Incorporation of Knowledge for Network-based Candidate Gene Prioritization

by

Chad Kimmel

B.S., Mount Union University, 2006

M.S., University of Pittsburgh, 2008

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2012

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Chad Kimmel

It was defended on

July 10, 2012

and approved by

Naftali Kaminski, M.D.

Professor, Department of Medicine, University of Pittsburgh

Vanathi Gopalakrishnan, Ph.D.

Associate Professor, Department of Biomedical Informatics, University of Pittsburgh

Madhavi Ganapathiraju, Ph.D.

Assistant Professor, Department of Biomedical Informatics, University of Pittsburgh

Dissertation Advisor: Shyam Visweswaran, M.D., Ph.D.

Assistant Professor, Department of Biomedical Informatics, University of Pittsburgh

Incorporation of Knowledge for Network-based Candidate Gene Prioritization

Chad Kimmel

University of Pittsburgh, 2012

In order to identify the genes associated with a given disease, a number of different high-throughput techniques are available such as gene expression profiles. However, these high-throughput approaches often result in hundreds of different candidate genes, and it is thus very difficult for biomedical researchers to narrow their focus to a few candidate genes when studying a given disease. In order to assist in this challenge, a process called gene prioritization can be utilized. Gene prioritization is the process of identifying and ranking new genes as being associated with a given disease. Candidate genes which rank high are deemed more likely to be associated with the disease than those that rank low. This dissertation focuses on a specific kind of gene prioritization method called network-based gene prioritization. Network-based methods utilize a biological network such as a protein-protein interaction network to rank the candidate genes. In a biological network, a node represents a protein (or gene), and a link represents a biological relationship between two proteins such as a physical interaction.

The purpose of this dissertation was to investigate if the incorporation of biological knowledge into the network-based gene prioritization process can provide a significant benefit. The biological knowledge consisted of a variety of information about a given gene including gene ontology (GO) functional terms, MEDLINE articles, gene co-expression measurements, and protein domains to name just a few. The biological knowledge was incorporated into the network's links and nodes as link and node knowledge respectively. An example of link knowledge is the degree of functional similarity between two proteins, and an example of node

knowledge is the number of GO terms associated with a given protein. Since there were no existing network-based inference algorithms which could incorporate node knowledge, I developed a new network-based inference algorithm to incorporate both link and node knowledge called the Knowledge Network Gene Prioritization (KNGP) algorithm.

The results showed that the incorporation of biological knowledge via link and node knowledge can provide a significant benefit for network-based gene prioritization. The KNGP algorithm was utilized to combine the link and node knowledge.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I have so many people to thank for this dissertation. First, I thank my committee members, Naftali Kaminski, M.D., Madhavi Ganapathiraju, Ph.D., and Vanathi Gopalakrishnan, Ph.D. for their steady support and guidance.

I am also grateful for the financial support provided by the National Library of Medicine (Medical Informatics Training Grant number 5 T15 LM007059-24) and the Clinical and Translational Research Institute at the University of Pittsburgh for my graduate studies.

I especially want to thank my adviser Shyam Visweswaran, M.D., Ph.D. for his guidance and assistance with my research. We have both spent a countless number of hours on this dissertation together, and it's an indelible moment to see it come to finality. I thank Shyam so much for all of the time and work he has put into my experience at Pitt, and I can't tell you how proud I am to be his first doctoral student.

And lastly, I want to thank my loving parents who have been very supportive of me throughout my educational career – from pre-school to graduate school. I can't tell you how proud I am to be their son, and I could not have achieved this accomplishment without them. They provided me invaluable support when I ran out of funding as a grad student, and I could not have achieved this accomplishment without them. I love my parents to death. Thank you mom and dad!

I came to Pittsburgh six years ago straight from my under-grad – very inexperienced, but energetic and youthful at the same time. I was truly excited beyond the stars at the opportunity to study at Pitt, and after a long and adventurous road, here I am six years later: ready to graduate with my PhD – hard to believe. I have had so many amazing and unforgettable experiences at Pitt, and I have grown in so many ways: academically, professionally, socially, and spiritually to name just a few. I will especially remember the many memorable experiences I had with the Catholic Newman Center at Pitt – a big thanks to the Oratorians and all my friends at the Newman Center for being that rock of friendship and support during my many years at Pitt. Pitt will always be in my heart forever, and I will never forget the city and University that I love. H2P!

GLOSSARY

Candidate gene prioritization – is the process of identifying and ranking new genes as potential candidates of being associated with a disease or phenotype.

Knowledge network – is a graph that consists of nodes and links between pairs of nodes where nodes represent entities and the links represent a variety of pair-wise relations that can exist among the entities. For example, in a protein-protein interaction knowledge network, nodes represent proteins, and the links represent pair-wise interactions among the proteins.

Link weight – is a value assigned to a link in the knowledge network. The link weight is a number that characterizes a relationship between a pair of nodes.

Node weight – is a value assigned to a node in the knowledge network. The node weight is a number that characterizes a node property.

Root nodes – are nodes known to be associated with a given concept (e.g., disease) in a knowledge network. The set of root nodes is denoted by *R*.

Root set – is a set of genes or proteins known to be associated with a given disease.

Candidate nodes – are nodes in a knowledge network that a user wants to rank or prioritize relative to a set of root nodes. The set of root nodes is denoted by *C*.

Candidate set – is a set of genes or proteins that a user wants to rank or prioritize relative as being associated with a given disease.

Inference – in a network is the process of ranking nodes relative to a set of root nodes. Examples of network inference algorithms include PageRank and PageRank with Priors.

PageRank – is a network inference algorithm that is widely used by internet search engines.

PageRank with Priors – is an extension of the PageRank algorithm which incorporates a prior probability vector for nodes of the network.

Knowledge network gene prioritization (KNGP) algorithm – is a new network inference algorithm that was developed in this dissertation and allows the incorporation of both link weights and node weights. In contrast, the PageRrank and PageRank with Priors algorithms are able to incorporate only link weights.

1.0 INTRODUCTION

Understanding the genetic and biological mechanisms of diseases is an ongoing challenge. Common diseases such as Alzheimer's disease and asthma that occur relatively frequently in the population are likely to have complex and multifactorial underlying mechanisms. Moreover, common diseases likely arise from a combination of several genetic factors that interact with environmental factors. In recent years, several high-throughput techniques that survey a large number of genes or even the entire genome have been developed for elucidating the genetic factors of common diseases. Such techniques include, for example, gene expression profiling, genotyping of single nucleotide polymorphisms (SNPs), and whole genome sequencing. One challenge with such techniques is that they typically produce hundreds of candidate genes associated with the disease of interest. In this dissertation, I focus on one approach to reduce the number of candidate genes for a disease of interest that can then be examined in detail by the biomedical researcher. This approach integrates several types of knowledge and information about genes in general with knowledge of genes already known to be associated with the disease of interest and produces a small set of candidate genes.

Candidate gene prioritization is the process of identifying and ranking new genes as potential candidates of being associated with a disease or phenotype. Genes that rank higher are more likely to be associated with the disease and more worthy of further biological investigation compared to those genes that rank lower. Most candidate gene prioritization methods rely on a

set of genes that are already known to be associated with the disease (the root set) to rank the other genes. Developing excellent methods for candidate gene prioritization is important, because such methods can save biomedical researchers a significant amount of time, effort and resources by allowing them to focus on a relatively small set of promising genes to be studied in depth. Thus, candidate gene prioritization has enormous potential for accelerating progress in translational bioinformatics and in the development of new therapies.

Many gene prioritization methods rank candidate genes based on the similarity between the candidate genes and the genes in the root set. Similarity between genes is typically computed from known knowledge and information about genes such as the function and the cellular location of the corresponding protein. Such knowledge is obtained from functional annotations [1], sequence data [2] and gene expression data [3]. Information from non-human sources have also been shown to be useful in a recent study that incorporated mouse phenotype information [4].

1.1 OVERVIEW OF NETWORK-BASED APPROACH

More recently, network-based approaches have been applied to candidate gene prioritization. In the network-based approach, biological knowledge about genes is represented as a network. A network is a mathematical object that consists of nodes and links between pairs of nodes where nodes represent entities and the links represent a variety of pair-wise relations that can exist among the entities. For example, in a protein-protein interaction network (PPIN), nodes represent proteins, and the links represent pair-wise interactions among the proteins. In a co-expression network, nodes represent genes measured in a microarray experiment, and the links may

represent correlations between expressions of pairs of genes. A network that incorporates knowledge is called a knowledge network.

In network-based gene prioritization, an inference algorithm is applied to the knowledge network to rank the genes relative to the root set of genes (or proteins) associated with a disease of interest. The premise underlying this approach is that genes in the network that are in close proximity to genes in the root set are more likely to be associated with the disease than those that are further away. Proximity between genes in a network can be defined and computed using a variety of inference methods that have been developed for social- and Web-network analysis such as PageRank [5] and Hyperlink-Induced Topic Search (HITS) [6].

In this dissertation, I investigate - in depth - the network-based approach for the candidate gene prioritization problem. In particular, I investigate how a variety knowledge sources about genes can be incorporated into the network, and if such incorporation is useful for improving the network-based gene prioritization process. In the past, researchers have investigated the incorporation of only a single type of knowledge in network-based gene prioritization. My hypothesis is that combining and incorporating several types of knowledge in the network will outperform a network that incorporates only a single type of knowledge. A major challenge in incorporating several knowledge sources is to design a suitable network representation that denotes combined knowledge. In conjunction with this, I investigated two ways of representing knowledge in a network – namely – as nodes and links. An example of knowledge that can be represented as a link between two nodes is the degree of functional similarity between two genes where a node denotes a gene and a link denotes the degree of functional similarity between a pair of genes. Additional examples of link knowledge include known protein-protein interactions, gene expression information, and gene functional

information. An example of knowledge that can be represented in a node is the number of MEDLINE articles associated with a given gene of interest. Additional examples of such node knowledge include the number of gene ontology annotations associated with a gene and the number of functional domains on the corresponding protein derived from a gene. In addition to investigating the utility of each type of knowledge, I also investigate combining several types of knowledge and representing them in the network. Figure 1 illustrates a small network that represents both node and link knowledge in the form of node and link weights.

Inference in a network is the process of ranking nodes relative to a set of root nodes. Examples of network inference algorithms include PageRank and PageRank with Priors. These algorithms can do inference only on a network containing link knowledge - not on networks which contain both link and node knowledge. Because of this limitation of existing inference algorithms, I developed a new inference algorithm called the Knowledge Network Gene Prioritization (KNGP) algorithm which is a generalization of the PageRank and PageRank with Priors algorithms. The PageRank with Priors inference algorithm takes as input a network and a root set and then computes a relative importance score for each of the remaining nodes in the network. This relative importance score is a measure of how likely the corresponding gene is to be associated with the disease of interest. The PageRank algorithm is a link analysis algorithm that was developed by Larry Page and is used by the Google Internet search engine [7]. It assigns a relative importance score to each webpage of a hyperlinked set of webpages with the purpose of measuring its relative importance within the set. The PageRank with Priors algorithm is a generalization of the PageRank algorithm. The Knowledge Network Gene Prioritization inference algorithm is an important contribution of this dissertation.

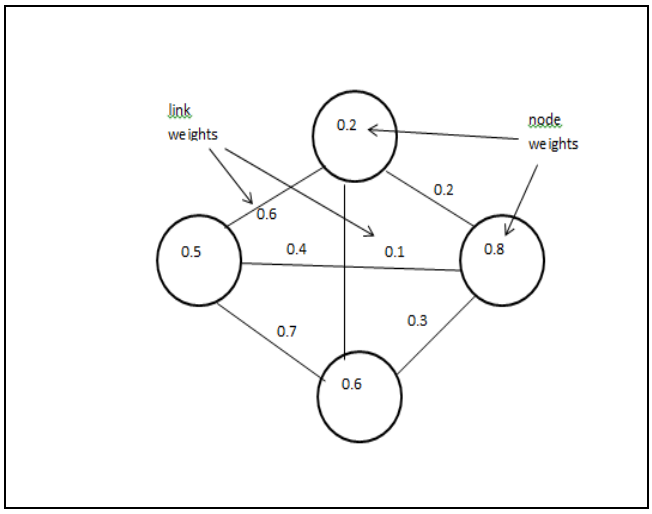

Figure 1. Example of a network with link weights and node weights.

1.2 MAIN AIMS

The main aim of this dissertation is to determine if the incorporation of knowledge is helpful in network-based gene prioritization. The incorporation of knowledge in the form of node knowledge, link knowledge, a combination of link knowledge, and a combination of link and node knowledge together was investigated. In order to determine if the knowledge added any significant benefit, the PPIN was used as the baseline, because this is the type of network that is typically used in network-based gene prioritization. The null and alternative hypotheses for the four main aims are the following:

The null (H_0) and alternate hypotheses (H_1) for the first aim are as follows:

H_0: The incorporation of link knowledge from a single source does not provide a benefit for network-based gene prioritization. The link knowledge investigated include gene

function information from the biological process ontology, molecular function ontology, cellular component ontology, MEDLINE and gene expression measurements.

H_1: Some forms of link knowledge provide a benefit for network-based gene prioritization.

The null (H_0) and alternate hypotheses (H_1) for the second aim are as follows:

H_0: The combination of different types of link knowledge does not provide a benefit for network-based gene prioritization. The types of link knowledge that were combined for investigation included gene functional information and predicted protein-protein interactions.

H_1: The combination of some forms of link knowledge provides a benefit in network-based gene prioritization.

The null (H_0) and alternate hypotheses (H_1) for the third aim are as follows:

H_0: The incorporation of node knowledge does not provide a benefit for network-based gene prioritization. The node knowledge investigated include the number of MEDLINE articles associated with each gene, the number of gene ontology annotations for a given gene, and the number of domains associated with each protein.

H_1: Some forms of node knowledge provide a benefit in network-based gene prioritization.

The null (H_0) and alternate hypotheses (H_1) for the fourth aim are as follows:

H_0: The incorporation of node knowledge and link knowledge together does not provide a benefit for network-based gene prioritization.

H_1: The combination of node and link knowledge together can provide a benefit for network-based gene prioritization.

1.3 CONTRIBUTIONS

There are two major contributions of this dissertation.

The first major contribution is the development of a network-based inference algorithm that can utilize both link and node knowledge for network-based gene prioritization. Currently, there are no network-based inference algorithms which can incorporate node knowledge into the network-based gene prioritization process. Thus, in order to incorporate node knowledge, a new algorithm had to be developed, and this algorithm is called the Knowledge Network Gene Prioritization algorithm.

The second major contribution is the investigation of whether biological knowledge can successfully be used to significantly benefit the network-based gene prioritization process. Both the introduction of link and node knowledge was investigated. The aims listed in the previous section directly address this contribution. In order to determine if the biological knowledge added any significant benefit, the knowledge sources were compared to the protein-protein interaction network, because this network does not represent the incorporation of any new knowledge into the network-based gene prioritization process and represents what is traditionally used for network-based gene prioritization.

1.4 OVERVIEW OF DISSERTATION

Chapter 2 provides the relevant background for gene prioritization. The background includes a broad overview of network and non-network based gene prioritization methods and a detailed review of the common network inference algorithms including PageRank and PageRank with Priors.

Chapter 3 provides a detailed description of the Knowledge Network Gene Prioritization algorithm including its components and computational complexity. Chapter 4 provides details of the experimental methods and describes the knowledge sources used in creating the knowledge networks, the creation of root sets for the experimental diseases and the evaluation protocol. Chapter 5 provides the experimental results including results on networks derived from synthetic data and networks derived from real biological knowledge. Chapter 6 summarizes the contributions and discusses some limitations and future work.

2.0 BACKGROUND

This chapter provides background on gene prioritization methods and inference methods for networks. Section 2.1 gives a general overview of the different types of gene prioritization methods including similarity and network-based methods. Section 2.2 describes the representation of knowledge in networks, and Section 2.3 describes two common network-based inference algorithms including Page Rank and Page Rank with Priors algorithms. Section 2.4 briefly reviews previous work that uses several knowledge sources in networks.

2.1 GENE PROPRITIZATION METHODS

The gene prioritization methods described in the literature can be broadly classified into two groups: similarity-based and network-based methods. Section 2.1.1 provides a review of similarity-based gene prioritization methods and Section 2.1.2 summarizes the literature on network-based similarity methods. The section ends with an overview of network-based methods.

2.1.1 Similarity-based methods

Similarity-based methods attempt to identify those candidate genes whose features are most similar to genes known to be associated with a particular disease. Examples of such features include expression patterns [3], sequence features [1], and functional annotations [8] to name just a few. The following are a rather exhaustive description of the similarity based papers in the literature.

Radivojac et al. [9] constructed a Support Vector Machine classifier using similarity features to predict a gene's association with a disease. Three different types of features were constructed: Protein-protein interaction disease ontology (PPI-DO), protein-protein interaction gene ontology (PPI-GO), and sequence, physiochemical, and other predicted properties (SPP-GO). The PPI-DO and PPI-GO features were constructed by counting the number of disease and gene ontology terms at various protein interaction distances in the PPIN. The interaction distance was defined as the shortest distance between two nodes in the PPIN. The SPP-GO features were constructed from physiochemical or predicted properties (intrinsic disorder, hydrophobic moment, prediction of helix, sheet, coil, predictable surface area, etc.). The classifier was used to predict associated genes for 422 diseases, and the mean area under the ROC curve (AUC) was 73.1%. The authors showed that a state-of-the art classifier using similarity features was able to identify candidate genes with reasonable performance.

Rossi et al. [10] created Transcriptomics of OMIM (TOM). TOM first uses sequence information to identify candidate genes at a given chromosomal area of interest. The candidate genes are then filtered based on their expression profile and GO annotation similarity to the genes already known to be associated with the disease. The algorithm is available online, and it

allows the user to associate gene mapping and functional annotations in the search for candidate genes.

In one of the earliest approaches, Perez-Iratxeta et al. [11] created a gene prioritization system based on a fuzzy set theory. The system calculated gene-disease associations by linking phenotype to genotype and filtered the candidate genes through a MEDLINE search. The system was used to discover gene-disease relationships for 455 genetically inherited diseases.

Chen et al. [4] constructed an application called ToppGene that uses a fuzzy similarity based score which measures the similarity between sets of feature annotations for two genes. The types of annotations used for computing the measures of similarity were GO, Mammalian Phenotype, Pathway, Protein Domain, MEDLINE, and Protein Interactions. In ToppGene, those candidate genes with more similar annotation sets to the known disease related genes were deemed more likely to be associated with the disease.

Adi et al. [1] constructed a gene prioritization tool called PROSPECTR that uses a wide variety of sequence features such as the sequence's percent protein identity to a rat homolog, the number of exons, and whether the protein has a predicted transmembrane domain. These sequence features were input into a decision tree classifier to predict a gene's likely involvement in a disease. The authors showed that PROSPECTR was able to expand the set of genes thought to be implicated in the disease from the root set two-fold 77% of the time. The same authors later created SUSPECTS [2]. SUSPECTS is a freely available web service which combines sequence and annotation purposes for the purpose of gene prioritization. Most notably, the method is able to limit the effect of annotation bias by combining the precision of annotation-based methods with the better recall of sequence-based methods. In conclusion, the authors found that SUSPECTS was an improvement over PROSPECTR.

Aerts et al. [3] constructed Endeavor to prioritize genes for human diseases. Endeavor attempts to integrate multiple data sources which annotate a variety of protein and gene characteristics including functional annotations (Gene Ontology), microarray experiments, and pathway membership. Candidate test genes are then ranked according to their similarity with the training set of genes based on the above characteristics. The authors obtained an AUC of 0.866 by integrating all of the data sources. Endeavor is also able to prioritize genes in biological pathways sets using similar methodology.

George et al. [12] created a methodology called Common Module Profiling (CMP) to prioritize genes in a specific locus. CMP uses a method called SSEARCH – an implementation of the Smith and Waterman alignment algorithm – to calculate the similarity between the domains of the candidate proteins and the known diseased proteins, and this similarity is utilized to identify novel disease genes. With 170 diseased genes for 29 diseases, a specificity of 0.69 and sensitivity of 0.59 was obtained.

Hua et al. [13] proposed an ensemble learner in combination with a bootstrapping method to impute the missing expression values in a microarray. The new expression vectors were then used to prioritize a list of genes using the student's t-test. The authors compared their methodology to a common non-ensemble approach, and they showed that their method was better able to control the false positives in gene prioritization.

De Bie et al. [14] constructed a novel kernel based method for gene prioritization based on a number of data sources including the GO InterPro Domains (IP), and KEGG pathways to name just a few, and the authors showed that a combination of these data sources performed more adequately than using just one data source. The authors showed that their new kernel method outperformed the previous Endeavor methodology mentioned previously.

Hutz et al. [15] created CANDID. CANDID is a gene prioritization tool to output rankings of candidate genes for a given disease. CANDID uses several data sources to produce the rankings including publications, protein domain descriptions, gene expression profiles, cross species conservation measures, and protein-protein interactions. Each candidate gene then receives a score based on the gene's similarity to the traits associated with the desired disease for each data source. For instance, if a candidate gene is very similar to a given disease's protein domains, the gene will receive a high score. In order to produce a final ranking, a user-defined weight for each for each data source is defined, and the scores in conjunction with the weights are then coalesced to produce a final ranking for each gene. CANDID was tested on several known diseases from the Online Mendelian Inheritance of Man (OMIM) and performed adequately. The approach taken by these authors to incorporate the knowledge sources is somewhat similar to the approach taken in this dissertation, even though it was not in the network-based context.

All of the preceding similarity approaches consider a very large set of proteins as the candidate gene set to prioritize for a given disease. However, there are some similarity based methods which only prioritize a small subset of genes. For instance, some prioritization methods only prioritize genes within a given quantitative trait locus (QTL). A QTL is a small stretch of DNA which is suspected to be linked to a given trait through an experimental measure (SNP testing). The following similarity methods prioritize a small subset of genes.

Gauton et al. [16] developed a freely available gene prioritization service called CEASER. CEASER combines data and text mining to rank genes according to a given biological process (such as a disease). CEASER consists of three steps. CEASER uses ontologies to exploit the knowledge of complex traits in the literature; this knowledge is then semantically mapped to

trait and protein-centric information from a variety of data sources such as protein-protein interactions, metabolic pathways and tissue-specific gene expression. CEASER was tested on 18 susceptibility genes for 11 complex traits and shown to be rather successful. The test genes were ranked higher than about 96% of all genes on average.

Shriner et al. [17] utilized an approach which was highly dependent on the GO.. The GO is a network of functional terms with links which describe the function of a given gene. This dissertation utilizes the GO, and the GO is further explained in Section 3.1. Using the GO for gene prioritization, the typical approach is to try to find the genes which are most represented in a given set of interesting genes (i.e., differentially expressed genes). This is called *gene enrichment*. However, the problem with most gene enrichment approaches is that there are several inherent correlations of the terms within the GO, and these correlations are not accounted for in the statistical machinery. In order to alleviate this correlation concern, the authors developed a dimension reduction method through Principle Component Analysis. The authors then applied this method in conjunction with a novel scoring scheme to prioritize genes within a given Quantitative Locus Region (QTL). This method was called Commonality of Functional Annotation (CFA). The method was applied to two complex traits: Alzheimer's disease and Body Mass Index.

Linghu et al. [18] integrated 16 different genomic data sources including protein-protein interactions and expression data among others to create a functional-linkage network. In order to create the functional linkage network, a naïve Bayes classifier was used to compute functional links between all possible gene pairs. The functional links (or weights) represented the probability of the gene pair sharing in the same biological process (i.e., disease) after summing over all the data sources. After the functional link weights were created, a linkage weight cutoff

score (or threshold) was chosen such as to determine if the pair of genes retained a link. This threshold was used to determine if the overall evidence supported the functional linkage. The threshold retained edges with more evidence for functional association and removed edges with more evidence against functional association. In order to prioritize the genes, a similarity-based neighborhood weighting scoring scheme was utilized which prioritized a given gene according to the sum of its weights with the neighboring root genes. The authors used their gene prioritization methodology to predict new candidate disease genes for 110 diseases. Furthermore, the authors showed that the integration of multiple data sources outperformed the use of just individual data sources. It is important to note that the authors used a local based inference algorithm instead of a global based inference algorithm. The difference between these two types of algorithms is described in Section 2.2.2.

2.1.2 Network-based methods

Network-based gene prioritization methods primarily use the topology of a knowledge network where the nodes represent entities, and the links represent relationships between the entities. The most common type of network utilized is a PPIN, but other types of networks such as co-expression networks may also be utilized. Network-based gene prioritization methods make the assumption that the genes associated with a disease are likely to be topologically close together in the network. For example, in a PPIN, the assumption is that the proteins related to a disease are likely to reside in the same sub-network. This section describes papers which utilize the network-based approach to gene prioritization. If possible, the relationship of the paper to this dissertation will be discussed.

Chen et al. [19] compared ToppGene (an integrated functional-similarity based method described in the Section 2.1) to several network-based methods. The authors found that similarity-based methods performed better than any of the network-based methods and concluded that network-based methods are not as effective as integrated functional annotation-based methods. However, the authors also found that the network-based methods performed better than any individual knowledge sources using the similarity-based methods and are easier to apply in practice than integrated functional annotation based methods. Thus, the authors concluded that network-based methods can effectively be used for candidate gene prioritization.

Oti et al. [20] used a PPIN to search for genes associated with a given disease. First, the authors identified the protein-protein interaction partners of a given disease gene on a PPIN. If an interacting gene was found to be within one or more chromosomal loci of a disease gene, then it was considered to be a candidate for the disease. In total, about 300 disease candidate gene predictions were made, and the accuracy of the predictions was tested using a benchmark set of known diseased genes. Of these 300 disease gene predictions, about 10% or more were expected to be genuine disease genes which represented a 10-fold enrichment. Even though the methodology in this paper was very simple, it is significant, because it was one of the first papers to show that protein-protein interactions can indeed be used to discover disease candidate genes.

In an earlier paper, Chen et al. [21] used a network-based gene prioritization algorithm to rank each protein in the Online Predicted Human Interaction Database (OPHID) according to the protein's association with Alzheimer's. Any gene which directly interacted with a root gene on the PPIN was considered to be in the candidate gene list – this is known as a "nearest neighbor" based approach. In order to prioritize the genes, a relevancy score was introduced which utilized the PPIN. Even such a simple gene prioritization approach was shown to be effective. For

example, a protein, *B Catenin*, was predicted to be associated with Alzheimer's disease which had not been previously implicated in the disease. The authors validated the novel finding by showing that, in the literature, B-Catenin was related to Alzheimer's disease via a signaling pathway. This dissertation – in a very similar fashion – utilized the species (e.g., human, fly, worm) that a given interaction was derived from as a knowledge source.

Gonzalez et al. [22] constructed an interaction network using only the interactions involved with genes in the root set (the set of genes already known to be associated with the disease). The genes that interacted with the root set of genes were the candidate set of genes. The candidate genes were then prioritized based on the confidence level of the interactions with the proteins in the root set and how relevant the gene was for maintaining the local inter-connectivity of the protein network. A high degree of local inter-connectivity has been previously shown to identify sets of functionally related proteins [23, 24]. Greater confidence was given to genes derived from curated sources than those derived from a natural language processing system. Using this prioritizing schema, a novel protein, PRKCG (protein kinase C), was found which had not been previously implicated in the disease atherosclerosis. The authors validated the finding by pointing out that protein kinase C is known to play a role in the action of cytokines, and other cytokines – such as IL1 and IL6 – are known to have a strong relationship with atherosclerosis. This paper was one of the first to show that the property of inner-connectivity in a biological network can successfully be used to prioritize proteins.

Kohler et al. [25] constructed a PPIN and used a random walk and diffusion kernel network algorithm to prioritize the candidate genes. The random walk method models the probability of a random walker being at a given protein in a network based on an adjacency matrix. The diffusion kernel method is based on a similar but slightly different method from the

random walk method. The authors compared the random walk and diffusion kernel methods to the previously mentioned PROSPECTR and Endeavor methodologies using a family of diseases caused by a single disease (monogenic), a family of cancer diseases, and a family of diseases caused by multiple genes (polygenic). For the polygenic and cancer families, the authors showed an improvement for the random walk and kernel methods over the previously mentioned PROSPECTR and Endeavor methodologies. This paper showed that global based inference algorithms for gene prioritization are better than local based inference algorithms. This dissertation uses a global based similarity measure. The major difference between global and local based measures is explained in Section 2.2.2.

Wu et al. [26] constructed a PPIN for the purpose of gene prioritization. The authors also utilized phenotype similarity scores in the manner of van Driel et al. [27] which represented how similar two phenotypes (or diseases) were. The authors used the phenotypes (diseases) defined in the OMIM database, and the phenotype similarity scores were calculated from the Medical Subject Heading (MeSH) terms. The MeSH represents the topical associations for a given MEDLINE article. In order to prioritize the genes, a novel method called CIPHER (Correlating protein Interaction network and PHEnotype network to pRedict disease genes) was constructed. The CIPHER is dependent on a novel concordance score. The authors used the prioritization method to successfully rank known disease genes in 709 out of 1444 linkage intervals; the method was shown to be effective for prioritizing genes with little genetic basis.

The vast majority of network-based gene prioritization algorithms described in the literature utilizes a PPIN. However, other types of networks such as co-expression networks have also been utilized. I now briefly describe the literature on the use of other types of networks.

Ala et al. [28] constructed a co-expression network from microarray data to prioritize genes within a specific disease loci. The analysis of the co-expression network began with the construction of co-expression clusters. A co-expression cluster was defined as a given gene plus all of the nearest neighbors in the co-expression network. These clusters were then used to identify new diseased genes in a specific genetic locus. The authors applied their methodology to 850 OMIM phenotype entries where mapped disease loci existed but no diseased genes could be identified. For validation, the authors noted that for 3 of the OMIM phenotypes, the prediction included genes that had already been found to be mutated in the disease, but were not correctly annotated in OMIM. Consequently, this dissertation also uses microarray measurements as a knowledge source.

Nitch et.al. [29] created a gene prioritization method to overcome the problem of insufficient knowledge about genes associated with a given disease in question. To overcome this problem, the authors used differential gene expression data between healthy and disease affected samples. Using a gene/protein network, a candidate gene was accessed by considering the degree of differential expression in the local neighborhood around the gene under the assumption that candidate genes tend to be surrounded by differential expressed neighbors. A weight-based inference algorithm was then used to score and rank all the candidate genes. The authors demonstrated their approach on four monogenic diseases. The authors then later approved upon this method [30] with various machine learning approaches including ridge regression, a Heat Kernel Diffusion Ranking [31], the Arnoldi algorithm [32] and an average based neighborhood ranking method. Using a functional annotation and a PPIN, the authors found that the Heat Kernel Diffusion Ranking method performed the best. The authors also

created PINTA: a web resource for candidate gene prioritization using a PPIN [33]. As input, PINTA requires expression data and the resource is freely available online.

Karni et al. [34] introduced an approach which combined the use of PPINs with gene expression data. Unique to the paper, instead of creating a set of gene-disease associations (the root set), the authors made use of the use of genes which change their expression level within a given affected tissue. These are then designated as the *disease-related* genes. The authors' gene prioritization algorithm relies on the assumption that in the disease state, the causal genes are disrupted which leads to expression level changes downstream in the signaling pathways of the expression network. In order to uncover the causal genes, the authors then attempt to find the smallest set of genes which could best explain the expression level changes in the genes. In simulations, the authors were able to show that their algorithm was very effective and outperformed a naïve algorithm which ranked disease associated genes based on their distances in the network. Essentially, these authors combined the use of pathway and expression knowledge in their prioritization algorithm.

2.2 KNOWLEDGE REPRESENTATION IN NETWORKS

This section describes the construction of knowledge networks and the types of representation used in such networks.

2.2.1 Network construction

Network-based methods first construct a knowledge network from data that is a graph with nodes and links which may be un-weighted or weighted. Typically, for gene prioritization, the knowledge network takes the form of a PPIN where the nodes represent proteins, and the links between the nodes represent protein-protein interactions. The node and link weights are derived from properties associated with the nodes and/or links and are deduced from domain knowledge. Hence, knowledge is incorporated into the network either through the nodes (as node weights) or through the links (as link weights). For example, a node weight may represent the number of biological pathways that the protein is involved in, and a link weight may represent the GO molecular functional similarity between the two interacting proteins. The final knowledge network thus consists of nodes, links, node weights, and link weights. A network-based inference algorithm is then applied to the knowledge network to rank a set of candidate nodes of interest to the user given a set of root nodes (those genes or proteins known to be associated with the disease). A candidate gene or protein that is ranked high asserts that it is topologically closely associated with the known disease genes or proteins, and this topological association is assumed to imply a high degree of association with the disease itself. It is important to note that in order for the prioritization algorithm to be considered a network-based inference algorithm, the algorithm *must* explicitly use a network-based property of the candidate node. An example of a network-based property would be the shortest distance to another node, the degree distribution, or the neighbors of a given candidate node.

Figure 2 provides the flow chart for network-based gene prioritization. A network is created from data which consists of binary interactions between proteins. An inference engine then queries the network to output a rank-ordering of nodes.

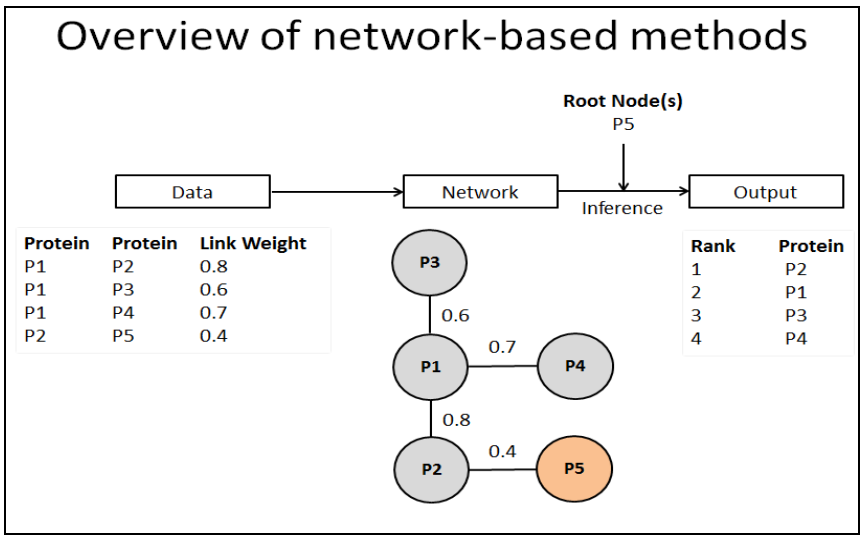

Figure 2. Overview of network-based gene prioritization.

The following two sections describe global versus local based inference algorithms and continuous versus binary inference algorithms.

2.2.2 Global versus local inference algorithms

A network-based inference algorithm can be either a global or local based inference algorithm. The PageRank with Priors algorithm is an example of a global based inference algorithm. In a global based inference algorithm, in order to determine whether a given gene is associated with a given disease, information from all other genes *can* used to determine whether the given gene is associated with the given disease. This is in contrast to a local based inference algorithm. In a local based inference algorithm, in order to determine whether a given gene is associated with a given disease, only local information around the gene or subset of proteins is utilized. For instance, in a PPIN, only those proteins which directly interact with the given protein of interest

are often utilized. In this case, the protein-protein interactions determine the locality around a given protein.

It has been shown in the literature that global based approaches are much more effective than local based approaches for gene prioritization. For instance, Kohler et al. [25] applied a random walk and diffusion kernel inference method for network-based gene prioritization – both of which capture global relationships with a network – and showed that these two methods are vastly superior to two other local based similarity measures. The two local based similarity measures were the direct interactions method of Oti et al. [20] and the single shortest path method of George et al. [12]. Both of these papers were discussed in the background section.

2.2.3 Continuous versus binary link weights

A network-based inference algorithm can utilize either binary link weights (a link weight is either 0 or 1) or continuous link weights (a link weight can take any value between 0 and 1). The approach utilized in this dissertation uses continuous link weights, which is contrary to what typically is done in the literature for network-based gene prioritization. For example, almost all authors – when creating a co-expression network – use a threshold on the correlation coefficient scores to create the co-expression network. In order for a link to exist between the two proteins, the coefficient score between the two proteins has to be above a given threshold. In other words, the author is essentially creating a binary network for the given knowledge source. However, when one does this, information in the link weights is essentially lost, because the continuous link weights are converted to a binary value. For example, in Figure 3, which represents a network with continuous link weights, node *A* has a much greater link weight with node *B* than node *C*: 0.99 versus 0.50 respectively. If this network was converted to a binary network with a

threshold of 0.5, both link weights would equivocally convert to a value of 1.0. Thus, the difference in link weights between the two values would be lost which would also subsequently result in a loss of information. This dissertation alleviates this concern by retaining the continuous link weights.

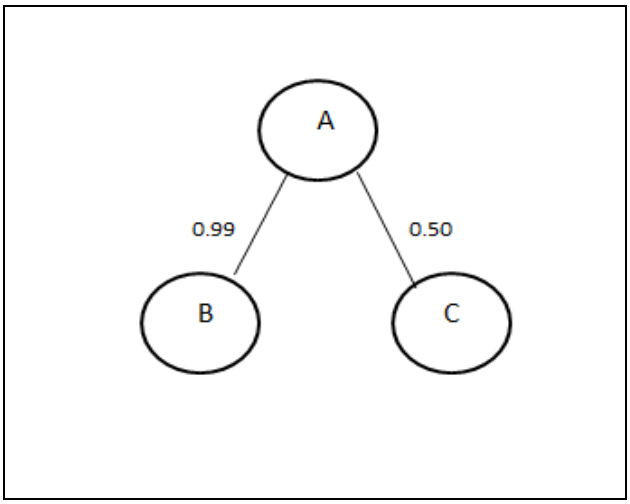

Figure 3. A network with continuous link weights.

2.3 INFERENCE IN NETWORKS

This section describes the common algorithms used in network-based inference including the PageRank and the PageRank with Priors algorithms.

2.3.1 The PageRank algorithm

The basic PageRank algorithm as described by Brin and Page [35] was developed to compute the importance of a webpage, and the importance is then used to rank the webpages. An intuitive description of the PageRank and the PageRank algorithm is provided followed by a formal description.

PageRank is a numeric value that represents the importance of a webpage on the Web. The intuition is that when a webpage links to another webpage, it can be considered as casting a vote for the other webpage. The importance of the webpage that is casting the vote determines how important the vote itself is. The more votes that are cast for a webpage, the more important the webpage is considered to be.

The PageRank algorithm computes the PageRank for each webpage on the Web. Imagine a web surfer who starts at a webpage and moves from one webpage to another by clicking on a hyperlink in a random fashion. The sequence of webpages visited by such a random web surfer is used to compute the PageRank of a webpage. After the random web surfer has visited a long sequence of webpages, the PageRank of a webpage is proportional to the number of visits to that webpage.

2.3.2 Random walk on a directed graph

More formally, the web is represented by a graph; the sequence of webpages visited by a random web surfer is called a random walk and is represented by a Markov chain model. The relative number of visits to a webpage is obtained by computing the stationary probability of the Markov chain for that webpage. These concepts are discussed in further detail below.

The Web is represented as a directed graph where the nodes are webpages and a directed link between two nodes represents the corresponding hyperlink from one webpage (node) to the other node (webpage).

A Markov chain is a stochastic model describing a sequence of events in which the probability of each event depends only on the state attained in the previous event. The sequence of events is called a random walk. In the context of the random surfer, a Markov chain describes a sequence of webpage visits in which the probability of visiting a webpage depends only on the webpage visited in the previous step.

A Markov chain consists of:

- A set S, the state space. The elements of S are called states and is represented by the set $\{v_1, v_2, \ldots, v_n\}$. A walk is a sequence of events $x_1, x_2, \ldots, x_t$ where $x_i \in S$ and the event x_i denotes the state at x_i.
- The walk is a sequence of events where each event corresponds to visiting a node v_i.
- A starting probability vector $Po^{(0)}$ where the ith element po_i is the probability the walk starts in state v_i.
- A transition probability matrix $\boldsymbol{Q}$ where the element at ith row and jth column is denoted by q_{ij}. All q_{ij} are ≥ 0 and $\sum_i q_{ij} = 1$. This matrix denotes that a walk which is now at v_i will be at v_j after the next step with probability q_{ij}.

The key property of the Markov chain is that the choice of the state for the next event in a walk only depends on the state in the current event and not on the states achieved in events previous to the current event.

2.3.3 The stationary distribution

Given the starting probability vector $Po^{(0)}$ and the transition probability matrix $\boldsymbol{Q}$, the probability vector after the first step is given by:

$$Po^{(1)} = \boldsymbol{Q} \times Po^{(0)} \quad (1)$$

where $Po^{(1)}$ is a probability vector where element po_i denotes the probability of the walk being at state v_i after the first step. This equation can be applied sequentially to generate a new probability vector at step i. Under certain conditions, a Markov chain will have a stationary probability distribution. This occurs if at some step $i+1$ the probability vector remains unchanged from step i for a given state. Typically, $Po^{(0)}$ is set to the uniform probability distribution.

Once the stationary probability distribution is reached, the *Po* vector does not change with further steps, and the Markov chain is said to have converged to a stationary distribution. The stationary probability distribution denotes the fraction of time that a surfer spends at any one node during the random walk and can be interpreted as the importance of the node relative to the other nodes in the network.

The basic PageRank algorithm uses an iterative algorithm to compute the stationary distribution from the prior probability vector and the transition probability matrix. For each iteration of the algorithm, a new probability vector is computed from the probability vector in the previous step and the transition matrix. The algorithm terminates when the change in the probability vector from one iteration to the next is below a specified tolerance threshold. The final probability vector is provided as output.

2.3.4 PageRank with Priors algorithm

White and Smyth [36] extended PageRank for estimating relative importance in networks to PageRank with Priors. Let *G* be a graph with a set of nodes *V* and a set of links *L*. Given *G*, a set of nodes *C* where $C \subseteq V$ and a set of root nodes *R* where $R \subseteq V$, the goal is to rank the nodes in *C* with respect to *R*. To do this, they compute a measure called node importance $I(c \mid R)$ for all $c \in C$ so that the largest values can be said to have the highest importance and conversely for the smallest values.

The authors defined a vector *Pr* of prior probabilities $Pr = \{pr_1, pr_2, ..., pr_{|R|}\}$ such that the probabilities sum to 1, and *Pr* represents the prior relative importance attached to node *v*. Specifically, they defined the prior as:

$$Pr_v = \frac{1}{|R|} \quad for\ v \in R$$

$$Pr_v = 0 \quad for\ v \notin R \tag{3}$$

where *R* is the set of root nodes. In this equation, all of the root nodes have equal prior probability. Thus, PageRank with Priors differs from PageRank in the prior probability vector. In PageRank, this vector is uniform over all the nodes; while for PageRank with Priors, this vector is uniform over root nodes and 0 for the non-root nodes. In addition, PageRank with Priors also defines a "back probability" β, $0 \le \beta \le 1$ which determines how often the algorithm jumps back to the set of root nodes. The iterative stationary probability equations for PageRank with Priors are of the form:

$$Po_v^{(i+1)} = (1-\beta)\left(\sum_{u=1}^{d_{in}(v)} p(v \mid u) \times Po_u^{(i)}\right) + \beta \times Pr_v \quad (4)$$

where, $Po_v^{(i+1)}$ denotes the probability attached to node *v* at the (*i*+1) iteration. When *Po* reaches the stationary distribution, the algorithm terminates. The relative importance is then obtained as $I(v \mid R) = Po_v$ after convergence; this relative importance is biased towards the set *R* due to the second term on the right hand side in Equation 4.

Intuitively, this equation represents a Markov chain for a random surfer who transitions "back" to the root set *R* with probability β at each time-step. This is similar in spirit to the use of weighted paths as follows: we are evaluating the probability of landing on a node in the modified Markov chain where a random graph surfer starts in the set *R* (with appropriate prior probabilities) and executes a random walk that ends stochastically with probability β (at which point the process restarts). This process defines an (infinite) set of walks of variable length starting at the root set (they 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.

Chen et al. [19] successfully applied the PageRank with Priors algorithm to the gene prioritization problem. The algorithm was applied to 19 different disease data sets and used a PPIN derived from OPHID as the knowledge network. An AUC of 0.8 was obtained. Perhaps most importantly, the authors showed that network-based methods used to study primarily social and web networks can be successfully applied to gene prioritization.

2.4 INCORPORATION OF KNOWLEDGE IN NETWORKS

This section provides details about papers from the literature which have previously incorporated various knowledge sources for the purpose of network-based gene prioritization.

The following papers integrate multiple data sources for the purpose of network-based gene prioritization. It is important to note that the primary objective in these papers was not to compare the performance of the individual knowledge sources against each other, and thus the methodology does not reflect this objective. In order to make this comprehensive comparison, the results from the various data sources need to be compared across *multiple* diseases on a disease-by-disease basis - none of the papers did this. Thus, the authors cannot make the general statement that one type of knowledge source is more useful than another one for the purpose of network-based gene prioritization. Furthermore, only the incorporation of link knowledge was investigated – not the incorporation of node knowledge as was done in this dissertation.

Frank et al. [37] created a PPIN derived from several true positive interaction data sources for the purpose of gene prioritization. Because the true positive data source only included a limited number of interactions between genes, the authors used some other data sources to predict interactions between the remaining gene pairs. These data sources include the GO, microarray expression measurements, predicted protein-protein interactions and true-positive (known) protein-protein interactions. A Bayesian classifier was used to predict the interactions, and this classification scheme could be used to combine knowledge from the aforementioned knowledge sources. For each network, a link weight represented the evidence of interaction for a gene pair, and this was learned from a Bayesian classifier. A Guassian kernel function was then applied to prioritize all the genes in a given locus which utilized the shortest distance between two given genes on the network. Of all the networks created, the network

which performed the best was the network with supplemented knowledge from the GO and microarray expression measurements. Using this network, the authors were able to detect at least one known disease gene in 54% of the diseases studied and this represented a 2.8-fold increase over random selection.

Chen et al. [38] created a network-based gene prioritization framework which can utilize several data sources including protein-protein interactions, expression data, and pathway data. For each data source, a separate *binary network* was created (a *binary network* is one where the link weights are all either one or zero). A variety of network-based inference algorithms were then used to prioritize genes for each network. In order to integrate the various knowledge sources, a novel data integration rank (DIR) score was produced. The DIR score assures that only the most informative networks – derived from the previously mentioned binary networks – will contribute to the final disease-gene relationship for a given candidate gene. A network is considered to be more informative if the disease genes in the network are more closely connected in terms of their binary interactions. The results showed that the DIR score improved when multiple data sources were utilized compared to a single data source. In conclusion, the authors showed that their approach out-performed two previous gene prioritization programs: ENDEAVOR and random walk with restarts.

The aforementioned papers all showed a benefit when multiple knowledge sources were utilized for the network-based gene prioritization process. The papers essentially incorporated the knowledge in the form of link weights, and the papers seemed to show the most benefit when the link knowledge from multiple data sources were combined.

3.0 KNOWLEDGE NETWORK GENE PRIORITIZATION ALGORITHM

This chapter describes the Knowledge Network Gene Prioritization (KNGP) algorithm. This algorithm was developed to overcome the main limitation of existing inference algorithms such as PageRank and PageRank with Priors algorithms. These algorithms can successfully incorporate link knowledge but not node knowledge. The KNGP algorithm includes both link knowledge and node knowledge for inference and generalizes the Page Rank with Priors algorithm.

3.1 OVERVIEW OF THE KNGP ALGORITHM

The KNGP algorithm creates a knowledge network from biological knowledge related to genes. The biological knowledge is represented in two ways: 1) knowledge related to a gene is represented as a **node weight** of the corresponding node (e.g., the number of articles in MEDLINE associated with a gene), and 2) knowledge related to a pair of genes is represented as a **link weight** (e.g., whether the products of a pair of genes interact) of the link connecting the corresponding nodes in the network. The algorithm computes a ranking for the nodes in the network relative to a set of genes already known to be associated with a disease of interest which is specified as the **root node set** in the network. Computing the ranking is called inference. More specifically, inference on the knowledge network outputs a number called the posterior node

importance for each gene in a set of genes of interest which is specified as the **candidate node set** in the network. The posterior node importance of a node is a measure of how likely the corresponding gene is to be associated with the disease of interest. The development of the KNGP algorithm was motivated and is based on the PageRank with Priors algorithm (see Section 2.3.4). The main advance of the KNGP algorithm is its ability to combine node weights representing node knowledge with the ability to specify if a node is a member of the root node set by modifying its node weight. In contrast, the PageRank with Priors algorithm uses node weights only to specify whether a node is a member of the root node set or not.

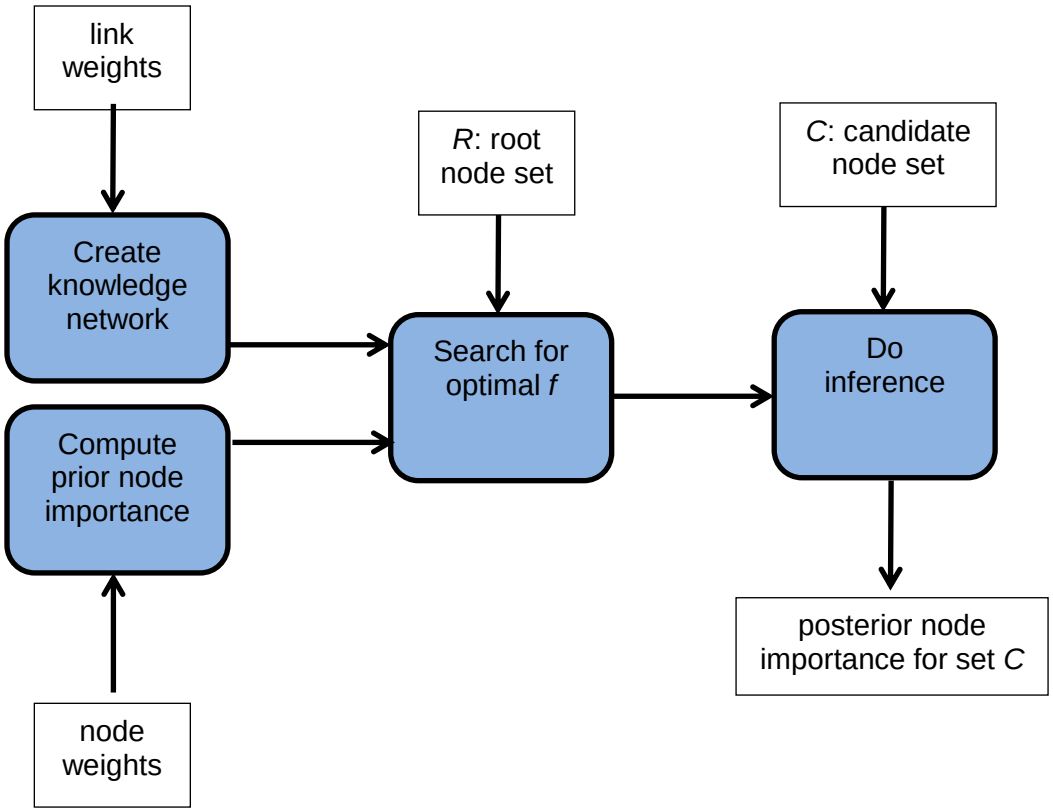

Figure 4. Components of the KGNP algorithm.

Figure 4 shows the components, inputs, and output of the KNGP algorithm. The four components of the algorithm include 1) creating the knowledge network, 2) computing the prior node importance, 3) searching for the optimal value of the parameter *f*, and 4) doing inference. The inputs include link weights, node weights, the set of root nodes *R* and the set of candidate nodes *C* and the output is the posterior node importance for each candidate node. The following sections describe the components of the KNGP algorithm in detail.

3.1.1 Create knowledge network

The knowledge network consists of a graph of nodes and links. The link knowledge is represented as link weights. There are two matrices that can be defined for the knowledge network: the link knowledge matrix and the transition probability matrix which is derived from the link knowledge network. The **link knowledge matrix** is a *n***n* matrix where *n* is the number of nodes in the knowledge network. An entry in the knowledge matrix represents the link weight between the nodes specified by the row number and the column number. For instance, an entry of 0.6 in the cell specified by row 1 and column 7 represents a link weight of 0.6 for the link between node number 1 and node number 7. Typically, the link weight takes a value between 0 and 1.0. D**e**tails of how the link weights can be obtained from knowledge sources are provided in Chapter 4. The **transition probability matrix** is a *n***n* matrix and is derived from the link knowledge matrix. An entry in this matrix gives the transition probability of going to one node (represented by the row number) from another node (represented by the column number) in the network. The transition probability of going to node *v* from node *u* is given by:

$$p(v \mid u) = \frac{lw(u,v)}{\sum_{i=1}^{neighbors(u)} lw(i,v)} \quad (5)$$

where $lw(u,v)$ is the link weight between node u and v obtained from the link knowledge matrix, and $neighbors(u)$ is the set of neighboring nodes for node u.

3.1.2 Compute prior node importance

The prior node importance of a node represents how likely the corresponding gene is to be associated with a given disease. The prior node importance is defined by two vectors: the node knowledge vector and the prior probability vector which is derived from the node knowledge vector. The **node knowledge vector** is a n dimensional vector where n is the number of nodes in the knowledge network. An entry in the vector represents the node weight associated with the corresponding node. For instance, the entry in position 7 represents the node weight for node number 7. Typically, the node weight takes a value between 0 and positive infinity. For example, for the MEDLINE knowledge source, the node weight may represent the number of articles associated with a given gene. Details on how the node weights may be obtained from a given knowledge source are provided in Chapter 4. The **prior probability vector** is also a n dimensional vector and is derived from the node knowledge vector. This vector contains the prior probabilities or prior importance for the nodes. I extend the prior probability vector used in the PageRank with Priors algorithm (see Section 2.3.5) for the KNGP algorithm to incorporate node knowledge into the network-based gene prioritization process. The prior probability vector is defined as:

$$Pr_v = \frac{fw_v}{\sum_{v \in R} fw_v + \sum_{v \notin R} w_v} \quad for\ v \in R$$

$$Pr_v = \frac{w_v}{\sum_{v \in R} f w_v + \sum_{v \notin R} w_v} \quad for\ v \notin R \tag{6}$$

where R is the set of root nodes, w_v is the node weight associated with the node v that is obtained from the node knowledge vector, and f is a parameter that takes a value between 0 and positive infinity. The next section describes the f parameter in more detail and how the optimal value of f is obtained.

3.1.3 Search for optimal *f*

After the knowledge network is created and the prior node importance is computed, the next step is to search for the optimal f value for a given knowledge network and root node set. In order to search for the optimal f value, a double cross validation methodology is utilized which consists of two loops: an inner and outer loop. The purpose of the outer loop is to iterate through all of the f values (defined by the user) and the purpose of the inner loop is to iterate through each of the root set members and produce an AUC for each individual f value from the outer loop. At the very end, the f value with the best AUC is returned. The optimal f value which is obtained depends on the relative distribution of the link and node weights between the root node and candidate node sets. Further explanation is provided in Section 5.1 which presents result about the behavior of the KNGP algorithm using synthetic data.

3.1.4 Do inference

After the optimal f value is determined, inference is performed to rank the candidate nodes which are the nodes that are of interest to the user. Inference produces a **posterior probability vector**

which is a *n* dimensional vector where *n* is the number of nodes in the network. The posterior probability vector represents the relative probability of a given node being associated with a disease after the node weights and link weights from the knowledge network are utilized. The posterior probability is taken to be the **relative importance** of the node with respect to the set of root nodes. The posterior probability vector is computed using the following iterative equation:

$$Po^{(i+1)} = (1-\beta)\times(\mathrm{Q}\times Po^{(i)}) + \beta \times Pr \qquad (7)$$

where *Pr* is the prior probability vector (an *n**1 dimensional vector), ***Q*** is the transitional probability matrix (an *n***n* matrix) and *Po* is the posterior probability vector (an *n**1 vector). At i=0, the *Po* vector is set to an *n* dimensional vector of all 0s. The term ß is a constant, $0 \leq$ ß ≤ 1, which represents how often the Markov process jumps back to the set of root nodes. At iteration (*i*+1), the *Po* is updated by multiplying the *Po* at iteration (*i*) with the matrix ***Q***. This equation is imputed for several iterations until the stationary distribution is reached. The stationary distribution occurs when the difference in the sum of the probabilities of *Po* at (*i*+1) and *Po* at (*i*) are less than some small constant.

The posterior probability vector includes a probability for every node in the network. After the stationary posterior probability vector is obtained, the KNGP algorithm ranks the candidate nodes and outputs them along with the posterior node importance (which is equal to the posterior probability). Often times, the candidate nodes will consist of all nodes in the network that are not in the root node set.

3.1.5 Illustrative example of inference

The following provides a simple example of the difference between the PageRank with Priors algorithm and the KGNP algorithm. In this example, the domain consist of 5 genes of which the

first two are root nodes associated with a fictitious disease of interest and the remaining three are candidate nodes that we want to prioritize.

The two algorithms differ in the specification of the prior probability vector *Pr* and in this example, they are specified as follows. For PageRank with Priors, the root set is defined to have uniform probability distribution and the non-root nodes have probabilities of 0. In this example, *Pr* = [0.5, 0.5, 0, 0, 0] which shows that the two root nodes have prior probabilities of 0.5 and the non-root nodes have probabilities of 0. For KNGP, the *Pr* entries are obtained by combining node weights (representing node knowledge) with information about whether a node belongs to the root node set or not. In this example, the node knowledge vector representing node knowledge is [20, 40, 20, 20, 40]. A key function of the KNGP algorithm is to identify the optimal *f* value for a given disease of interest. However, in this simple illustration, we assume that the optimal *f* = 2 and applying Equation 6 we obtain *Pr* = [0.2, 0.4, 0.1, 0.1, 0.2]. Note that the PageRank with Priors algorithm does not use the node knowledge vector.

Inference for both algorithms is done by applying Equation 7. The link weight matrix was the following: [[1.0,0.3,0.3,0.1,0.3], [0.3,1.0,0.6,0.2,0.3], [0.3,0.6,1.0,0.1,0.4], [0.1,0.2,0.1,1.0,0.2], [0.3,0.3,0.4,0.2,1.0]]. The transition probability matrix, **Q**, was derived from the link weight matrix and was the following: [[0.5,0.13,0.13,0.06,0.14], [0.15,0.42,0.25,0.12,0.14], [0.15,0.25,0.42,0.06,0.18], [0.05,0.08,0.04,0.59,0.09], [0.15,0.13,0.17,0.12,0.45]]. The *Pr* vector is specified as described in the preceding paragraph. The *Po* vector for iteration 0 is set to the 0 vector and the back-probability β is set to 0.5. The stationary distribution which is the vector *Po* in the final iteration is reached after 17 iterations in this example. The final *Po* for PageRank with Priors is [0.21,0.24,0.21,0.13,0.19] and for KNGP is [0.19,0.,24,0.22,0.15,0.21].

3.2 PSEUDOCODE

\

```
// Knowledge network gene prioritization (KNGP)
KNGP (knowledge, R)
input:   knowledge in the form of prior node importance and link weights
         R is a set of root nodes for disease of interest
output:  C is the set of candidate nodes with posterior node importance

network ← knowledge network created from knowledge
F ← set of f values
N ← set of nodes in network
best_f ← find_best_f (network, R, F)
prior ← compute prior node importance for all nodes in N using best_f
C ← inference (network, prior, R, N / R)          // N / R denotes set difference
return C


// search for optimal f
find_best_f (network, R, F)
input:   network is a knowledge network
         R is set of root nodes
         F is set of f values
output:  best_f which is the f that has highest AUC

N ← set of nodes in network
best_f = null
best_auc = -infinity
for each f in F:
    for each node i in R:
        mix i with 99 nodes drawn randomly from N / R to create set S
        prior ← compute prior node importance for all nodes in N using f
                posterior ← inference (network, prior, R, S)
        store posterior
    end for
    auc ← compute AUC from all posteriors
    if auc > best_auc:
        best_auc = auc
        best_f = f
    end if
end for
return best_f
```

Figure 5. Pseudocode for the KGNP algorithm**.**

The pseudocode for the KGNP algorithm is given in Figure 5. The top level procedure is called *KGNP* which takes as input link and node knowledge and *R* the set of root nodes. It outputs the posterior node importance for the candidate nodes *C*. The *find_best_f* procedure finds the best value of *f* from a set of possible values. It does so by performing inference on each *f* parameter

```
// do inference
inference (network, prior, R, S)
input:   network is a knowledge network
         prior is prior node importance for all nodes in network
         R is a set of root nodes
         S is a set of nodes for which to compute posterior node importance
output:  S with posterior node importance

Pr       ← prior          // prior probability vector
Q ← getTransProbMatrix(network) // transitional probability matrix
Po ← 0 // initialize posterior probability vector to 0

ß = 0.5
threshold = 0.00001
delta = 1
Po_prev = Po

do while delta > threshold:
    Po_curr ← (1- ß)*(Q*Po_prev) + ß*Pr
    delta = abs(Po_curr - Po_prev)
    Po_prev ← Po_curr
end while

return S from Po
```

Figure 5 (continued). Pseudocode for the KGNP algorithm.

value and then chooses the value that maximizes the AUC. The *inference* procedure implements Equation 7 and computes the posteriori probability from the prior probability vector and the transitional probability matrix. Note that the *inference* procedure is called for every value of *f* that is evaluated by the *find_best_f* procedure and called once by the *KGNP* procedure for the optimal value of *f*.

3.3 COMPUTATIONAL COMPLEXITY

This section provides an analysis of the time and space complexity of the KGNP algorithm.

The time complexity of the *inference* procedure is $O(n^2t)$ where n is the number of nodes in the network and t is the number of iterations needed to converge to the stationary distribution. In each iteration, the complexity is dominated by the multiplication of matrix Q which is of size $n*n$ with vector Po which is of size n; the complexity of this operation is $O(n^2)$. Since this operation is done t times, the overall complexity of the *inference* procedure is $O(n^2t)$.

The *find_best_f* procedure calls the inference procedure once for each value of f in F and each node i in R. Thus, the time complexity of *find_best_f* procedure is $O(n^2t|F||\mathrm{R}|)$ where $|F|$ is the number of values in F and |R| is the number of nodes in the root set.

The KNGP procedure calls the *find_best_f* and *inference* procedures once. Hence, the time complexity of the KNGP procedure is $O(n^2t|F||\mathrm{R}|) + O(n^2t)$. Since the first term in the sum dominates, the time complexity of the KNGP algorithm is $O(n^2t|F||\mathrm{R}|)$.

The space complexity comes mainly from the transitional probability matrix, the prior probability vector, and the posterior probability vectors. The complexity of storing the matrix and the vectors is $O(n^2) + O(n) + O(n)$. Since the first term in the sum dominates, the overall space complexity is $O(n^2)$.

In my experiments, n=17,631 which lead to a fairly large time and space requirement to execute the KGNP algorithm. Thus, running times for the experiments were fairly long.

4.0 EXPERIMENTAL METHODS

This chapter describes the experimental methods. Section 4.1 gives details of the knowledge sources used to create various knowledge networks. The knowledge networks that were created include networks created from a single link knowledge source, networks created from multiple link knowledge sources, networks created from a single node knowledge source, and networks crated from both link knowledge and node knowledge sources. Section 4.2 describes the creation of the root sets for the diseases used in the experiments. Section 4.3 gives details of the evaluation protocol.

4.1 CREATION OF KNOWLEDGE NETWORKS

The UniProt database provides a comprehensive catalog and annotation of all known proteins [39]. This annotation includes information such as the protein's name and description, the amino acid sequence, taxonomic data, cross-reference data, experimental data, and biological ontology information. The UniProt database has been utilized in several biological and bioinformatics research projects such as the study and structure of kinases [40], the construction of rule based models for cell signaling systems [41] and in the analysis of interactome networks [42].

I downloaded all 17,691 unique proteins from UniProt in March of 2011. For each knowledge source, I created a knowledge network whose nodes were the set of 17,691 proteins obtained from UniProt. For each pair of proteins (or corresponding genes), *u* and *v*, I calculated a link weight between proteins *u* and *v* that was specific to the knowledge source. The link weight ranged between 0 and 1 where 0 represents the notion that the corresponding proteins are dissimilar, and 1 represents the notion that the corresponding proteins are similar or interact with each other. The following sections define the link weights derived from various knowledge sources.

4.1.1 Knowledge networks created from a single link knowledge source

IID Link Weight Network: The Interologous Interaction Database (IID) contains 102,740 human, protein-protein, experimental interactions from a number of model organisms including human, mouse, rat and fly [43] [44]. The IID database extracts their interaction information from a variety of interaction databases such as the Human Protein Reference Database (HPRD) [45] and the Molecular Interaction Database (MINT) [46]. For the link weight, if an interaction was present in IID, I assigned a weight of 1.0 to the corresponding link. Otherwise, I assigned a weight of 0 to the corresponding link. Thus, for the IID Link Weight Network, a link weight was either 0 or 1. In total, this network resulted in 77,410 interactions between 10,487 proteins.

Species Link Weight Network: Every experimental interaction is derived from a related organism (e.g., yeast two-hybrid assay), and this data is available in the IID. In one of the earliest network-based gene prioritization papers, Chen et al. [21] used the species that a given protein-protein interaction was derived from as a knowledge source. For the link weight, generally

speaking, greater confidence was given to those experiments which were human based since I was trying to model the human cell condition. If the interaction came from a human based experiment, then the species link feature was assigned a value of 0.9. If the experiment came from a mammalian based experiment, then the value was 0.6, and if the experiment came from a non-mammalian based organism, the value was 0.3. If the species was unknown, a value of 0.0 was entered. A species value was calculated for each of the 77,410 interactions for the IID Network.

GO Molecular Function Link Weight Network: The Gene Ontology (GO) [47] is a set of controlled vocabularies which describes the functions of proteins within the cell. The ontology is constructed as a graph with nodes and edges where the nodes represent functional terms and the edges represent hierarchal relationships between the nodes. As one goes down the graph, the terms become more specific. For instance, at the very top, a very general functional term such as "Biological Process" may be defined. As a child term, more specific terms may be defined such as "Cell Proliferation" and under that may be "Muscle Cell Proliferation." The child-to-parent semantic relationships in the GO are defined as being "is-a" or "part-of". For instance, Cell Proliferation (child) "is-a" Biological Process (parent). provides a simple example of the gene ontology. Each protein within UniProt is associated with a set of GO Ontology Terms based on the protein's function, and this associated information is available from the Gene Ontology Consortium [47].

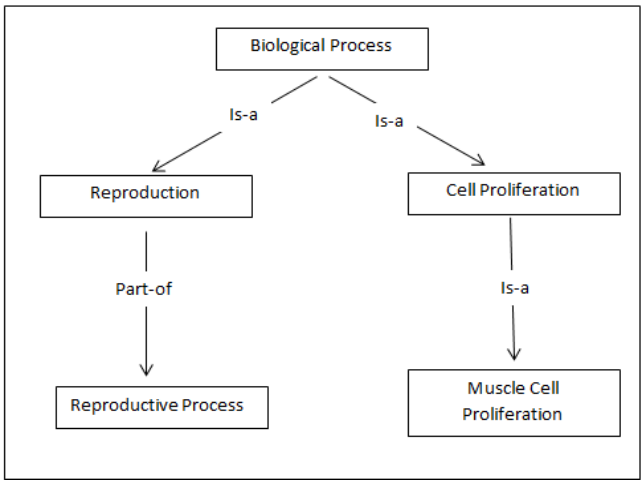

Figure 6. A simple example of the Gene Ontology.

The gene ontology is divided into three separate ontologies: Molecular Function, Biological Process, and Cellular Component. The subsequent sections on GO link weights describe each of these ontologies in more detail. This link weight deals with the GO Molecular Function ontology.

The GO Molecular Function ontology annotates single event activities which occur at the molecular level [47]. Two examples of activities are binding and transporter activities. I used the Wang et al. [48] similarity measure as the GO Molecular Function link weight. The similarity measure calculates the similarity between two sets of GO terms associated with the two genes, and the similarity measure ranges from 0 to 1 where 0 represents the smallest degree of molecular function similarity between two genes, and 1 represents the greatest degree of molecular function similarity between two genes. Most importantly, the measure takes into account the distance of the GO terms in the ontological graph from each other, and the depth of the GO terms. For instance, dealing with depth, two terms which are very close together and higher up in the ontology will get a smaller similarity score than if they were farther down in the ontology graph, because the terms with greater depth in the gene ontology are more specifically

defined. This is in contrast to the commonly used Jaccard similarity measure which just simply takes the intersection of the two sets of GO terms – GO_1 and GO_2 – and divides it by the union:

$$Jaccard(GO_1, GO_2) = \frac{|GO_1 \cap GO_2|}{|GO_1 \cup GO_2|} \quad (8)$$

This measure will give the same score to two sets of GO Ontology terms which are the same distance apart from a given parent node on the ontology regardless of their depth on the ontology graph. For example, in Figure 6, ontology terms *B* and *C* would receive the same similarity score as ontology terms *F* and *G* using the Jaccard similarity score, because they are the same distance apart from a given parent node on the ontology graph. However, the Wang similarity measure will correctly give a greater similarity score to ontology terms *F* and *G*, because they are farther down on the ontology graph and more specifically defined. The Wang similarity measure also utilizes the semantic relationships between the GO terms in the gene ontology.

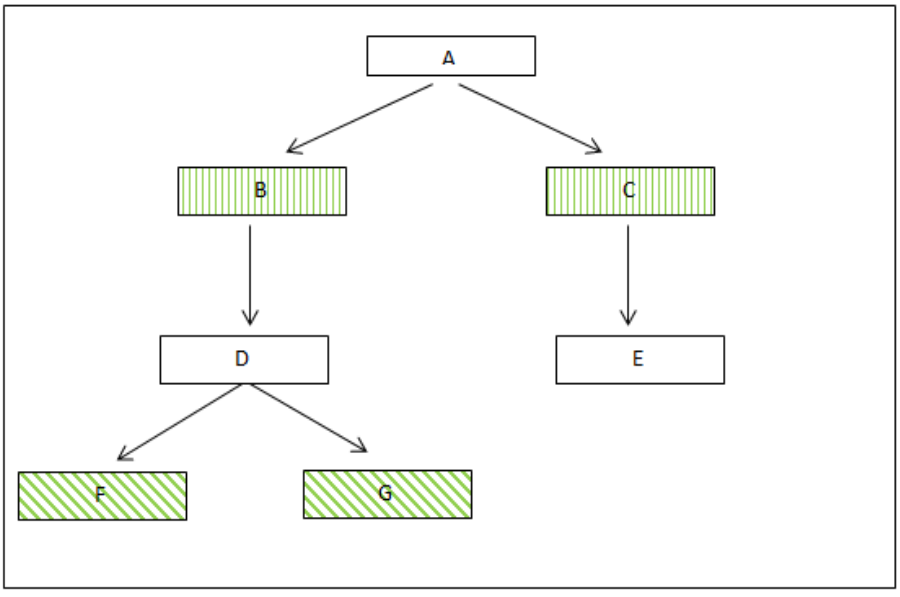

Figure 7. Two sets of ontology terms at different locations on the ontology graph.

In total, 100,997,587 GO Molecular similarity calculations were computed between 14,215 proteins.

GO Biological Process Link Weight Network: I computed the GO Biological Process link weight in the same manner as the GO Molecular Function link weight using the GO molecular function ontology. A biological process is a series of events carried out by assemblies of molecular functions. Two such examples are cellular localization and cell adhesion. In total, 88,411,753 GO Biological similarity calculations were computed among 13,297 proteins.

GO Cellular Component Link Weight Network: I computed the GO Cellular Component link weight in the same manner as the GO Molecular Function link weight using the GO Cellular Component ontology. A cellular component describes the location of a biological process. Two such examples are extracellular region and the organelle part. In total, 110,342,940 GO Cellular similarity calculations were computed between 14,855 proteins.

MEDLINE Link Weight Network: Currently, MEDLINE contains more than 18 million records from about 5,000 journals in a variety of health science fields. Each publication in MEDLINE has a unique MEDLINE identifier, and the Jaccard similarity measure between the two proteins was used as the MEDLINE Citations link weight. Given sets *A* and *B* which are the MEDLINE articles associated with the two proteins, the Jaccard similarity measure is computed as: $Jaccard(A,B) = |A \cap B| / |A \cup B|$. Thus, the Jaccard similarity measure counts the number of articles associated with both proteins and divides it by the total number of articles associated

with either of the proteins. I used NIH David [49, 50] to identify the MEDLINE articles associated with the two proteins. In total, 137,640,715 MEDLINE similarity calculations were calculated among 17,658 proteins.

Co-Expression Link Weight Network: The explosion of high-throughput technologies in recent years has elevated the study of genomics and proteomics. There are a variety of platforms currently available such as mass spectrometry and high throughput sequencing data. This network utilized high throughput microarray expression values from the Beer et al [51]. Only the 4206 proteins in the Beer data set were utilized instead of the full complement of proteins from UniProt. Furthermore, only the healthy samples were used – not the diseased samples – since I was trying to model a healthy human organism. For the link weight, the Spearman correlation between the two protein's expression profiles was calculated and the absolute value was taken. Thus, the link weight represented the degree of similarity between the expression profiles for the two sets of genes. In total, there were 8,847,321 co-expression calculations among 4,206 proteins.

4.1.2 Knowledge networks created from a combination of link knowledge sources

Predicted Protein-Protein Interaction (PPPI) Link Weight Network: The protein-protein interactions used in the IID Network represent experimentally derived interactions. I postulated that the addition of predicted protein-protein interactions may improve performance. I obtained predicted protein-protein interactions from the human protein-protein interaction (HPPI) database and created the PPI network from the union of experimental interactions and predicted interactions [52, 53] . The HPPI database contains over 79,000 predicted interactions and has

little overlap with the IID database. If the interaction was present within the IID or the HPPI, a link weight of 1.0 was entered. Otherwise, a value of 0.0 was entered. Essentially, this network turns some of the zeros in the IID Network to ones. In total, this network resulted in 126,668 interactions between 11,259 proteins. Thus, there were a total of 49,258 interactions added from the IID Network to the PPI Network. In other words, about 49,000 of the 0s in the IID Network changed to 1s in the PPI Network. This network combined knowledge from the HPPI and IID knowledge sources.

PPPI + Gene Ontology Molecular Function (GOM) Link Weight Network: I speculated that combining predicted protein-protein interaction knowledge with the GO knowledge source should lead to better gene prioritization performance. This network utilized the GO Molecular Function Ontology. The PPI+GOM link weight was calculated as follows. If the PPI link weight as computed in the PPI Link Weight Network was 0, I assigned a value of 0 as the link weight. Otherwise, I assigned the value from the GO Molecular Function Link Weight Network. This network combined information from the HPPI, PPI, and GOM knowledge sources.

PPPI + Gene Ontology Biological Process (GOB) Link Weight Network: This link weight was computed in the same manner as the PPI+GOM Link Weight Network except the GO Biological Process Ontology was utilized. This network combined information from the HPPI, PPI, and GOB knowledge sources.

PPPI + Gene Ontology Cellular Component (GOC) Link Weight Network: This link weight was computed in the same manner as the PPPI+GOM Link Weight Network except the GO Cellular

Component Ontology was utilized. This network combined information from the HPPI, PPI, and GOC knowledge sources.

4.1.3 Knowledge networks created from node knowledge sources

InterPro Node Weight Network: InterPro [54] is an integrated database of protein signatures used for the classification and annotation of proteins and genomes. Among the types of annotations annotated by InterPro are the functional domains for a given protein. This node weight represented the number of InterPro domains associated with a given protein.

GO Node Weight Network: As previously mentioned, the GO is a set of controlled vocabularies which describes the functions of proteins within the cell as previously described. This node weight represented the number of gene ontology associations for a given protein. The ontology associations were summed across all of three different types of gene ontologies previously mentioned: cellular, molecular, and functional.

4.1.4 Knowledge network from node and link knowledge sources

PPPI+GOC Link Weight and GO Node Weight Network: This network represented the incorporation of both link and node knowledge. The link weights were the same as those used in the PPPI+GOC Link Weight Network, and the node weights were the same as those used in the GO Node Weight Network.

4.1.5 Overview of knowledge networks

Table 1 gives details of the node and link knowledge used in the networks for Aims 1 through 4. The subsequent sections explain how the knowledge was incorporated into the knowledge networks, and why the individual knowledge sources were added.

Table 1. Networks associated with each aim.

Aims	**Networks**
Aim 1: Link knowledge	IID Link Weight Network, Species Link Weight Network, GO Molecular Function Link Weight Network, GO Biological Process Link Weight Network, GO Cellular Function Link Weight Network, MEDLINE Link Weight Network, Co-Expression Link Weight Network
Aim 2: Combination of link knowledge	Predicted Protein-Protein Interaction (PPPI) Link Weight Network, PPPI+GOM Link Weight Network, PPPI+GOB Link Weight Network, PPPI+GOC Link Weight Network
Aim 3: Node knowledge	InterPro Node Weight Network, GO Node Weight Network
Aim 4: Node and link knowledge	PPPI+GOC Link Weight and GO Node Weight Network

4.1.5.1 How knowledge is represented in the network

As previously mentioned, the knowledge is incorporated into the network through link and node weights. The knowledge is being represented through the link weights via the calculation of similarity scores between protein pairs. This approach makes sense, since in a network, a high weight traditionally means a greater degree of similarity between two nodes than a low weight score, and this is the assumption that the network inference algorithm works on. Furthermore, there is backing for this link weight approach in the literature. Sharma et al. [55] created a gene prioritization approach which integrates weights in a network similar to the type of weights used in this dissertation. The new approach was able to enrich the candidate list for type 2 diabetes by 6.8 fold.

The knowledge in the network is also represented through node weights by counting the number of associations (GO, and InterPro) for a given protein. For network-based inference algorithms, the node weight represents the prior node importance, and a greater node weight means the corresponding protein is assigned greater prior importance. This makes sense, since proteins which are studied more will have more associations, and should thus be more likely to be associated with a disease.

It is interesting to note that Navlakha et al. [56] in their comparison of various network-based inference based algorithms found that disease related proteins which were spread far apart on the PPIN, the inference algorithms suffered with low precision and recall. The authors showed that predictions made for more homophilic diseases (*homphily* represents the closeness of a set of nodes on a graph) were of much greater quality than those that had less homphily. In order to alleviate this concern, *the authors suggested the use of different knowledge sources and for more computational efforts to be directed in this area*. The various networks created in this dissertation – formed from the disparate knowledge sources – are an effort in this direction.

4.1.5.2 An explanation for the various knowledge sources

For network-based gene identification, if two interacting genes share a feature in common and would also be more likely to share a similar disease in common, then it would make sense to add the feature. In other words, given that genes *G1* and *G2* interact, if the two genes share a similar feature *F* in common, then the two genes would also be more likely to share a similar disease *D* in common if it's known that one of those two genes is already associated with the disease.

It has been known in the literature for some time that diseased genes tend to share functional characteristics in common, and several gene prioritization papers in the literature reflect this. In one of the earliest papers, Freudberg et al. [8] created clusters of diseases and their

respective causative genes, and scored potential disease genes according to their functional similarity to genes in the clusters. Furthermore, Jimenez-Sanchez et al. [57] found a strong correlation between gene function and certain disease features such as age of onset and mode of inheritance. Thus, including the GO Ontology as a link feature should add a significant benefit for network-based gene identification.

It has also been shown in the literature that diseased genes also tend to share similar co-expression patterns. For instance, Alu et al. [28] showed systematically that the integration of expression profiles in human and mouse – in conjunction with a phenotype similarity map – allowed for the identification of disease genes in very large genotypic regions. Oti et al. [58] had a similar result. The authors showed that evolutionary conserved co-expression patterns can be used to prioritize candidate genes effectively. Interestingly, the authors also showed that co-expression across multiple species (fly, rat, yeast, etc.) are a better predictor of candidate disease genes than using just human alone.

It would also seem to make sense to add the species and MEDLINE knowledge sources for the link weights. Chen et al. [21] successfully utilized the species link knowledge source in one of the earliest known network-based gene prioritization papers. The authors decided to utilize the species knowledge source in place of using just protein-protein interactions, so this evidentially would imply that the authors thought that the species knowledge source was more useful than using just protein-protein interactions. Furthermore, it would also seem to make sense to utilize the MEDLINE knowledge source. Since protein-disease associations are reported in MEDLINE, if two proteins have similar literature trails, they may be implicated in the same or at least similar diseases. There is backing for the use of MEDLINE articles for gene prioritization in the literature. For instance, Hritizvoski et al. [59] created a gene prioritization system called

BITOLA which was almost totally dependent on the use of MEDLINE articles. The system attempted to discover new relations between a given starting concept of interest (disease) and other concept (i.e., disease-related gene) by automatically mining MEDLINE. The authors showed that BITOLA could successfully be used for the purpose of gene prioritization. However, in general, I was not as confident about the inclusion of the MEDLINE knowledge source, because a given protein can be associated with a MEDLINE article for a variety of reasons and may not even be related to the main topic of interest for the article.

The predicted protein-protein interaction knowledge source should also add a benefit. Many gene prioritization papers have utilized predicted protein-protein interactions with success, and it would thus seem to make sense that adding predicted protein-protein interactions should provide a benefit. It is interesting to note that Franke et al. [37] employed the use of predicted protein-protein interactions in his paper with success. The predicted interactions were derived from microarray measurements and the gene ontology.

I also considered additional knowledge sources that were not included in this dissertation. These sources included the KEGG Pathway knowledge source – which provides a comprehensive catalog of biological pathways for every gene – and PROSITE – which provides a comprehensive database of protein domains, families and functional sites. However, neither of these knowledge sources provided the sufficient annotation coverage for the full list of 17,691 unique proteins downloaded from UniProt. Both of these knowledge sources had annotation coverage of less than 50%. Given the small annotation coverage, it is not possible to provide a fair comparative evaluation for these knowledge sources against the other knowledge sources.

Even though several knowledge sources could not be included due to insufficient annotation coverage, it is important to note that the predicted interactions from the human

protein-protein interaction (HPPI) database combined information from several knowledge sources including gene co-expression, orthology, co-occurrence of domains, post-translational modifications, co-localization of the proteins within the cell and analysis of the local topology of the predicted PPIN. The authors used a naïve Bayes model in the fashion of Scott and Barton [53] to predict the probability of a given pair of proteins interacting. First, probabilities were obtained for each disparate knowledge source, and then the probabilities were combined to give an overall likelihood of interaction for each pair of proteins. Thus, even though several knowledge sources could not be included in this dissertation directly because of insufficient annotation coverage, they were included indirectly through the predicted protein-protein interactions.

4.2 CREATION OF ROOT NODE SETS

The root nodes consisted of proteins known to be associated with the disease of interest. For my experiments, I chose 19 diseases and created a set of genes for each disease that are known to be associated with that disease. I call such a set as a *root set* for the disease of interest. I obtained the root sets for the 19 experimental diseases from the Gene Association to Disease (GAD) database. The GAD contains both positive and negative gene-disease associations. A positive association asserts that the protein is associated with the disease of interest and a negative association asserts that the protein is not associated with the disease of interest. I selected 19 experimental diseases such that each disease had a root set of 5 or more genes. For a gene to be eligible to be included in the root set, the gene had to have two more positive associations than

negative associations with the respective disease in the GAD. Table 2 provides some statistics for the genes extracted from the GAD.

Table 2. Statistics for the genes extracted from the GAD.

Total # of genes with at least one positive disease association in the GAD	3562
Total # of genes with at least one positive association for the 19 experimental diseases	845
Total # of genes with two more positive than negative associations for the 19 experimental diseases	229
Total # of genes with two more positive than negative associations for the 19 experimental diseases which are associated with more than one of the 19 diseases	58

Table 3 provides the list of 19 experimental diseases and number of genes associated with each disease. Appendix A provides a list of all the root set genes (using UniPort identifiers) associated with each disease.

Table 3. Number of genes known to associated with each of the 19 experimental diseases.

Disease	**Number of genes**
Rheumatoid Arthritis	24
Parkinson's Disease	21
Celiac Disease	16
Esophageal Cancer	8
Hepatitis C	8
Crohn's Disease	17
Breast Cancer	27
Asthma	29
Alzheimer's Disease	21
Ulcerative Colitis	24
Endometriosis	5
Lymphoma	7
Osteoarthritis	8
Epilepsy	6
Atherosclerosis	43
Pancreatitis	6
Cirrhosis	7
Myocardial Infarction	32
Tuberculosis	12

4.3 EVALUATION

This section describes the evaluation protocol used. The evaluation protocol is shown in Figure 8. In the protocol, it is important to note that for aims 1 and 2, there is no search for the optimal *f* parameter value in the KGNP algorithm since in these aims node knowledge is not used. Hence, for these two aims, the *find_best_f* procedure from the pseudocode (Section 3.2) was not executed.

1. Iterate through each of the proteins in the root set in a leave-one-out cross validation manner.
2. Take the node in step 1 and mix it with 99 other nodes randomly chosen from the set of non-root nodes. Call the node that was selected as the left-out node.
3. Using the full set of root nodes (excluding the left-out node), apply the KNGP algorithm to the network and rank order the 100 nodes selected in Step 2. Compute the AUC. (In Aims 1 and 2, do not search for optimal *f* parameter value).
4. Repeat Steps 2 and 3 for each of the nodes chosen in Step 1.
5. Repeat steps 1 through 4 for a total of 10 times

Figure 8. Evaluation protocol.

The protocol generates a total of *m**10 (where *m* is the size of the root node set) rank ordered lists of 100 nodes each with a left-out node that is embedded in 99 non-root nodes. A threshold rank (for example, the 5^{th} rank) for such a list separates those nodes that are ranked above it from those that are ranked below it. For a given threshold rank, sensitivity is defined as the percentage of lists where the left-out node was ranked above the threshold and specificity as the percentage of lists where the left-out node was ranked below the threshold. Varying the threshold rank produced a series of sensitivity and specificity values from which a ROC curve was constructed, and the corresponding AUC was calculated.

4.3.1 Link and node weights

For a given knowledge source, the extracted link weights were represented as the link knowledge matrix. For instance, if the GO Molecular Function is the knowledge source, then the link weight represented the GO Molecular Function similarity between two proteins. For the node weight networks (Aim 3), the link weights were the same as the IID link weight network.

The node weights derived from a knowledge source were represented as the node knowledge vector. For aims 3 and 4, the node weight was assigned the value from the respective node weight knowledge source. For instance, for the GO knowledge source, the node weight for a node represents the number of GO tersm associated with the corresponding protein. Aims 1 and 2 do not utilize the node knowledge vector and thus are not initially assigned node weights. Rather, they use the same prior weights as the Page Rank with Priors algorithm.

4.3.2 Wilcoxon paired-samples signed-rank test

The Wilcoxon paired-samples signed-rank test was used for comparing the performance of the knowledge sources. This test is a nonparametric procedure used to test whether there is sufficient evidence that the median of two probability distributions are significantly different [60]. In evaluating knowledge sources, it can be used to test whether two knowledge sources differ significantly in performance on a specified measure such as the AUC. The IID Network was used as the baseline since it is commonly used in the literature and does not represent the incorporation of any new knowledge.

5.0 EXPERIMENTAL RESULTS

This chapter provides experimental results and also discusses the results. Section 5.1 provides results from synthetic data experiments to characterize the behavior of the *f* parameter in the KNGP algorithm. Section 5.2 describes the results for the experimental diseases using real knowledge networks. These include results from single link knowledge networks, combined link knowledge networks, node knowledge networks, and combined link and node knowledge networks.

5.1 RESULTS OF SYNTHETIC DATA EXPERIMENTS

This section describes synthetic data experiments and the results from them that I conducted to explore the behavior of the KNGP algorithm. My goal was to examine how the node weights interacted with the link weights to influence the AUCs at different *f* values in the KGNP algorithm. The goal was to understand under which circumstances the node weights are more important than the link weights, link weights are more important than the node weights and links weights and node weights are roughly equally important in determining the AUCs.

A set of synthetic datasets were created as follows. Each dataset contained 1000 nodes of which nodes 1 to 100 are designated as root nodes and the remaining nodes are designated as

candidate nodes (or non-root nodes). To assign node weights and link weights, the 1000 nodes were partitioned into the following 5 groups (see Table 11):

- Group 1 consisted of root nodes 1 through 50
- Group 2 consisted of root nodes 51 through 10
- Group 3 consisted of candidate nodes 101 through 150
- Group 4 consisted of candidate nodes 151 through 200
- Group 5 consisted of candidate nodes 201 through 1000

Four datasets were generated in the following manner:

- In dataset 1, each of the 1000 nodes was assigned a random node weight between 0 and 1. Thus, root nodes and candidate nodes had similar node weights. The links among the root nodes (i.e., node groups 1 and 2) were assigned a random weight between 0.5 and 1 and the links among the candidate nodes and among the root nodes and the candidate nodes were assigned a random weight between 0 and 0.5. Thus, links among root nodes had higher weights than other links.
- In dataset 2, the root nodes (i.e., groups 1 and 2) were assigned a random node weight between 0.5 and 1, and the candidate nodes (i.e., groups 3, 4 and 5) were assigned a random node weight between 0 and 0.5. Thus, root nodes had higher node weights than all of the candidate nodes. All links were assigned a random link weight between 0 and 1. Thus, links among root nodes, links among candidate nodes and links among root nodes and candidate nodes had similar weights.
- In dataset 3, the root nodes were assigned a random weight between 0.9 and 1.0, and the candidate nodes were assigned a random weight between 0.5 and 1.0. Thus, the root nodes, on average, had higher node weights than the candidate nodes, but some of the

candidate nodes could have had greater node weights. The link weights between the root nodes were assigned a value between 0.55 and 1.0, and the link weights between the candidate nodes were assigned a value between 0.5 and 1.0. Thus, the links between the root nodes were, on average, were higher than the link weights between the candidate nodes, but some of the candidate node link weights could have been higher.

- In dataset 4, the root nodes were assigned a random node weight between 0.95 and 1.0, and the candidate nodes were assigned a random node weight between 0 and 1.0. Thus, the root nodes, on average, had higher node weights than the candidate nodes, but some of the candidate nodes could have had greater node weights. The link weights between the root nodes were assigned a value between 0.1 and 1.0, and the link weights between the candidate nodes were assigned a value between 0 and 1.0. Thus, the links between the root nodes were, on average, higher than the link weights between the candidate nodes, but some of the candidate node link weights could have been higher.

For each of the 4 datasets, the KNGP algorithm was run using the evaluation protocol for a range of *f* parameter values. The *f* parameter values tested were the following: 0, 1, 15, 100, 10,000 and 1 trillion (which represents infinity). At *f*=0, the prior probability of the root nodes becomes 0.0, and at *f*=1 trillion, the prior probability of the root nodes approach infinity, and the prior probability for the candidate nodes approaches 0.0. Table 4 provides the link weights utilized for each dataset. Table 5 and Table 6 provide the link weights for each individual group and between the groups respectively. Table 7 provides the AUCs for each data set. The highest AUC is in bold font.

Table 4. Specification of node weights for each group.

Dataset	**Node Weights**				
	Group 1	Group 2	Group 3	Group 4	Group 5
1	rand(0,1)	rand(0,1)	rand(0,1)	rand(0,1)	rand(0,1)
2	**rand(0.5,1)**	**rand(0.5**	rand(0,0.5	rand(0,0.5)	rand(0,0.5)
3	**rand(0.9,1)**	**rand(0.9,1)**	rand(0.5,1)	rand(0.5,1)	rand(0.5,1)
4	**rand(0.95,1)**	**rand(0.95,1)**	rand(0,1)	rand(0,1)	rand(0,1)

Table 5. Specification of link weights for each group.

Dataset	**Link Weights**				
	Group 1	Group 2	Group 3	Group 4	Group 5
1	**rand(0.5,1)**	**rand(0.5,1)**	rand(0,0.5)	rand(0,0.5)	rand(0,0.5)
2	rand(0,1)	rand(0,1)	rand(0,1)	rand(0,1)	rand(0,1)
3	**rand(0.55,1.0)**	**rand(0.55,1.0)**	rand(0.5,1)	rand(0.5,1)	rand(0.5,1)
4	**rand(0.1,1.0)**	**rand(0.1,1.0)**	rand(0,1)	rand(0,1)	rand(0,1)

Table 6. Specification of link weights between groups.

Dataset	**Link Weights**									
	Group 1-2	Group 1-3	Group 1-4	Group 1-5	Group 2-3	Group 2-4	Group 2-5	Group 3-4	Group 3-5	Group 4-5
1	**rand(0.5,1)**	rand(0,0.5)	rand(0,0.5)	rand(0,0.5)	rand(0,0.5)	rand(0,0.5)	rand(0,0.5)	rand(0,0.5)	rand(0,0.5)	rand(0,0.5)
2	rand(0,1)	rand(0,1)	rand(0,1)	rand(0,1)	rand(0,1)	rand(0,1)	rand(0,1)	rand(0,1)	rand(0,1)	rand(0,1)
3	**rand(0.55,1.0)**	rand(0.5,1)	rand(0.5,1)	rand(0.5,1)	rand(0.5,1)	rand(0.5,1)	rand(0.5,1)	rand(0.5,1)	rand(0.5,1)	rand(0.5,1)
4	**rand(0.1,1.0)**	**rand(0.1,1.0)**	rand(0,1)	rand(0,1)	rand(0,1)	rand(0,1)	rand(0,1)	rand(0,1)	rand(0,1)	rand(0,1)

Table 7. AUCs for each dataset.

Dataset	***f*=0**	***f*=1**	***f*=15**	***f*=100**	***f*=10,000**	***f*=INF**
1	0.602	0.651	0.991	**1.000**	**1.000**	**1.000**
2	**1.000**	0.999	0.996	0.877	0.467	0.461
3	0.898	0.901	0.941	**0.977**	0.924	0.922
4	0.974	0.978	**0.991**	0.975	0.897	0.895

As Table 7 shows, the optimal *f* value (i.e., the *f* value that achieved the highest AUC) depends on the degree to which the link and node weights are biased towards the root nodes versus the non-root nodes. In this context, the bias indicated how much greater the node or link

weights were for the root nodes versus the non-root nodes. If the link weights were considerably more biased towards the root nodes than the non-root nodes – as in dataset 1 – than the highest AUC was obtained at the largest *f* value. Conversely, if the node weights were considerably more biased towards the root nodes than the non-root nodes – as in dataset 2 – than the highest AUC was obtained at the smallest *f* value. When the bias towards the root nodes was more balanced between the node weights and link weights – as in datasets 3 and 4 – than the highest AUC was obtained at a *f* value between the two extremes.

These synthetic experiments provide some intuition for the *f* parameter in the KGNP algorithm. The *f* parameter represents the tradeoff in the importance between the link weights and the node weights in determining the relative importance of nodes. If the optimal *f* value is high then it implies that the link weights dominate over the node weights in determining the relative importance. In other words, the connectivity of the root nodes according to the link weights matters a great deal, and the node weights contribute little – if any – benefit at this extreme. Conversely, if the optimal *f* value is low then it implies that the node weights dominate over the link weights in determining the relative importance. In other words, the connectivity of the link weights between the root nodes matters very little – if at all – and network-based gene prioritization is thus not useful at this extreme, because the network itself (characterized by the links) is not being utilized. These results imply that in order for node knowledge to contribute to determining the relative importance, the optimal *f* value *must* occur between the two extremes.

5.2 RESULTS OF DISEASE DATA EXPERIMENTS

This section provides results from the application of the KNGP algorithm to 19 diseases on a variety of link knowledge networks, node knowledge networks and combined link knowledge and node knowledge networks. First, I describe the results from the incorporation of link knowledge from single data sources (Aim 1); second, I describe the results from the incorporation of link knowledge from multiple data sources (Aim 2); third, I describe the incorporation of node knowledge (Aim 3); and fourth, I describe the incorporation of both node and link knowledge together (Aim 4).

5.2.1 Incorporation of single link knowledge source

Table 8 shows the AUCs for each network and knowledge source based on the incorporation of link knowledge from a single knowledge source. The performance of the IID Network was used as the baseline. The last row in the table provides the average AUC obtained by averaging the AUCs for the 19 diseases.

Table 8. AUCs for networks using single link knowledge.

Disease	**IID**	**Species**	**GO Molecular**	**GO Biological**	**GO Cellular**	**MEDLINE**	**Expression**
Rheumatoid Arthritis	0.699	0.699	0.597	0.766	0.600	0.522	0.592
Parkinson's Disease	0.631	0.639	0.572	0.736	0.582	0.224	0.552
Celiac Disease	0.772	0.774	0.662	0.792	0.606	0.413	0.716
Esophageal Cancer	0.857	0.842	0.742	0.870	0.693	0.734	0.673
Hepatitis C	0.721	0.731	0.440	0.810	0.437	0.300	0.468
Crohn’s Disease	0.814	0.812	0.619	0.719	0.615	0.443	0.464
Breast Cancer	0.834	0.839	0.702	0.782	0.561	0.651	0.502
Asthma	0.774	0.778	0.595	0.726	0.649	0.500	0.539
Alzheimer’s Disease	0.835	0.838	0.575	0.685	0.638	0.256	0.497
Ulcerative Colitis	0.672	0.672	0.550	0.676	0.610	0.482	0.480
Endometriosis	0.772	0.776	0.486	0.856	0.595	0.663	0.677

Lymphoma	0.830	0.828	0.630	0.880	0.521	0.724	0.352
Osteoarthritis	0.753	0.755	0.880	0.686	0.836	0.475	0.473
Epilepsy	0.578	0.579	0.611	0.794	0.634	0.225	0.921
Atherosclerosis	0.798	0.800	0.638	0.817	0.764	0.453	0.457
Pancreatitis	0.767	0.780	0.537	0.548	0.352	0.637	0.433
Cirrhosis	0.600	0.564	0.368	0.474	0.600	0.448	0.165
Myocardial Infarction	0.865	0.867	0.662	0.770	0.748	0.448	0.379
Tuberculosis	0.664	0.672	0.662	0.851	0.604	0.745	0.583
Average	**0.747**	**0.750**	**0.600**	**0.749**	**0.613**	**0.492**	**0.522**
p-value	ref.	<0.90	<0.99	<0.90	<0.99	<0.99	<0.99

Among the networks constructed from a single knowledge source, the GO Biological Network had the highest average AUC but its performance was not statistically significantly better than the IID knowledge source based on the Wilcoxon paired-samples signed-rank test. Furthermore, none of the networks based on the other types of knowledge sources – including the GO Cellular and GO Component Networks – did significantly better than the IID Network. Overall, these results suggest that gene functional information, MEDLINE information, species information, and co-expression knowledge – by itself – are *not more useful* for network based gene prioritization than protein-protein interaction knowledge.

The result for the GO Networks was somewhat surprising since it has been reported in previous publications that disease genes tend to share a high degree of functional similarity. In one of the earliest papers, Freudberg et al. [8] created clusters of diseases based on their respective causative genes, and scored potential disease genes according to their functional similarity to genes in the clusters. Furthermore, Jimenez-Sanchez et al. [57] found a strong correlation between gene function and certain disease features such as age of onset and mode of inheritance. Both these papers suggest that disease genes tend to share common functionality. Given this literature trail, one would think that including the GO may add a significant benefit for network-based gene identification, but compared to using just protein-protein interactions, it did not.

It was also surprising that the Co-Expression Link Weight Network did not perform significantly better than the IID Network. Ala [28] showed that genes involved in similar diseases tend to share the same expression pattern, and given this observation, one may think that the Co-Expression Link Weight Network would add a significant benefit, but it did not.

And lastly, it was not too surprising that the Species Network did not perform significantly better than the IID Network. It is known that interactions from several highly related species like fly and yeast tend to be very similar to human protein-protein interactions. Thus, by just simply giving more confidence to human protein-protein interactions, I did not expect a significant benefit, but it was still worth looking at.

5.2.1.1 Topological explanation for AUCs

One interesting question was why the AUCs in Table 8 for one type of disease and link weight network were greater than another. For instance, why was the AUC using the GO Cellular Component Network (0.856) significantly greater than the AUC using the GO Biological Process Network (0.595) for endometriosis? I provide some explanations based on the network topology in the following sections.

Node strength of root nodes and relationship to relative importance

The degree of a node *v* is defined as the number of links that *v* has to other nodes in the network. In a weighted network (that has weighted links), the **node strength** of a node *v* is obtained by summing the weights on the links that *u* has to other root nodes and dividing it by the number of root nodes:

$$str(v) = \frac{\sum_{u \in R} lw(v,u)}{|R|} \qquad (9)$$

where R is the set of root nodes and $lw(v, u)$ is the link weight between nodes v and u.

I conjectured that the greater the relative importance (or posterior probability, see Section 3.1.4) assigned by the KNGP algorithm to a root node the larger its node strength. Figure 9, Figure 10 and Figure 11 plot the relative importance versus the node strength for the root nodes of rheumatoid arthritis for the following knowledge networks: the GO Molecular Function Network, the GO Biological Function Network, and the GO Cellular Component Network.

In all three plots, as the node strength increased, the relative importance also increased, and this correlation was significant for all three networks (p-value < 0.01). Similar results were observed for diseases other than rheumatoid arthritis (data not shown). Thus, a node that has high link weights to other nodes in the root set tends to obtain a higher relative importance.

.

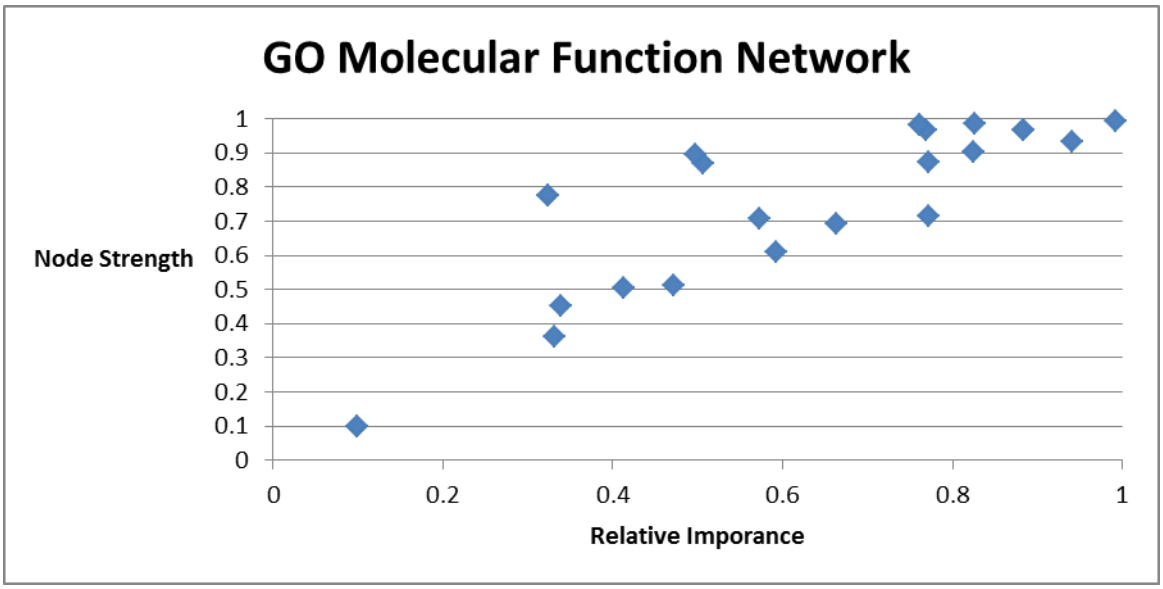

Figure 9. Relative importance versus node strength for the GO Molecular Function Network for rheumatoid arthritis.

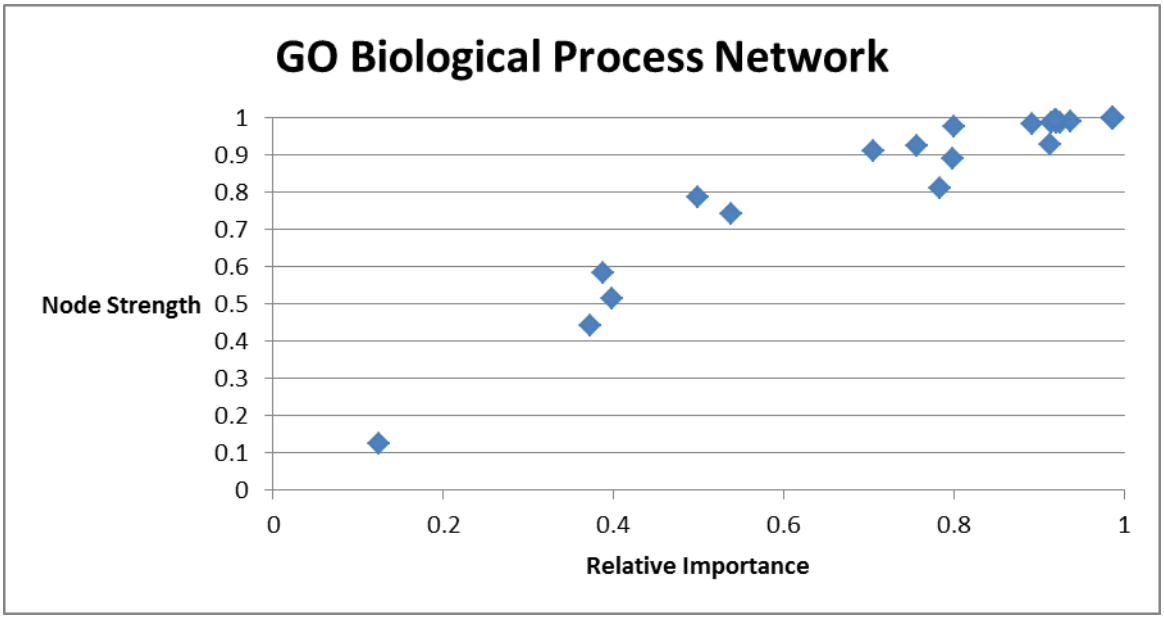

Figure 10. Relative importance versus node strength for the GO Biological Process Network for rheumatoid arthritis.

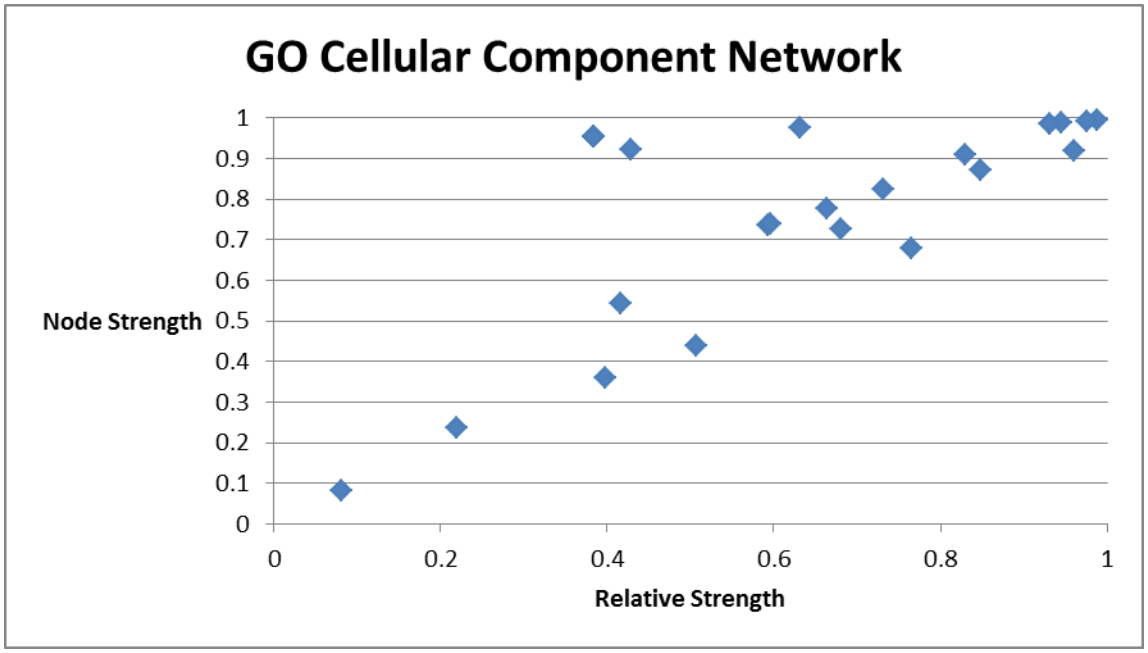

Figure 11. Relative importance versus node strength for the GO Cellular Component Network for rheumatoid arthritis.

Average node strength

The **average node strength** for the set *R* of root nodes is defined as the average of the node strengths of the nodes in *R* and is given by the following equation:

$$avg\_str(R) = \frac{\sum_{v \in R} str(v)}{|R|}. \quad (10)$$

Similarly, the average node strength for the set *C* of candidate nodes is given by the following equation:

$$avg\_str(C) = \frac{\sum_{v \in C} str(v)}{|C|}. \quad (11)$$

For a disease *D*, the difference between *avg_str*(*R*) and *avg_str*(*C*) is given by the following equation:

$$diff(D) = \frac{\sum_{v \in R} str(v)}{|R|} - \frac{\sum_{v \in C} str(v)}{|C|}. \quad (12)$$

I conjectured that greater the *diff*(*D*) for a disease *D* the higher will be the AUC obtained from the application of the KNGP algorithm to *D*. Figures 12, 13, 14 and 15 plot the *diff*(*D*) versus the AUCs for the 19 experimental diseases using the following knowledge networks respectively: the IID Network, the GO Molecular Network, the GO Biological Network, and the GO Cellular Network.

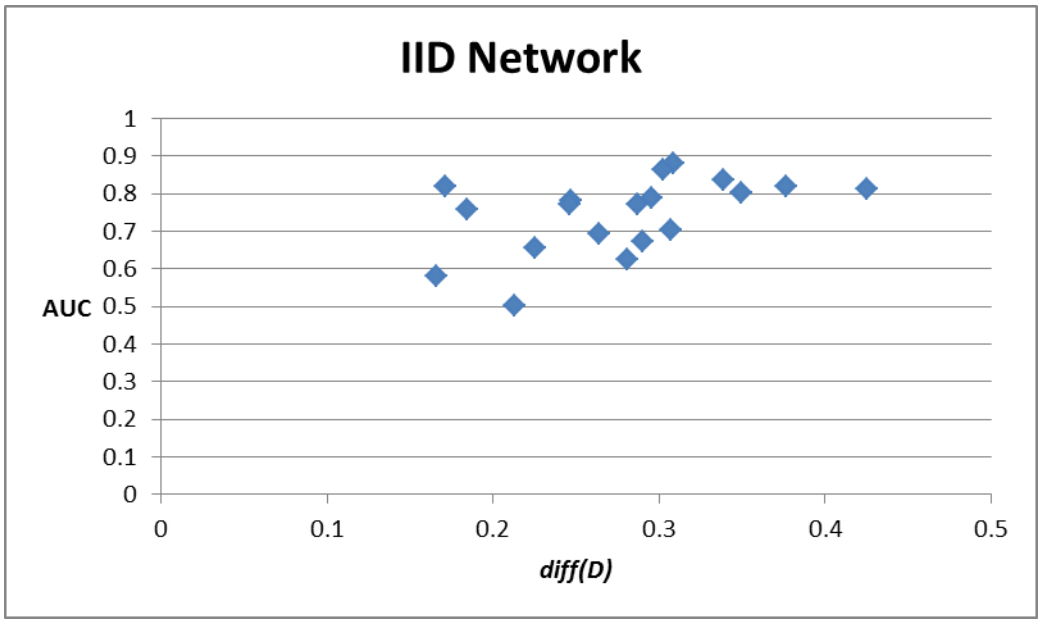

Figure 12. AUCs versus *diff*(*D*) for the IID Network.

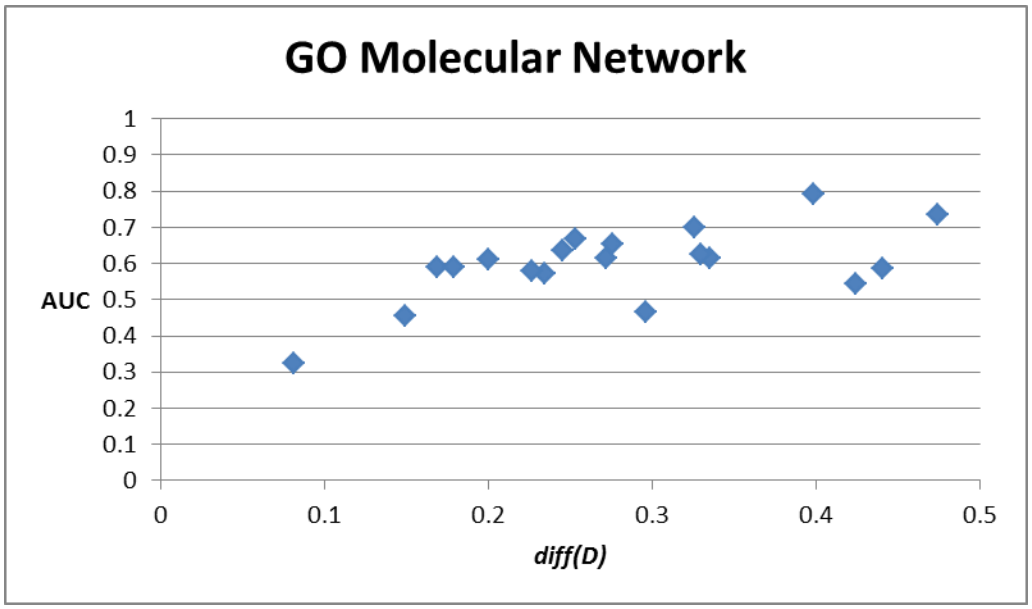

Figure 13. AUCs versus *diff*(*D*) for the GO Molecular Network.

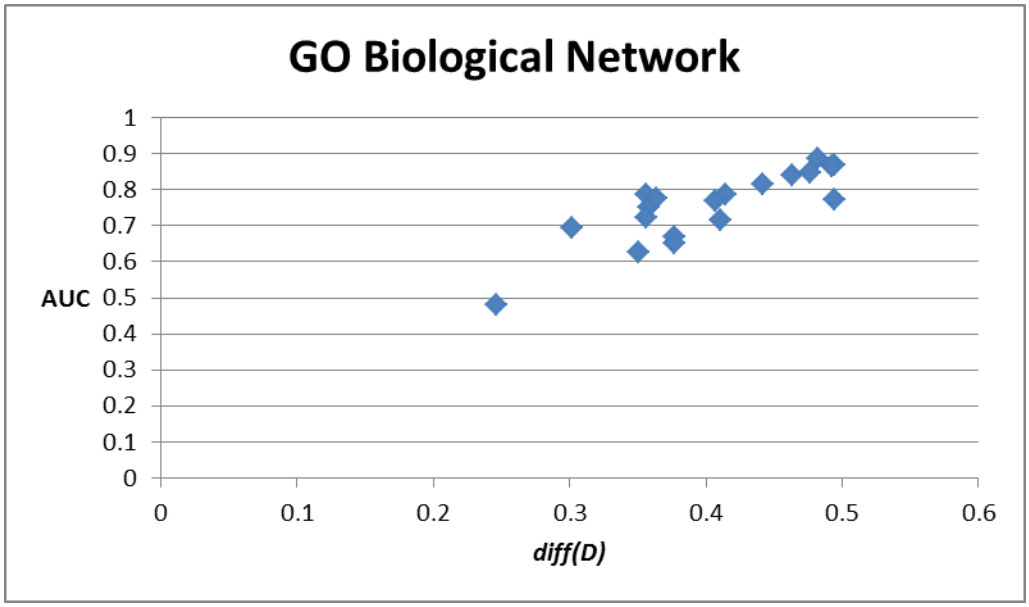

Figure 14. AUCs versus *diff*(*D*) for the GO Biological Network.

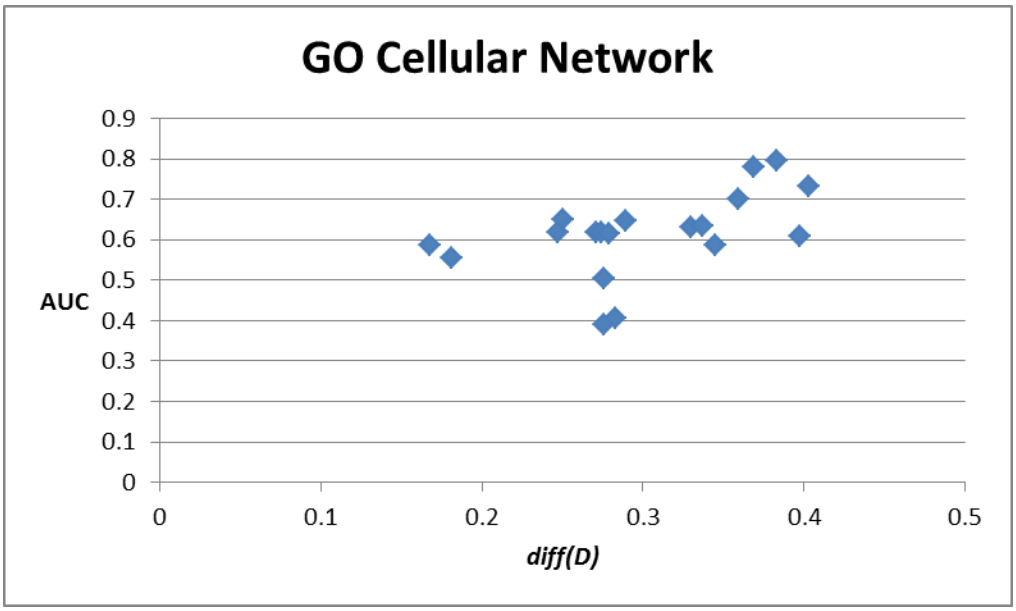

Figure 15. AUCs versus *diff*(*D*) for the GO Cellular Network.

In all four networks, there was a significant positive correlation between *diff* and AUC (p-value < 0.01), thus indicating that as the AUC increased, so did *diff*. Since *diff* is an indicator of the difference in similarity between the root and candidate genes, these results indicate *that*

root genes which are more similar (or more close together) to each other according to a given knowledge source will result in a greater AUC. In other words, the reason why the AUC for the GO Cellular Component Network (0.856) was greater than the GO Cellular Component Network (0.595) for pancreatitis is that the root proteins are more similar according to the GO Cellular Component network compared to the GO Biological Process network. Thus, when choosing a knowledge source for network-based gene prioritization, *one should choose the knowledge source which would naturally provide the greatest amount of similarity among the genes known to be associated with the disease*.

Table 9 below provides some of the link weight values between the endometriosis root proteins for the gene ontology cellular component network and the gene ontology biological process network. The first column provides a sample of some of the root protein pairs, and the second and third column provide the link weight values for the GO Cellular and GO Biological networks respectfully. The last row provides the average link weight values between the root proteins. As the table shows, the link weights for the GO Cellular Network were greater than the link weights for the GO Biological Process Network. The average link weight value for the GO Cellular Component Network was 0.55 and the average link weight value for the GO Biological Process network was 0.2. This up-weighted the transition probability values between the root proteins for the GO Cellular network vs. the GO Biological network which also increased the AUC values, because the KNGP algorithm utilizes the transition probability values. In other words, the average node strength for the GO Cellular network was greater than the average node strength for the GO Biological Process network which – as just previously mentioned and demonstrated – also increased the AUC values.

Table 9. Link Weight Values for the GO Cellular and GO Biological Component Networks

Root Protein Pair	GO Cellular Component Network	GO Biological Process Network
P24394-Q9BXN1	0.58	0.0
P24394-P02458	0.53	0.12
P24394-P11473	0.23	0.27
P24394-P43026	0.60	0.25
P24394-P02452	0.76	0.22
....	….	……
Average	**0.55**	**0.20**

5.2.2 Incorporation of combined link knowledge sources

Table 10 gives the AUCs for each network and knowledge source based on the incorporation of link knowledge from a combination of knowledge sources. The AUCs for the IID Network are shown for comparison. It is important to note that the IID Network contains only experimental interactions whereas the PPPI Network contains both experimental and predicted interactions.

Table 10. AUCs for networks with link weights from combination of sources.

Disease	IID	PPPI	PPPI+ GOM	PPPI+ GOB	PPPI+ GOC
Rheumatoid Arthritis	0.699	0.806	0.750	0.830	0.798
Parkinson's Disease	0.631	0.648	0.652	0.668	0.668
Celiac Disease	0.772	0.837	0.744	0.814	0.795
Esophageal Cancer	0.857	0.846	0.840	0.871	0.858
Hepatitis C	0.721	0.749	0.502	0.764	0.759
Crohn’s Disease	0.814	0.837	0.850	0.862	0.846
Breast Cancer	0.834	0.859	0.866	0.872	0.865
Asthma	0.774	0.861	0.797	0.856	0.825
Alzheimer’s Disease	0.835	0.835	0.807	0.843	0.828
Ulcerative Colitis	0.672	0.725	0.740	0.706	0.738
Endometriosis	0.772	0.940	0.747	0.953	0.944
Lymphoma	0.830	0.837	0.770	0.875	0.872

Osteoarthritis	0.753	0.823	0.840	0.778	0.837
Epilepsy	0.578	0.573	0.579	0.622	0.612
Atherosclerosis	0.798	0.820	0.880	0.840	0.827
Pancreatitis	0.767	0.847	0.852	0.715	0.865
Cirrhosis	0.600	0.667	0.525	0.689	0.683
Myocardial Infarction	0.865	0.878	0.884	0.892	0.880
Tuberculosis	0.664	0.871	0.800	0.887	0.876
Average	**0.747**	**0.805**	**0.757**	**0.807**	**0.809**
p-value	ref.	<0.05	<0.30	<0.05	<0.05

The PPPI Network performed well, and its AUCs were significantly greater than the IID Network (p-value < 0.00001) and the GO Biological Network (p-value < 0.03). Of the three PPPI+GO Networks, two of the three – the PPPI+GOB and PPPI+GOC Networks – performed significantly better than the IID Network (p-value < 0.001). Of all the networks, the network which performed the best was the PPPI+GOC Link Weight Network. These results underscore the *importance of combining knowledge from multiple sources when performing network based gene prioritization and knowing which data sources to combine*.

The fact that the PPPI Network performed significantly better than the IID Network was not surprising. The results from the incorporation of link knowledge from single knowledge sources (Section 4.1) showed that protein-protein interactions, by themselves, perform quite well, and it was thus not surprising that adding predicted interactions would also provide a benefit – especially since the predicted interactions came from a good source and had little overlap with the interactions from the IID.

Overall, the finding that multiple sources of knowledge can be combined to improve network based gene prioritization is not too surprising given the literature trail. For instance, Sun et al. [61] created a weighting scheme to combine information for gene prioritization from several genetic data sources for the disease schizophrenia. The genetic data sources included more than two-thousand association studies, genome-wide linkage scans, and gene expression

studies. The authors showed that their approach can be promising for gene prioritization and had some success. Even though the authors did not use a network-based approach as was used in this dissertation, the authors showed that the combination of knowledge can be useful for gene prioritization for at least one disease.

To understand the superior performance of PPPI, I examined a number of additional links among the root nodes in the PPPI Network compared to the IID Network. I calculated a disease statistic $ds(D)$ that measured the difference in the number of recorded interactions between the PPPI and IID Networks among the root nodes for a disease D. The disease statistic was defined as follows:

$$ds(D) = PercRootInt(PPPI) - PercRootInt(IID) \quad (12)$$

where *PercRootInt* is a procedure which outputs the percentage of links that are present among all possible root links for a given binary network. In other words, if the percentage is 0.10, this means that 10 percent of all possible root-root links were actually recorded as interactions in the given binary network. The first term on the right hand side of the equation uses the PPPI Network and the second term uses the IID Network. This disease statistic has a lower bound of 0, because the interactions in the IID Network always exist in the PPPI Network.

If there was a substantial difference in the AUC between the PPPI and IID Networks for a given disease (Table 10), one would expect the $ds(D)$ to be high. In order to test this, the difference in the AUCs between the PPPI and IID Networks were obtained along with the $ds(D)$ for each disease. Figure 16 shows the AUC differences versus $ds(D)$ for the experimental diseases.

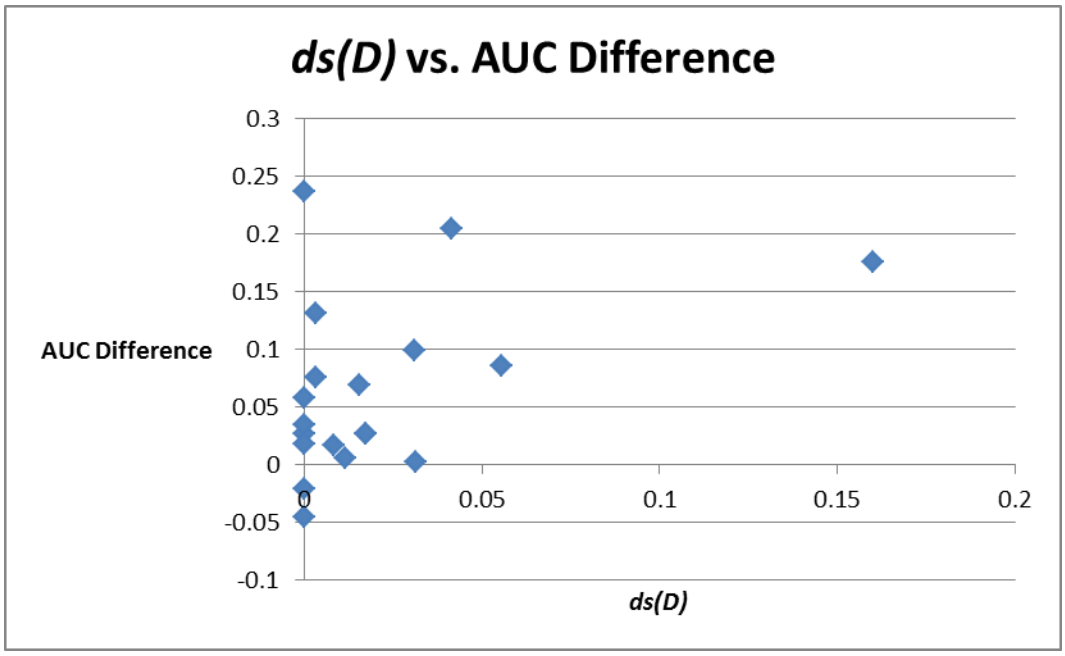

Figure 16. AUC difference versus *ds*(*D*).

There was a statistically significant correlation between *ds*(*D*) and the AUCs (p-value < 0.05). In other words, the large difference in the AUCs between the IID and PPPI Networks were due in large part to the additional interactions added to the PPPI Network among the root nodes.

5.2.3 Incorporation of node knowledge source

This section presents the results of networks that incorporate only node knowledge. Table 11 gives the AUCs for the following node weight networks: GO and InterPro. The AUCs for the IID Network are included for comparison.

Table 11. AUCs for node weight networks.

Disease	**IID**	**GO**	**InterPro**
Rheumatoid Arthritis	0.699	0.770	0.765
Parkinson's Disease	0.631	0.724	0.715

Celiac Disease	0.772	0.775	0.794
Esophageal Cancer	0.857	0.876	0.846
Hepatitis C	0.721	0.774	0.707
Crohn's Disease	0.814	0.808	0.832
Breast Cancer	0.834	0.855	0.841
Asthma	0.774	0.794	0.834
Alzheimer's Disease	0.835	0.868	0.854
Ulcerative Colitis	0.672	0.701	0.702
Endometriosis	0.772	0.758	0.880
Lymphoma	0.830	0.910	0.851
Osteoarthritis	0.753	0.803	0.790
Epilepsy	0.578	0.710	0.672
Atherosclerosis	0.798	0.885	0.821
Pancreatitis	0.767	0.755	0.809
Cirrhosis	0.600	0.579	0.673
Myocardial Infarction	0.865	0.885	0.868
Tuberculosis	0.664	0.833	0.745
Average	**0.749**	**0.793**	**0.789**
p-value	ref.	<0.05	<0.05

As Table 11 shows, the GO Node Weight Network had the highest average AUC. Both the GO and InterPro networks were significantly greater than the IID Network (p-value < 0.05). These results show that adding node knowledge – derived from a variety of different data sources – *can significantly benefit the network-based gene prioritization process*.

The reason that the two node weight networks performed significantly better than the IID Network is that the root proteins tend to have more GO and InterPro associations (and a correspondingly higher prior probabilities) than candidate proteins. This resulted in a prior probability vector where the root nodes had higher probabilities than the candidate nodes. For the GO Node Weight Network, the median number of GO associations for all 17,658 proteins (both root and candidate) was 36. However, the median number of GO associations for all 229 root proteins was 80. It is expected that root proteins would have more associations than candidate proteins, because disease-related proteins are probably researched considerably more than non-

disease-related proteins. In summary, the three node weight networks are taking advantage of the implicit property of the root proteins having more associations than the candidate proteins.

The plot in Figure 21 shows the distribution of GO associations for all proteins and the plot in Figure 22 shows the distribution of GO associations for only the root proteins. Thus, Figure 21 represents all 17,691 proteins, and Figure 22 represents the 229 unique root proteins associated with the 19 diseases.

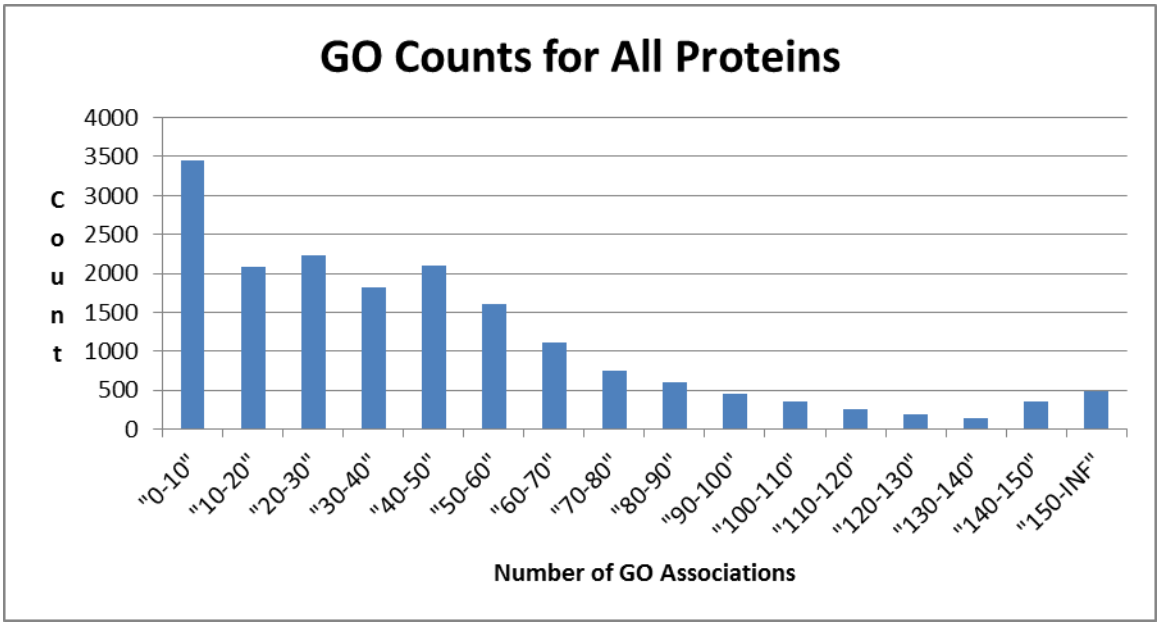

Figure 17. Histogram of GO associations for all proteins.

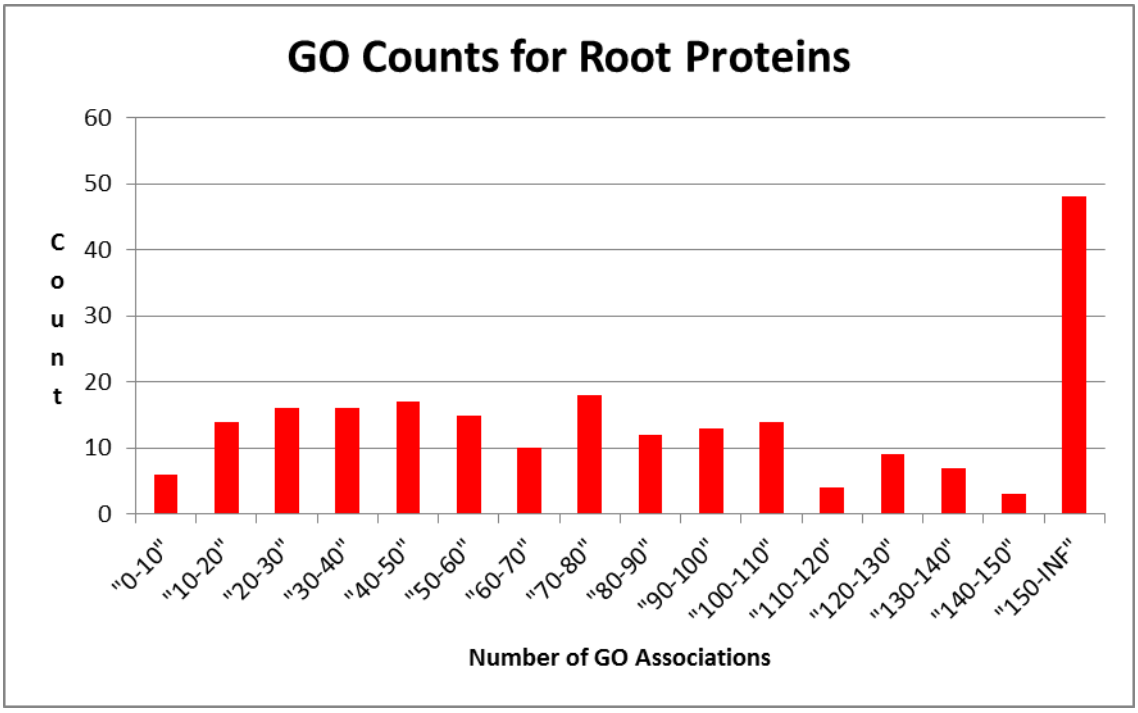

Figure 18. Histogram of GO associations for root proteins.

Form these two figures, it can be seen that the number of GO associations for the root proteins was – on average – larger than the number of GO associations for all 17,691 proteins. This resulted in the node weights (and subsequent prior weights) for the root proteins to be higher than the node weights for the candidate nodes resulting in greater AUC scores, because the relative importance of a given node in the KGNP algorithm is derived in part from the prior probabilities. This same relationship existed for the InterPro knowledge source as well.

5.2.4 Incorporation of link and node knowledge sources

The results in the preceding sections showed that incorporation of link knowledge in the form of predicted protein-protein interactions and GO Cellular Component knowledge combined (PPPI+GOC) and incorporation of node knowledge in the form of GO associations led to

improved performance. This section presents the results of a network that combines link weights from the PPPI+GOC Link Weight Network and node weights from the GO Node Weight Network. Table 12 below provides the AUCs for the PPPI+GOC Link Weight and GO Node Weight Network along with the AUCs for the PPPI+GOC Link Weight Network and the GO Node Weight Network for comparison. The PPPI+GOC Link Weight Network represented the optimal network using link knowledge, and the GO Node Weight Network represented the optimal network using node knowledge.

Table 12. AUCs for combined link and node weight networks.

Disease	**PPPI+ GOC Link Weight Network**	**GO Node Weight Network**	**PPI+GOC Link Weight and GO Node Weight Network**
Rheumatoid Arthritis	0.798	0.770	0.835
Parkinson's Disease	0.668	0.724	0.734
Celiac Disease	0.795	0.775	0.807
Esophageal Cancer	0.858	0.876	0.853
Hepatitis C	0.759	0.774	0.756
Crohn’s Disease	0.846	0.808	0.847
Breast Cancer	0.865	0.855	0.867
Asthma	0.825	0.794	0.845
Alzheimer’s Disease	0.828	0.868	0.863
Ulcerative Colitis	0.738	0.701	0.740
Endometriosis	0.944	0.758	0.986
Lymphoma	0.872	0.910	0.918
Osteoarthritis	0.837	0.803	0.858
Epilepsy	0.612	0.710	0.718
Atherosclerosis	0.827	0.885	0.896
Pancreatitis	0.865	0.755	0.878
Cirrhosis	0.683	0.579	0.666
Myocardial Infarction	0.880	0.885	0.907
Tuberculosis	0.876	0.833	0.943
Average	**0.809**	**0.793**	**0.838**
p-value	ref	ref	<0.05 / <0.05

The AUCs for the PPPI+GOC and GO Node Weight Network were significantly greater than the AUCs for the best link weight only network and the best node weight only network (p-value < 0.05). This shows that *the incorporation of both link and node knowledge together can significantly benefit the network-based gene prioritization process*. This combined link and node weight network represents the optimal network for the purpose of network-based gene prioritization.

5.2.5 Validation for asthma

For each of the 19 diseases, I scored and ranked all 17,691 proteins using the full set of root nodes and the PPPI+GOC Link Weight and GO Node Weight Network which was the best performing network. Table 13 gives the 5 top ranking candidate proteins (identified by UniProt identifier) for the disease asthma, and Appendix B provides the top 10 ranking candidate proteins for all 19 diseases. I searched the literature and found evidence for the two highlighted proteins in Table 13 being associated with asthma. Both of these proteins ranked low using the IID Network.

Table 13. Top five ranked candidate proteins for asthma.

Q01113 (IL9R)
Q13224 (GRIN2B)
P24394 (IL4R)
P29460 (IL12B)
P48357 (LEPR)

Kauppi et al. [62] genotyped several alleles from the IL9R gene and compared results between a large cohort of patients with asthma and healthy-control samples. The results were studied using linkage analysis, transmission disequilibrium, and homozygosity analyses. The authors showed that a IL9R allele – sDF2*10 – was more likely to be transmitted among patients

with asthma and was found homozygotic among asthma patients more often than expected. Furthermore, a specific X chromosome haplotype was found to be more associated for patients with asthma. This gene was ranked 1st out of approximately 17,500 proteins using the PPPI+GOC Link Weight and GO Node Weight Network but 926th using the IID Network. In order to test the hypothesis that the IL12B gene contains polymorphisms associated with asthma, Randolph et al. [63] performed a genotype analysis for polymorphisms in the IL12B gene between patients with asthma and their parents. In the results, the authors showed that one of the alleles of the IL12B gene was under-transmitted to children with asthma. Furthermore, the authors showed that a polymorphism of the IL2B gene may be significantly associated with asthma severity in whites. The IL12B gene was ranked 4th using the PPPI+GOC link weight and GO Node Weight Network but 290th using the IID Network.

Both the IL9R and IL12B genes were found to have a high likelihood of being associated with their respective diseases from the literature. This supports the validation and use of the PPPI+GOC Link Weight and GO Node Weight Network over the IID Network alone (the baseline) since the two genes were ranked high with the PPPI+GAD Network but low with the IID Network.

6.0 CONCLUSIONS

This dissertation explored in depth the network-based gene prioritization approach. Section 6.1 summarizes the main contributions of this dissertation. Section 6.2 discusses some of limitations and section 6.3 provides some directions for future work.

6.1 CONTRIBUTIONS

The first major contribution was the development of a network-based inference algorithm in order to incorporate node knowledge into the network-based gene prioritization process, and I called this algorithm the knowledge network-gene prioritization algorithm (KNGP). This algorithm generalizes two current network-based inference algorithms: PageRank and PageRank with Priors. Previous network-based inference algorithms could be used to incorporate only link knowledge while the KNGP algorithm can incorporate both link and node knowledge. The KGNP algorithm can be used to incorporate knowledge for any general purpose which can use network-based inference – not just gene prioritization.

The second major contribution was the investigation of whether biological knowledge can successfully be used to benefit the network-based gene prioritization process. This contribution was enveloped into four aims. For the first aim, the null hypothesis is accepted that the incorporation of knowledge from a single source does not provide a benefit for network-

based gene prioritization. The results showed that the use of protein-protein interaction knowledge is equal or better than all of the other types of knowledge sources tested. For the second aim, the null hypothesis was rejected that the combination of knowledge cannot provide a benefit for network-based gene prioritization. Particularly, the use of predicted-protein interactions – by itself – and in combination with the Gene Cellular Component ontology performed significantly better than using just experimental interactions. For the third aim, the null hypothesis was rejected that the incorporation of node knowledge does not provide a benefit for network-based gene prioritization. Particularly, the incorporation of GO and InterPro associations was shown to provide a significant benefit given that the node weights for all of the proteins are utilized. For the fourth aim, the null hypothesis was rejected that the combination of node and link knowledge does not provide a benefit for network-based gene prioritization. Particularly, the incorporation of link knowledge in the form of predicted protein-protein interactions with the Gene Cellular Component and node knowledge in the form of GO associations was shown to add a significant benefit. The following table provides the summary of the results for each of the aims enveloped within the second contribution.

Table 14. The Summary of Results for Each Aim

Aim	Summary of Results
Aim 1: Link Knowledge From a Single Source	Species and GOB network equal to IID network
Aim 2: Link Knowledge From a Combination of Sources	PPPI, PPPI+GOB, and PPPI+GOC networks were significantly better than IID network
Aim 3: Node Knowledge	GO and InterPro networks were significantly better than IID network
Aim 4: Combined Link and Node Knowledge	Combined link and node knowledge network was best network

The results from this contribution are significant in the area of gene prioritization for several reasons. First, this dissertation is the first to comprehensively compare multiple knowledge sources for the purpose of network-based gene prioritization. Second, this dissertation marks the first time that node knowledge has been incorporated into the network-based gene prioritization process. Previous work has only incorporated link knowledge.

6.2 LIMITATIONS

The biggest limitation of incorporating knowledge into the network-based approach is that many of the knowledge sources that I considered had poor annotation coverage for the proteins. I applied the criterion that in order for a given knowledge source to be included, the knowledge source should have at least one known annotation for at least 75% of all proteins listed in UniProt. This criterion eliminated several knowledge sources including protein domain and pathway knowledge. One possible remedy for this problem is to include predicted knowledge similar to predicted knowledge that is available for protein-protein interactions.

The other major limitation is the high computational space and time requirements of the PageRank with Priors and KGNP algorithm. The PageRank with Priors algorithms could be implemented using matrix algebra, but with a total of approximately 17,500 proteins, this required a ~ 17,500 by 17,500 matrix of real numbers to be stored in memory (although only half of these numbers actually needed to be stored due to the symmetric nature of the matrix). Fortunately, a computer with a sufficient amount of RAM could be found to run the algorithm in sufficient space and – thanks to Python's multi-processor threading capabilities – in sufficient time.

6.3 FUTURE WORK

There are several possible extensions to the work described in this dissertation. For example, there are several additional knowledge sources which could have been added. There is a wealth of proteomic and genomic knowledge stored in a number of different systems biology databases. These systems biology databases store complex information about genes and proteins such as more in depth information about how various proteins interact (e.g., transcription, methylation, etc.). This in-depth systems type information could be used as an additional knowledge source for gene prioritization. However, it is not exactly clear how one would create a similarity measure for this type of information, because these types of databases tend to be fee-for-service software and thus the data would not be easy to download and obtain. There is also a wealth of predicted knowledge which could potentially be added to this dissertation. For instance, Troyanskaya et al. [64] constructed MAGIC ((Multisource Association of Genes by Integration of Clusters). MAGIC is based on a Bayesian system that combines evidence from heterogeneous data sources (mostly high throughput data) to predict whether two proteins are functionally related. The authors compared their predictive system to the Gene Ontology (GO) as the gold standard, and the system performed adequately. Dale et al. [65] used a series of machine learning methods – including naïve Bayes, decision trees, and logistic regression – to predict the pathways for a number of proteins. The authors showed that these machine learning methods performed better than several previously known pathway prediction algorithms. These predicted sources of biological knowledge could be useful. However, it may be difficult to obtain and use these algorithms to predict the biological data.

APPENDIX A

PROTEINS ASSOCIATED WITH EACH DISEASE BY UNIPROT ID

Rheumatoid Arthritis	Parkinson's disease	Celiac Disease	Esophageal Cancer	Hepatitis C	Crohn's Disease
P20039	P29475	P29459	P04818	Q8IZI9	P20039
P21580	P27338	P01920	P24385	P20591	P17706
P08700	P04062	O95256	P42898	Q8IZJ0	P08571
Q03519	P15559	Q9HBE4	P15559	P16410	P35408
Q96A65	Q5S007	Q13478	P05091	P10914	P54652
Q9UNS1	P00326	Q01638	P34896	P01130	Q8TAU0
P01909	P07339	Q08116	O14965	P30685	Q14116
P49279	Q9H1E3	Q04864	P04798	Q30201	Q9H015
P51681	P10635	P01909			O76082
Q15116	Q8IUH8	P16410			P08183
Q14116	Q92731	Q9Y6W8			P19438
P22301	P50406	P60568			P00738
P16410	P23560	P32246			P26927
O75015	P10636	P51677			P01375
P08637	P27169	Q2LD37			Q5VWK5
P01920	Q9BXM7	Q9UQQ2			P14174
P01584	P43354				Q676U5
Q7RTU3	P09488				
P18510	O60260				
P31939	P52824				
P19438	P06307				
Q96P31					
P01579					
P08254					

Breast Cancer	Asthma	Alzheimer's Disease	Ulcerative Colitis	Endometriosis	Lymphoma
P20815	P01024	P02654	P20039	P04440	P04637
P05121	Q9GZX7	P21397	P40879	P06401	P22301
P08183	P60022	O96008	P09622	P15692	P16410
P03372	P21731	P01375	P12318	P01909	P24394

P04637 P05164 P27169 P50225 P09211 P04179 P06401 P38398 Q16678 P05093 P22455 P21802 P39060 P16035 P33241 Q13233 P05106 P04798 P29474 P11473 Q14790 P01579 P08253	P05121 Q15746 P36222 P01375 P01920 P20930 Q9NQ38 Q9BZ11 P13500 P13501 P01909 Q14116 P35225 P04440 P05112 P10145 Q14765 P51677 P05106 P29475 P11684 Q9UIL8 P01011 P01579 P14780	P05164 P78380 Q15165 Q12800 P10909 Q9BZA7 P01584 Q03014 Q9Y6A2 P02649 P01034 P30533 P28223 P30456 P06276 P49768 P04406	P08571 P01903 Q9NZK7 Q8TAU0 Q9NPH9 P07942 O00206 Q14116 P22301 P20809 P08183 Q9H257 P26927 P01375 Q5VWK5 P14174 P01579 Q9NXI6 Q9UIR0 Q0VDK5	P03372	Q9UNQ0 P10415 P41182
Osteoarthritis	**Epilepsy**	**Atherosclerosis**	**Pancreatitis**	**Cirrhosis**	**Myocardial Infarction**
P24394 Q9BXN1 P02458 P11473 P43026 P02452 P01583 Q9UEF7	P01213 P23560 O95180 Q8N135 P35498 P18507	P13498 P02656 P29474 P08571 P16284 Q07869 P02741 P49238 P45452 P04035 P11150 P11597 P34913 Q6Q788 P02647 P02649 P09601 P05231 P23946 P05362 P12821 Q9BQB6 O60603 P07204	P20039 P07477 O00206 P00995 P05091 P13569	P05091 Q5SRN2 P01375 P01920 P05019 Q9UIR0 Q30201	P40225 P42772 P29474 P15692 P02741 P02649 P04114 P07359 P16442 P05362 P42771 P51681 P08514 P07204 P03372 P41597 P00488 Q15848 P12821 P07996 P00748 P16284 P05019 P30533

		P07203 Q15848 P30556 P04180 P01344 P35520 P16109 P78380 P27169 P35568 Q15165 P01303 P14780 P08253 P08254 Q9UEF7 P06858 Q8TE73 P01019			P78380 P27169 Q15165 P08254 P11712 P05121 P06858 P01019
Tuberculosis					
P20039 Q99572 P11473 P42701 P01909 P10145 P49279 O60603 P22301 P29460 P01920 P01579					

APPENDIX B

TOP 10 RANKING CANDIDATE PROTEINS FOR EACH OF THE 19 EXPERIMENTAL DISEASES

Rheumatoid Arthritis	**Parkinson's disease**	**Celiac Disease**	**Esophageal Cancer**	**HepatitisC**	**Crohns**
P01023	P31946	P01562	P00167	P15813	P11831
Q03518	P14672	P29460	P00387	P61769	Q15599
P29460	Q15796	Q9HBE5	Q9UBK8	Q9BXS5	O14745
P48357	Q92793	Q14213	P11142	P04114	Q5T2W1
Q01113	P06241	P23743	P14672	P02786	Q86UT5
O15533	P63104	P01589	P08107	P29016	P20333
P28062	P62158	P31785	Q9BZE4	P30456	P01374
P42702	P84022	P19397	Q9UL45	P17693	Q93038
P08253	Q9Y4K3	P14784	P04637	P30511	Q9Y5U5
P30685	P04637	P29353	Q92541	P42229	P08138
Breast Cancer	**Asthma**	**Alzheimer**	**Ulcerative colitis**	**Endometriosis**	**Lymphoma**
P00167	Q01113	P00740	P01023	O14786	P3194
P00387	Q13224	P02647	P11831	P49765	P14672
Q9UBE0	P24394	P02652	Q14626	P17948	Q15796
P16435	P29460	Q9UBK8	P43364	Q13275	Q92793
Q9UBK8	P48357	P04070	P19438	P49763	P06241
Q03135	P07477	P00742	Q08334	P22105	P63104
P29460	P42702	P02768	O95999	P29279	P62158
P48357	P02751	P00734	P20333	P35052	P84022
P08047	P27694	P13500	P29460	P09486	Q9Y4K3
P00451	P08887	P31946	Q15599	P04004	O15198
Osteoarthritis	**Epilepsy**	**Atherosclerosis**	**Pancreatitis**	**Cirrhosis**	**Myocardial Infarction**
Q16270	P31946	P00519	O60603	P17936	P00734
P12643	P14672	P42684	O60602	P24593	P00740
Q99985	Q15796	P00734	Q15399	P22692	P04070
P01584	Q92793	P04114	Q99836	P08833	P00742
Q99584	P06241	P00740	P01903	P18065	P02647

Q9NPH3	P63104	Q07954	Q9BXR5	P24592	P02760
P02751	P62158	Q02156	Q9Y2C9	P19438	P01008
P21810	P84022	P31946	P58753	P20333	P05155
P09486	Q9Y4K3	P00747	P08571	Q16270	Q04756
P07996	P04637	P02652	Q9NYK1	P01374	P04004
Tuberculosis					
P01023					
O95425					
P54852					
P16333					
P62736					
P00519					
P62993					
Q08334					
P42702					
P48357					

BIBLIOGRAPHY

1. Adie, E.A., et al., *Speeding disease gene discovery by sequence based candidate prioritization.* BMC Bioinformatics, 2005. **6**: p. 55.
2. Adie, E.A., et al., *SUSPECTS: enabling fast and effective prioritization of positional candidates.* Bioinformatics, 2006. **22**(6): p. 773-4.
3. Aerts, S., et al., *Gene prioritization through genomic data fusion.* Nat Biotechnol, 2006. **24**(5): p. 537-44.
4. Chen, J., et al., *Improved human disease candidate gene prioritization using mouse phenotype.* BMC Bioinformatics, 2007. **8**: p. 392.
5. L. Page, S.B., R. Motwani, and T. Winograd, *The PageRank citation ranking: Bringing order to the web.* Technical Report, 1998.
6. Kleinberg, J.M., *Authoritative sources in a hyperlinked environment.* The Journal of ACM, 1999. **46**(5): p. 604–632.
7. Page, L.a.B., Sergey and Motwani, Rajeev and Winograd, Terry, *The PageRank Citation Ranking: Bringing Order to the Web.* Stanford InfoLab, 1999.
8. Freudenberg, J. and P. Propping, *A similarity-based method for genome-wide prediction of disease-relevant human genes.* Bioinformatics, 2002. **18 Suppl 2**: p. S110-5.
9. Radivojac, P., et al., *An integrated approach to inferring gene-disease associations in humans.* Proteins, 2008. **72**(3): p. 1030-7.
10. Rossi, S., et al., *TOM: a web-based integrated approach for identification of candidate disease genes.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W285-92.
11. Perez-Iratxeta, C., P. Bork, and M.A. Andrade, *Association of genes to genetically inherited diseases using data mining.* Nat Genet, 2002. **31**(3): p. 316-9.
12. George, R.A., et al., *Analysis of protein sequence and interaction data for candidate disease gene prediction.* Nucleic Acids Res, 2006. **34**(19): p. e130.
13. Hua, D. and Y. Lai, *An ensemble approach to microarray data-based gene prioritization after missing value imputation.* Bioinformatics, 2007. **23**(6): p. 747-54.
14. De Bie, T., et al., *Kernel-based data fusion for gene prioritization.* Bioinformatics, 2007. **23**(13): p. i125-32.
15. Hutz, J.E., et al., *CANDID: a flexible method for prioritizing candidate genes for complex human traits.* Genet Epidemiol, 2008. **32**(8): p. 779-90.
16. Gaulton, K.J., K.L. Mohlke, and T.J. Vision, *A computational system to select candidate genes for complex human traits.* Bioinformatics, 2007. **23**(9): p. 1132-40.

17. Shriner, D., et al., *Commonality of functional annotation: a method for prioritization of candidate genes from genome-wide linkage studies.* Nucleic Acids Res, 2008. **36**(4): p. e26.
18. Linghu, B., et al., *Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network.* Genome Biol, 2009. **10**(9): p. R91.
19. Chen, J., B.J. Aronow, and A.G. Jegga, *Disease candidate gene identification and prioritization using protein interaction networks.* BMC Bioinformatics, 2009. **10**: p. 73.
20. Oti, M., et al., *Predicting disease genes using protein-protein interactions.* J Med Genet, 2006. **43**(8): p. 691-8.
21. Chen, J.Y., C. Shen, and A.Y. Sivachenko, *Mining Alzheimer disease relevant proteins from integrated protein interactome data.* Pac Symp Biocomput, 2006: p. 367-78.
22. Gonzalez, G., et al., *Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures.* Pac Symp Biocomput, 2007: p. 28-39.
23. Rives, A.W. and T. Galitski, *Modular organization of cellular networks.* Proc Natl Acad Sci U S A, 2003. **100**(3): p. 1128-33.
24. LaCount, D.J., et al., *A protein interaction network of the malaria parasite Plasmodium falciparum.* Nature, 2005. **438**(7064): p. 103-7.
25. Kohler, S., et al., *Walking the interactome for prioritization of candidate disease genes.* Am J Hum Genet, 2008. **82**(4): p. 949-58.
26. Wu, X., et al., *Network-based global inference of human disease genes.* Mol Syst Biol, 2008. **4**: p. 189.
27. van Driel, M.A., et al., *A new web-based data mining tool for the identification of candidate genes for human genetic disorders.* Eur J Hum Genet, 2003. **11**(1): p. 57-63.
28. Ala, U., et al., *Prediction of human disease genes by human-mouse conserved coexpression analysis.* PLoS Comput Biol, 2008. **4**(3): p. e1000043.
29. Nitsch, D., et al., *Network analysis of differential expression for the identification of disease-causing genes.* PLoS One, 2009. **4**(5): p. e5526.
30. Nitsch, D., et al., *Candidate gene prioritization by network analysis of differential expression using machine learning approaches.* BMC Bioinformatics, 2010. **11**: p. 460.
31. Chung F, Y.S., *Coverings, heat kernels and spanning trees.* Electronic Journal of Combinatorics, 1999. **6**.
32. Y, S., *Analysis of some Krylov subspace approximations to the matrix exponential operator.* SIAM Journal on Numerical Analysis (SINUM), 1992. **29**(1): p. 209-228.
33. Nitsch, D., et al., *PINTA: a web server for network-based gene prioritization from expression data.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W334-8.
34. Karni, S., H. Soreq, and R. Sharan, *A network-based method for predicting disease-causing genes.* J Comput Biol, 2009. **16**(2): p. 181-9.
35. Brin, S., Page, L., *The anatomy of a large-scale hypertextual Web search engine.* . Proceedings of the 7th International World Wide Web Conference, 1998: p. 107-117.
36. Scott White, P.S., *Algorithms for Estimating Relative Importance in Networks.* Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining\, 2003.

37. Franke, L., et al., *Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes.* Am J Hum Genet, 2006. **78**(6): p. 1011-25.
38. Chen, Y., et al., *In silico gene prioritization by integrating multiple data sources.* PLoS One, 2011. **6**(6): p. e21137.
39. Jain, E., et al., *Infrastructure for the life sciences: design and implementation of the UniProt website.* BMC Bioinformatics, 2009. **10**: p. 136.
40. Kalia, M. and A. Kukol, *Structure and dynamics of the kinase IKK-beta - A key regulator of the NF-kappa B transcription factor.* J Struct Biol, 2011. **176**(2): p. 133-42.
41. Chylek, L.A., et al., *Guidelines for visualizing and annotating rule-based models.* Mol Biosyst, 2011. **7**(10): p. 2779-95.
42. De Las Rivas, J. and C. Fontanillo, *Protein-protein interactions essentials: key concepts to building and analyzing interactome networks.* PLoS Comput Biol, 2010. **6**(6): p. e1000807.
43. Brown, K.R. and I. Jurisica, *Online predicted human interaction database.* Bioinformatics, 2005. **21**(9): p. 2076-82.
44. Brown, K.R. and I. Jurisica, *Unequal evolutionary conservation of human protein interactions in interologous networks.* Genome Biol, 2007. **8**(5): p. R95.
45. Prasad, T.S., K. Kandasamy, and A. Pandey, *Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology.* Methods Mol Biol, 2009. **577**: p. 67-79.
46. Ceol, A., et al., *MINT, the molecular interaction database: 2009 update.* Nucleic Acids Res, 2010. **38**(Database issue): p. D532-9.
47. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
48. Wang, J.Z., et al., *A new method to measure the semantic similarity of GO terms.* Bioinformatics, 2007. **23**(10): p. 1274-81.
49. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.
50. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.
51. Beer, D.G., et al., *Gene-expression profiles predict survival of patients with lung adenocarcinoma.* Nat Med, 2002. **8**(8): p. 816-24.
52. McDowall, M.D., M.S. Scott, and G.J. Barton, *PIPs: human protein-protein interaction prediction database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D651-6.
53. Scott, M.S. and G.J. Barton, *Probabilistic prediction and ranking of human protein-protein interactions.* BMC Bioinformatics, 2007. **8**: p. 239.
54. Hunter, S., et al., *InterPro: the integrative protein signature database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D211-5.
55. Sharma, A., et al., *Gene prioritization in Type 2 Diabetes using domain interactions and network analysis.* BMC Genomics, 2010. **11**: p. 84.
56. Navlakha, S. and C. Kingsford, *The power of protein interaction networks for associating genes with diseases.* Bioinformatics, 2010. **26**(8): p. 1057-63.
57. Jimenez-Sanchez, G., B. Childs, and D. Valle, *Human disease genes.* Nature, 2001. **409**(6822): p. 853-5.

58. Oti, M., et al., *Conserved co-expression for candidate disease gene prioritization.* BMC Bioinformatics, 2008. **9**: p. 208.
59. Hristovski, D., et al., *Using literature-based discovery to identify disease candidate genes.* Int J Med Inform, 2005. **74**(2-4): p. 289-98.
60. W, D., *Applied nonparametric statistics.* PWS-KENT Publishing Company, 1990. **2nd**.
61. Sun, J., et al., *A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases-schizophrenia as a case.* Bioinformatics, 2009. **25**(19): p. 2595-6602.
62. Kauppi, P., et al., *The IL9R region contribution in asthma is supported by genetic association in an isolated population.* Eur J Hum Genet, 2000. **8**(10): p. 788-92.
63. Randolph, A.G., et al., *The IL12B gene is associated with asthma.* Am J Hum Genet, 2004. **75**(4): p. 709-15.
64. Troyanskaya, O.G., et al., *A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae).* Proc Natl Acad Sci U S A, 2003. **100**(14): p. 8348-53.
65. Dale, J.M., L. Popescu, and P.D. Karp, *Machine learning methods for metabolic pathway prediction.* BMC Bioinformatics, 2010. **11**: p. 15.