5 Combining MI and VS to Construct Different Weighted Networks

Using the filtered data sets, we computed the mutual information \( MI(x, y) \) and VS scores (in 14 different feature networks \( VS_i(x, y), \ldots, VS_{14}(x, y) \)) for each of the filtered gene pairs. We combined the different VS scores with \( MI \) of each pair to construct 14 different weighted networks \( (N_1, \ldots, N_{14}) \). Since VS score is calculated based on a
cosine similarity, it ranges between [0, 1]. To facilitate a composite MI-VS score, we normalized the $MI$ of each pair also to [0, 1] scale by $\frac{MI}{MI_{MAX}}$, where $MI_{MAX}$ is the maximum $MI$ of the 29k and 4k pairs. Thus, the weight of pair $(x,y)$ in each $Ni$ is calculated by $w_{Ni}(x,y) = MI_{n}(x,y) + VS_{i}(x,y)$, ..., $w_{Ni}(x,y) = MI_{n}(x,y) + VS_{id}(x,y)$. By taking into account both topological and feature-based similarities, we compute a comprehensive and unbiased edge scoring method for each of the gene pairs. For example, a gene-gene pair with a low $MI$ may have high $VS$, and vice versa.

5.2.6 Clustering Different Weighted Networks and Meta Cluster Associating

Hypothesizing that the analysis of network connectivity or modularity may suggest novel combinations of targets or drugs that may result in identifying drug repositioning candidates, we further partitioned or “modularized” the different weighted network $Ni$ using ClusterONE, a state-of-art network clustering algorithm. The ClusterONE algorithm [142], available as a plug-in in the Cytoscape [143], looks for clusters of high cohesiveness based on a greedy strategy. We selected ClusterONE-identified top ranked clusters from each $Ni$, (clusters with $p<0.05$). Using the top-scored clusters from each of the 14 networks, we created a new gene-gene weighted network $P$ (Figure 5.4) where the nodes are genes and edges represent shared clusters. The weight of the edge was a Jaccard Index (JI) calculated with each gene pair’s intersection of their clusters divided by the union of their clusters. We repeated the ClusterONE partitioning on this shared clusters-based gene-gene network
to identify “metaclusters”. Since the edges in these metaclusters represent shared clusters which themselves have been identified from different feature networks, we believe that these clusters represent biologically relevant functional patterns and harbor potential repositioning candidates. See Figure 5.2 and 5.3 for a schematic representation of our drug repositioning framework.

Figure 5.2 Flow chart of the framework (part1)
Form gene to gene pairs with shared/unshared clusters (Jaccard Index)

Build new network $P$ ($|J| \geq 0.05$ as edge weight)

Re-Cluster network $P$

Form disease to drug combinations from each cluster

Find possible repositioning candidates from the combinations

Figure 5.3 Flow chart of the framework (part 2)
5.2.7 Recovery Rate and Literature Search for Validations

Because there is no reasonably good and ready-to-use “gold standard” for repositioned drugs, we were unable to undertake a systematic approach to validate our strategy. We therefore undertook 2 alternative measures as a test for validating or testing the utility of our approach.

First, we checked for the “recovery rate” of known drug-to-disease indications. The recovery rate is calculated as the number of known drug-to-disease combinations that can be found in all drugs to disease combinations generated in our method, divided by the all possible known indications from the disease and drug target genes involved in
the method. Specifically, the drug-to-disease combinations are formed from each of the top motifs derived from network $P_1$, and the all possible known indications are derived from all the genes’ drug and disease information in network $P_1$.

Second, we checked literature for co-citations of our framework-discovered disease-drug pairs. We used CoPub [144] for this. CoPub uses the entire Medline library to calculate robust statistics for keyword co-occurrence. It adopts an R-scaled score, which describes the strength of a co-citation between two keywords given their individual frequencies of occurrence [145], and the literature count, which is the number of co-cited publications between every keyword pair. We used the drug and disease combinations as the keywords, and performed the search in CoPub to check if there are any previous reports associating a drug and disease pair discovered in our analyses.

5.3 Results and Analysis

5.3.1 Known Indications Dataset

The metaclusters derived from the Known Indications dataset were mined to extract the cluster-associated disease-drug pairs. We compared these discovered disease-drug pairs with the original known indications to see how many of them were recovered. Out of a total possible known indications of 407, we were able to recover 130 with a recovery rate of 31.94% (130/407). The remaining ~68% (“unknown indications”) which were not in the original known indications data set were checked in literature
for presence of potential repositioning candidates. We also searched the recovered 130 known indications in the literature using CoPub. As expected these pairs were highly ranked (average number of the co-citation literature was 164.5 and the average R-scale score was 45.87). This justifies our approach of using a literature-based search for identifying and ranking potential repositioning candidates.

We next used the 68% of unknown indications drug-disease pairs from our analyses for literature search using CoPub. Interestingly, for about 60% (1,073/1,761) of the unknown indications, we could find co-citation records in CoPub, with an average co-citation number of 68.95 and R-scale number of 28.73. We manually checked these further to identify potential drug repositioning candidates.

Table 5.2 lists some of the interesting drug repositioning candidates discovered using our method. The disease and drug pairs in this table were not in our approved indications dataset but have been supported by multiple evidences (publications, technical reports, and ongoing clinical trials). However, not all the discovered disease-drug pairs are repositioning candidates and some of these may represent causal relationships and adverse events too. For instance, in table 5.2 the last two examples (osteoporosis-prednisone and osteoporosis-medroxyprogesterone) are cases of adverse events as prednisone and medroxyprogesterone predispose women to osteoporosis. Nevertheless, these relationships are also useful and could be potentially used to understand the molecular basis of adverse drug reactions and also help in
making informed decisions in selecting a drug especially in cases where drugs with relatively safer profiles are available.

Table 5.2 High-scored disease-drug pairs discovered from the analysis of known indications data set

<table>
<thead>
<tr>
<th>Disease</th>
<th>Drug</th>
<th>R-scale Score</th>
<th>Co-citation count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obesity</td>
<td>Rosiglitazone</td>
<td>39</td>
<td>196</td>
</tr>
<tr>
<td>Obesity</td>
<td>Pioglitazone</td>
<td>39</td>
<td>138</td>
</tr>
<tr>
<td>Obesity</td>
<td>Fluoxetine</td>
<td>32</td>
<td>101</td>
</tr>
<tr>
<td>Obesity</td>
<td>Topiramate</td>
<td>36</td>
<td>80</td>
</tr>
<tr>
<td>Obsessive-comulsive disorder</td>
<td>Risperidone</td>
<td>44</td>
<td>83</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>Simvastatin</td>
<td>32</td>
<td>28</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>Donepezil</td>
<td>37</td>
<td>42</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>Memantine</td>
<td>36</td>
<td>28</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>Prednisone</td>
<td>36</td>
<td>247</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>Medroxyprogesterone</td>
<td>39</td>
<td>108</td>
</tr>
</tbody>
</table>

5.3.2 OD-All-Drugs Dataset

The OD-All-Drugs dataset was also analyzed in a similar fashion. Briefly, after generating clusters-based network $P_2$ (genes connected based on shared top-ranked clusters from the 14 feature networks), we re-partitioned it using ClusterONE. We mined these metaclusters to generate OD-drug pairs (see Figures 5.5 to 5.7). In one of
the metacluster (Figure 5.5) with six genes, there were 12 possible disease-drug pairs (3 drugs multiplied by 4 diseases from this metacluster) discovered. The fact that some of these drugs are under different phases of clinical trials for their usage in ODs was encouraging. For instance, we could find several reports that indicate usage of orlistat with positive benefits for cholesteryl ester storage disease, Wolman disease, and hyperlipidemia (types I and II).

Likewise, from another metacluster (Figure 5.6), we found several novel disease-drug pairs of which imatinib has been reported to have therapeutic effectiveness for papillary or follicular thyroid carcinoma disease.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Node-Type</th>
<th>Node-Drug</th>
<th>Node-OD</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPL</td>
<td>Both-DrugTarget-ODMG</td>
<td>Tyloxapol</td>
<td>Hyperlipoproteinemia</td>
</tr>
<tr>
<td>LIPF</td>
<td>DrugTarget-Only</td>
<td>Orlistat</td>
<td>null</td>
</tr>
<tr>
<td>CES1</td>
<td>DrugTarget-Only</td>
<td>Oseltamivir</td>
<td>null</td>
</tr>
<tr>
<td>PNLIP</td>
<td>DrugTarget-Only</td>
<td>Orlistat</td>
<td>null</td>
</tr>
<tr>
<td>LIPA</td>
<td>ODMG-only</td>
<td>null</td>
<td>Cholesteryl ester storage disease, Wolman disease</td>
</tr>
<tr>
<td>LIPC</td>
<td>ODMG-only</td>
<td>null</td>
<td>Hyperlipidemia</td>
</tr>
</tbody>
</table>

Figure 5.5 Repositioning examples (1) from top clusters of network P2 from orphan disease to drug data set
In Figure 5.7, we found that gene CYP11A1’s target drug, aminoglutethimide, has clinical reports on curing prostate and breast cancers. Additionally, older publications indicate usage of this drug for congenital adrenal hyperplasia and familial hyperaldosteronism (blue highlighted cells).

While literature-mining (e.g., using CoPub) and/or clinical trials search can assist in prioritizing some of our discoveries, we could still miss potential novel repositioning candidates simply because there are no previous reports or they are recently reported and therefore some of the literature search engines miss them. One such example was abatacept and primary biliary cirrhosis (PBC). While CoPub and clinical trials search did not yield any “hits” for this pair, a very recent published study [146] reports that an optimized regimen with CTLA-4-Ig (commercially known as Abatacept;
Bristol-Myers Squibb) is a potential therapeutic candidate for patients with PBC.

What makes this discovery more interesting is the fact that the underlying disease-gene and drug-target pairs (IL12RB1 to CD86, and IL12RB1 to CD80) were not present in any of the original 14 feature networks. However, these inferred relationships (edge prediction) constituting our metaclusters represent a useful facet that is not just confined to known relationships between genes.

Surprisingly, the CoPub search results for our discovered OD-drug pairs from metaclusters were not encouraging with only 10% of the combinations being supported by a publication. However, we could trace the problem to the OD nomenclature. CoPub failed to recognize a majority of OD names and we believe that normalizing the OD name space will improve the publication recovery rate.

![Figure 5.7 Repositioning examples (3) from top clusters of network P2 from orphan disease to drug data set](image)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Node-Type</th>
<th>Drug</th>
<th>OD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP21A2</td>
<td>ODMG-only</td>
<td>metyrapone</td>
<td>Congenital adrenal hyperplasia</td>
</tr>
<tr>
<td>MSMB</td>
<td>ODMG-only</td>
<td>mitotane</td>
<td>Familial prostate cancer</td>
</tr>
<tr>
<td>RNASEL</td>
<td>ODMG-only</td>
<td>Nitazoxanide</td>
<td>Congenital adrenal hyperplasia</td>
</tr>
<tr>
<td>MXI1</td>
<td>ODMG-only</td>
<td></td>
<td>Familial prostate cancer</td>
</tr>
<tr>
<td>CHEK2</td>
<td>ODMG-only</td>
<td></td>
<td>Hereditary breast and ovarian cancer syndrome</td>
</tr>
<tr>
<td>BRCA1</td>
<td>ODMG-only</td>
<td></td>
<td>Familial prostate cancer, Hereditary breast and ovarian cancer syndrome, Primary peritoneal carcinoma, Hereditary breast and ovarian cancer syndrome</td>
</tr>
<tr>
<td>BRCA2</td>
<td>ODMG-only</td>
<td></td>
<td>Familial prostate cancer, Fanconi anemia, Hereditary breast and ovarian cancer syndrome, Nephroblastoma</td>
</tr>
<tr>
<td>PTEN</td>
<td>ODMG-only</td>
<td>Nitazoxanide</td>
<td>Bannayan-Riley-Ruvalcaba syndrome, Cowden syndrome, Familial prostate cancer, Hereditary breast and ovarian cancer syndrome, Juvenile polyposis of infancy, Lhermitte-Duclos disease, Macrocephaly, Proteus syndrome, Segmental outgrowth - lipomatosis - arteriovenous malformation - epidermal nevus, Squamous cell carcinoma of head and neck</td>
</tr>
<tr>
<td>CYP11B1</td>
<td>Both-DrugTarget-ODMGMetyrapone</td>
<td>Mitotane</td>
<td>Congenital adrenal hyperplasia, Familial hyperaldosteronism</td>
</tr>
<tr>
<td>CYP11A1</td>
<td>Both-DrugTarget-ODMGAminoglutethimide</td>
<td>Aminoglutethimide</td>
<td>Congenital lipoid adrenal hyperplasia, Disorder of sex development</td>
</tr>
<tr>
<td>POR</td>
<td>Both-DrugTarget-ODMGNitazoxanide</td>
<td></td>
<td>Antley-Bixler syndrome, Congenital adrenal hyperplasia</td>
</tr>
</tbody>
</table>
5.4 Conclusions

In this chapter, we proposed an integrative framework to quantify gene pairs, construct networks, and cluster the networks to predict drug repositioning candidates. Our method has been verified useful through known indication recovery, literature and clinical trials data mining. Two principal novel methodical contributions our approach has are:

First, we quantify the relationships between disease-genes and drug-targets based on their topological similarity and mutual information. By doing this, not only a gene pair can be quantified in a more unbiased way but can also be complementary in cases where the feature networks are sparse owing to lack of annotations.

Second, by generating metaclusters based on shared clusters, we predict potential relationships between a disease-gene and drug-target. In our method, after deriving the top clusters from the 14 weighted networks ($N_1$ to $N_{14}$), we compare and mine these clusters to generate a gene-gene network (and network $P$) wherein the edge represents shared top clusters from the 14 feature networks. For example, in the final shared-clusters-based gene-gene network, there are about 15% pairs (3768 out of 25270 pairs) that were not connected in any of the 14 networks. The fact that these predicted edges are supported by shared top clusters suggests that they could be functionally relevant and may also harbor potential drug repositioning candidates. For example, abatacept and primary biliary cirrhosis pair was based on two gene pairs
(IL12RB-CD86 and IL12RB1-CD80) which were predicted and not connected in any of the original 14 feature networks.

Our approach has some limitations. First, because our approach is gene-centric, diseases and drugs that do not have gene annotations will be ignored. Second, we currently do not have a systematic and efficient ranking strategy for the discovered disease-drug pairs. We used literature and clinical trials-based searches to “validate” or prioritize our predictions. However, this strategy may miss true positives simply because the records related to a novel association have not yet been added to the literature database (for example, abatacept and PBC). Third, not all discovered disease-drug pairs are repositioning candidates and some of them could be causative or adverse event related. We currently do not have a strategy to identify or separate these out. However, this information could also be useful in understanding the molecular basis of side effects or for making informed decisions in drug selection. Lastly, the lack of a “real” gold standard for drug repositioning makes it hard to validate our approach in a more systematic way.

Some of the future extensions we envisage for this project are:

1. Using phenotype (symptoms)-based disease networks (e.g., electronic health records) and side-effect-based drug networks (e.g., FDA’s Adverse Event Reports System).

2. Moving from gene-centric to pathway-centric approaches.
3. Building a Web-based interactive server to facilitate users query the integrated networks and obtain a list of drug repositioning candidates.
References

20. Palla, G., et al., *Uncovering the overlapping community structure of complex


55. Florez AF, P.D., Bhak J, Protein network prediction and topological analysis
127. Mei, H., et al., *Opportunities in systems biology to discover mechanisms and...*
146. Dhirapong, A., et al., Therapeutic effect of cytotoxic T lymphocyte antigen