TEXT AND NETWORK MINING FOR LITERATURE-BASED SCIENTIFIC DISCOVERY IN BIOMEDICINE

by

Arzucan Özgür

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in The University of Michigan
2010

Doctoral Committee:

Professor Dragomir R. Radev, Chair
Professor Brian D. Athey
Professor Satinder Singh Baveja
Professor Hosagrahar V. Jagadish
Associate Professor Steven P. Abney

To my family.

ACKNOWLEDGEMENTS

First and foremost, I am truly grateful to my advisor Dragomir Radev. His guidance and support have been essential not only in the development of this thesis, but also in my development as a scientist. I would also like to thank the other members of my dissertation committee, Steve Abney, Brian Athey, H.V. Jagadish, and Satinder Singh for their careful criticism and insightful feedback that improved the quality of this work.

I would like to thank all the past and current members of the Computational Linguistics and Information Retrieval (CLAIR) research group for being wonderful colleagues and friends. I am especially thankful to my collaborators Güneş Erkan, Thuy Vu, and Amjad abu Jbara. Thanks also to my office mates Kyla McMullen, Ahmed Hassan, Vahed Qazvinian and Pradeep Muthukrishnan.

I am very grateful to He Group, in particular to Yongqun "Oliver" He and Zuoshuang Xiang. Without our fruitful collaboration, Chapter V of this thesis would have not been possible.

I have had the opportunity of working with a number of outstanding researchers in the National Center for Integrative Biomedical Informatics (NCIBI), including Alex Ade, Glenn Tarcea, Jing Gao, H. V. Jagadish, Terry Weymouth, David States, James Cavalcoli, and Brian Athey. I am thankful for all their valuable comments and contributions to my research. I would also like to acknowledge the NIH Grant U54 DA021519 to the National Center for Integrative Biomedical Informatics that

supported this research.

My heartfelt thanks go to my friends who have been a family for me here in Ann Arbor, especially Hacer Karataş, Thuy Vu, Hande Koçak, Ayşe Büyüktür and Bilgen Ekim. Sincere thanks to Güniz Büyüktür for her invaluable mentoring. I would also like to thank all my professors and colleagues in the Computer Engineering Department of Boğaziçi University, and all my friends in Turkey.

This thesis would have not been possible without the support of my family. I would like to thank my in-laws, Ferhan and Hacı Türkmen, as well as Nimet and Burçin Yıldırım, who supported and encouraged me, even from miles of distance. The love and prayers of my grandparents have always been with me. Genuine thanks go to them as well as to my other relatives. Words cannot express my gratitude to my parents, Rahime and Haydar Özgür, and my brother Mürvet Ozan Özgür for unconditionally providing their love, guidance, and support, and making it possible for me to pursue my goals in life. Special thanks go to my husband Hamdi İlker Türkmen for his great support and patience, and for loving me so much that he could postpone his dreams in order for me to pursue mine.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Most of the new and important findings in biomedicine are only available in the text of the published scientific articles. The first goal of this thesis is to design methods based on natural language processing and machine learning to extract information about genes, proteins, and their interactions from text. We introduce a dependency tree kernel based relation extraction method to identify the interacting protein pairs in a sentence. We propose two kernel functions based on cosine similarity and edit distance among the dependency tree paths connecting the protein names. Using these kernel functions with supervised and semi-supervised machine learning methods, we report significant improvement (59.96% F-Measure performance over the AIMED data set) compared to the previous results in the literature. We also address the problem of distinguishing factual information from speculative information. Unlike previous methods that formulate the problem as a sentence classification task, we propose a two-step method to identify the speculative fragments of sentences. First, we use supervised classification to identify the speculation keywords using a diverse set of linguistic features that represent their contexts. Next, we use the syntactic structures of the sentences to resolve their linguistic scopes. Our results show that the method is effective in identifying speculative portions of sentences. The speculation keyword identification results are close to the upper bound of human inter-annotator agreement.

The second goal of this thesis is to generate new scientific hypotheses using the

literature-mined protein/gene interactions. We propose a literature-based discovery approach, where we start with a set of genes known to be related to a given concept and integrate text mining with network centrality analysis to predict novel concept-related genes. We present the application of the proposed approach to two different problems, namely predicting gene-disease associations and predicting genes that are important for vaccine development. Our results provide new insights and hypotheses worth future investigations in these domains and show the effectiveness of the proposed approach for literature-based discovery.

CHAPTER I

Introduction

1.1 Motivation

The post-genome era, which started with the completion of the Human Genome Project (Lander et al., 2001; Venter et al., 2001) has brought new research opportunities and challenges. The primary goal in this new era is interpreting the genome data, in other words understanding the functions of the genes, the proteins they code for, and their roles in the biological pathways. New techniques such as high-throughput experimental methods have been developed, which contributed to the generation of massive amounts of biomedical data and to the rapid increase in the number of published scientific articles in the domain.

The main system that provides access to the biomedical article citations and abstracts from MEDLINE and additional life sciences journals is the PubMed system[1], which is maintained by the U.S. National Library of Medicine and the National Institutes of Health. Some of the entries in PubMed include links to full-text articles from publisher web sites or PubMed Central[2], which contains nearly 2 million articles. Biomedical literature is growing at a double-exponential rate (Cohen & Hunter, 2004; Hunter & Cohen, 2006). In other words, both the total size of PubMed and the

[1] http://www.ncbi.nlm.nih.gov/pubmed/
[2] http://www.pubmedcentral.nih.gov/

growth rate of PubMed are increasing exponentially. Figure 1.1 shows the new entries added to PubMed each year between 1948 and 2008. While the average number of publications added to PubMed per day was only 194 in 1948, between 2000 and 4000 entries per working day have been added to PubMed since 2005[3]. For example, 811375 new entries were added to PubMed during 2008, which corresponds to an average of 2216 new entries per day. Figure 1.2 shows the total number of entries in PubMed at the end of each year between 1948 and 2008. While there were only 70871 entries in PubMed in 1948, currently there are over 19 million publications. Over 6.8 million of these publications were added in the last 10 years.

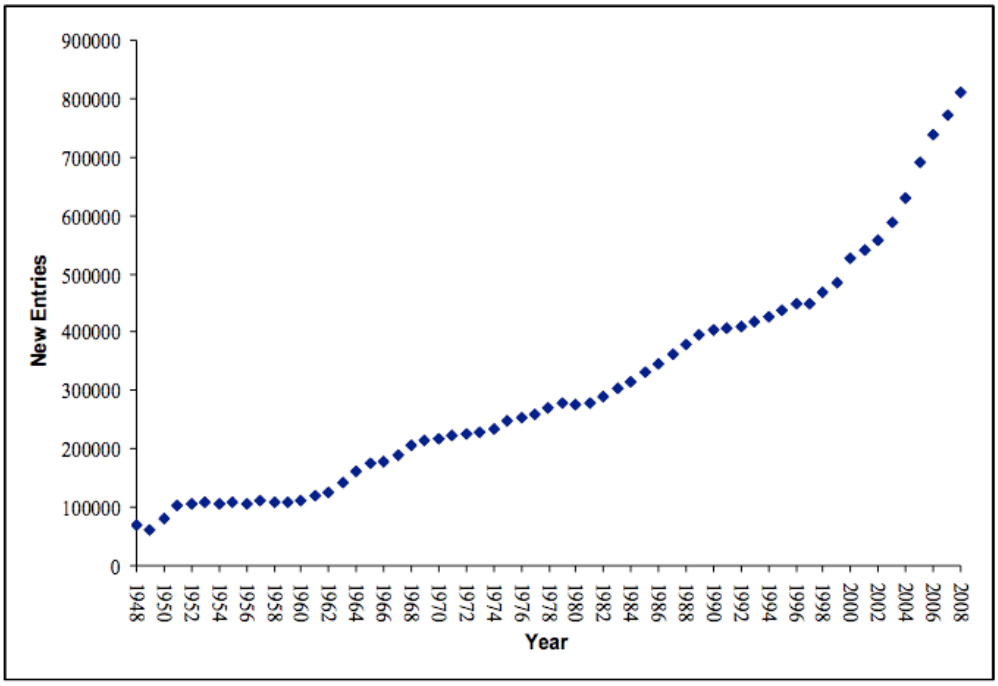


Figure 1.1: Growth in the Biomedical literature between 1948-2008. The plot shows the new entries added to the PubMed database each year.

The main way that biomedical researchers communicate their new findings is through scientific publications, written in natural language. Given the current amount and the growth rate of the biomedical literature, it is difficult or impos-

[3]http://www.nlm.nih.gov/pubs/factsheets/medline.html

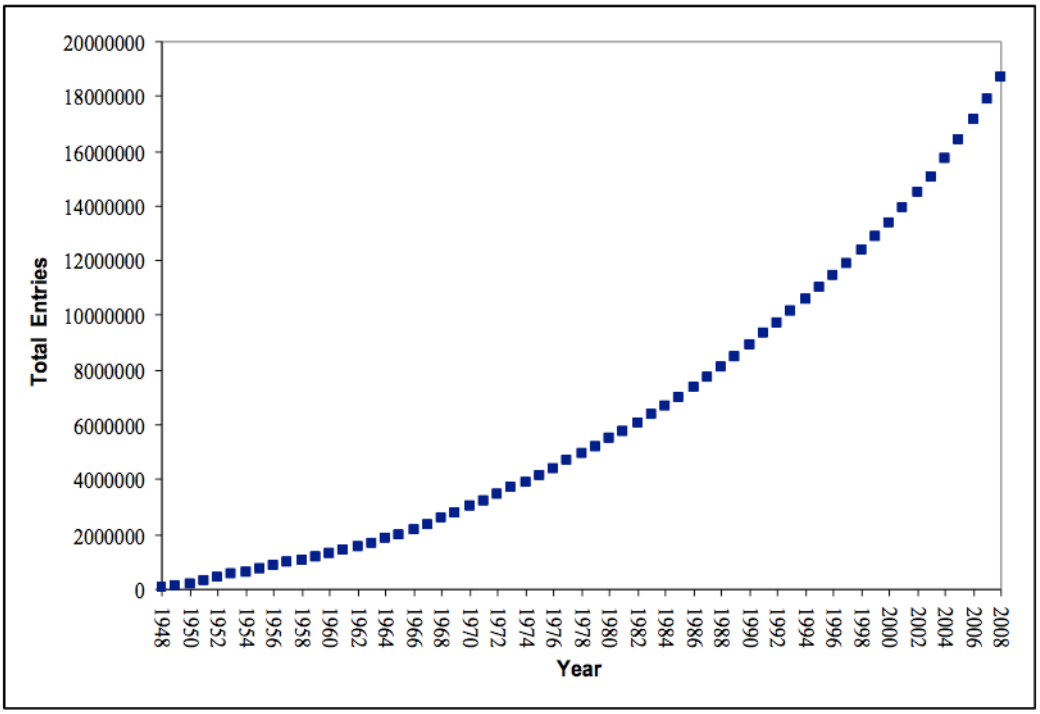


Figure 1.2: Growth in the Biomedical literature between 1948-2008. The plot shows the total entries indexed in the PubMed database at the end of each year.

sible for biomedical scientists to keep up with the relevant publications and utilize the knowledge contained in them. For example, consider a researcher interested in the interferon-gamma gene. A search in PubMed for "(ifn-gamma OR interferon-gamma)" will return 75464 articles[4]. Even if the researcher is interested in only certain aspects of the gene such as in its relatedness to vaccine development and restricts his search to "vaccine AND (ifn-gamma OR interferon-gamma)", the number of articles retrieved is 7536, which is still too high for reading manually.

There are a number of manually curated databases that store protein interactions, such as the Molecular INTeraction database (MINT) (Zanzoni et al., 2002), the Biomolecular Interaction Network Database (BIND) (Bader et al., 2003), SwissProt (Bairoch & Apweiler, 2000), and the Human Protein Reference Database (HPRD) (Keshava Prasad et al., 2009). Many databases also summarize results

[4]As of October, 2009

from publications about gene-disease relationships, such as the Online Mendelian Inheritance in Man (OMIM) (OMIM, 2007), the Pathogen-Host Interaction Data Integration and Analysis System (PHIDIAS) (Xiang et al., 2007), and the Genetic Association Database (GAD) (Becker et al., 2004). However, it usually takes a lot of time and effort before new discoveries are included in these databases. As a result, most of the knowledge remains hidden in the unstructured text of the published articles. Developing text mining techniques to uncover this knowledge is not only useful, but also necessary to facilitate biomedical research.

This thesis has two main objectives. The first goal is to design methods based on natural language processing and machine learning to extract information about genes, proteins, and their interactions from text. The second goal is to use the extracted information to build literature-mined protein interaction networks and to generate new scientific hypotheses by analyzing these networks.

The next section gives a brief introduction of protein interaction networks, biomedical information extraction and literature-based discovery. Work more closely related to ours is discussed in the related work sections of the subsequent chapters. We conclude this chapter with a summary of the remaining chapters.

1.2 Background

1.2.1 Protein Interaction Networks

The major goal in the post-genome era is not only to understand the functions of the genes and the proteins they code for, but also to understand their roles in the biological pathways, in other words the interactions among them. Proteins are the basic components of biological systems and most of the time they achieve their tasks by interacting with other proteins. These interactions might be permanent or transient (Zhang, 2009). An example of a permanent interaction is attaching of

proteins to each other to form a protein complex. On the other hand, transport of proteins across membranes involves transient protein interactions (Phizicky & Fields, 1995). For example, Importin is a type of protein that binds to another protein and transports it from the cytoplasm to the nucleus or vice versa. Post Translational Modifications (PTMs) such as phosphorylation, acetylation, and methylation are other examples of transient interactions. For example, in phosphorylation a protein kinase binds briefly to a target protein and adds a phosphate to the target protein. PTMs can affect the functions and interactions of the proteins in important ways. For instance, a protein might be able to bind to another protein, only if it has been phosphorylated.

Information about protein interactions is crucial for understanding the vital biological processes

Protein-protein interactions (PPIs) play important roles in many, if not all, vital biological processes such as metabolic and signaling pathways, cell cycle control, cell growth, and cell death (Phizicky & Fields, 1995). Information about these interactions is crucial not only for understanding these biological processes, but also for improving our understanding of diseases and developing approaches for their prevention and cure. As an example, consider apoptosis, which is the process of genetically controlled cell death. It plays an important role in tissue and organ development, which comprises the division and differentiation of a particular cell, and the apoptosis of the unwanted cells. For instance, the differentiation of the fingers in a developing human embryo is a result of the apoptosis of the cells between the fingers. While apoptosis is an important biological phenomenon in an organism's life cycle, defective apoptotic processes have been associated with various diseases. For example, damage in the apoptotic capabilities of cells might result in uncontrolled cell

proliferation, such as cancer. Figure 1.3 shows the pathway of apoptosis from KEGG[5] (Kanehisa & Goto, 2000; Kanehisa et al., 2006, 2010), which is a manually drawn pathway map of the currently known molecular interaction and reaction network for apoptosis. A chain of bio-molecular events that lead to apoptosis is outlined below.

1. Caspase-8 (CASP8) activation is triggered by death receptor engagement.
2. CASP8 activation regulates caspase-3 (CASP3) activation.
3. CASP3 activation leads to the degradation of cellular proteins that are required for cell survival.
4. Degradation of cellular proteins that are required for cell survival causes apoptosis.

Figure 1.3 shows that the Inhibitor of Apoptosis (IAP) family of proteins inhibit CASP3, thus suppressing apoptosis. Information about these interactions can help the development of strategies to treat insufficient amount of apoptosis (e.g. in cancer). For example, one approach could be identifying mechanisms that can inhibit IAP, thus preventing it from suppressing apoptosis (Danson et al., 2007; Fulda, 2008).

Size of protein interaction networks correlates with the biological complexities of the organisms

One of the most surprising results of the genome sequencing projects (Lander et al., 2001; Venter et al., 2001) was that the number of genes in organisms does not necessarily reflect their organismal complexity, which is often measured with the number of distinct cell types (Copley, 2008). It is interesting that the number of protein-coding genes in humans, which has been estimated as 20500 (Clamp et al., 2007), is similar to the number of protein-coding genes in the nematode Caenorhab-

[5] http://www.genome.jp/kegg/pathway/hsa/hsa04210.html (Accessed on October 4, 2009.)

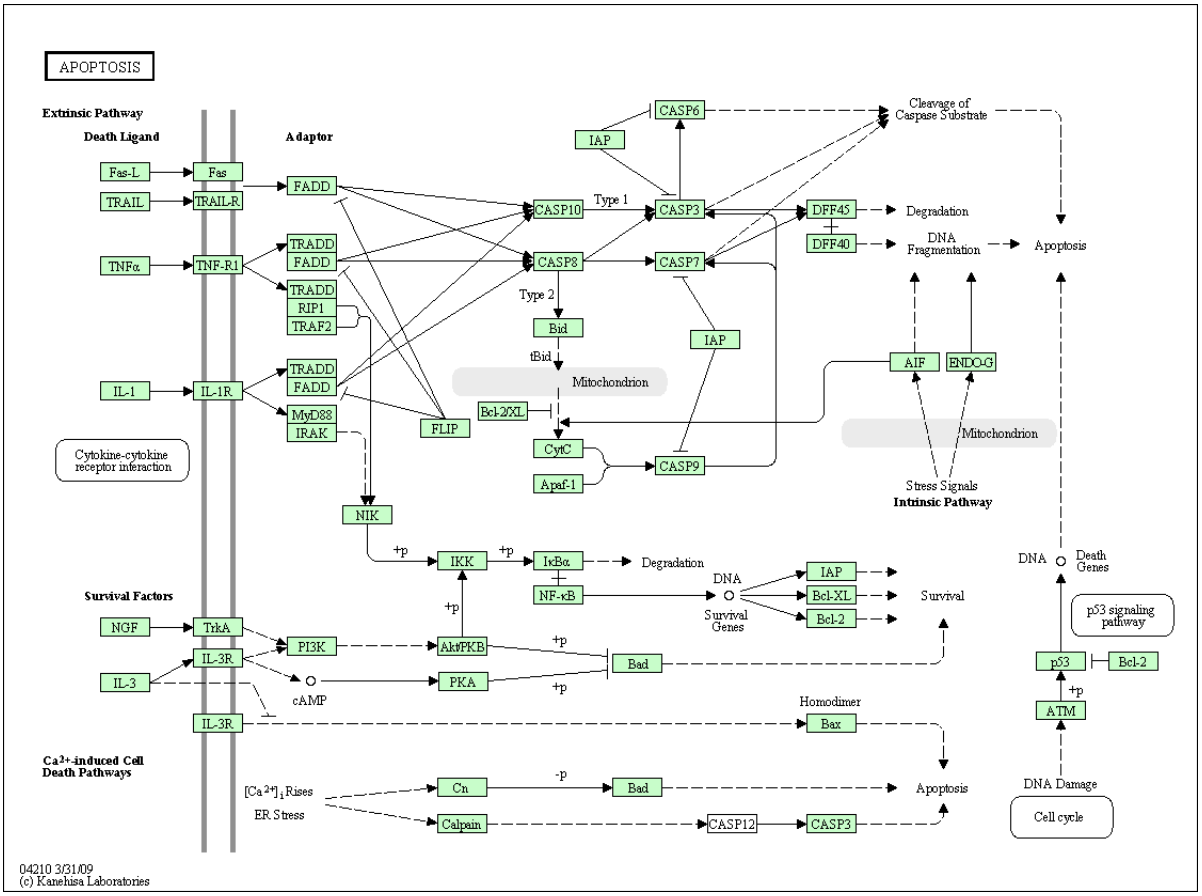


Figure 1.3: Apoptosis pathway from KEGG (http://www.genome.jp/kegg/pathway/hsa/hsa04210.html), which shows the map of the currently known molecular interaction and reaction network for apoptosis (Kanehisa & Goto, 2000; Kanehisa et al., 2006, 2010). The image is included with the permission of the GenomeNet team.

ditis elegans, which has been estimated as 19735 (Hillier et al., 2005). It has been suggested that the biological complexities of organisms does not only depend on the number of genes that they have, but on the number of bio-molecular interactions (Lander et al., 2001; Copley, 2008). A recent study has shown that the size of the protein interaction networks correlates with the biological complexities of the organisms (Stumpf et al., 2008). The number of protein interactions in humans has been estimated as 650000, which is around ten times more than the estimated number of interactions in fruit fly (Drosophila melanogaster) and around three times more than the estimated number of interactions in Caenorhabditis elegans (Stumpf et al.,

2008).

Many protein interactions are available only in the text of published scientific articles

The development of high-throughput experimental methods such as two-hybrid system, mass spectrometry, and protein chip technology has led to rapid increase in data and publications relevant to PPI (Zhang, 2009). A number of mostly manually curated databases that store PPI data in structured formats have been developed as a response. However, these databases cover only a small fraction of the available PPI data. For example, one of the most comprehensive databases for human protein interactions is the HPRD database (Keshava Prasad et al., 2009), which currently contains 38806 PPIs[6]. This number is much smaller than the total estimated number of human PPIs of 650000 (Stumpf et al., 2008). Although there might be many PPIs that are not uncovered yet, a recent study by Ramani et al. (2005), has demonstrated that there is only a small overlap ($< 0.1\%$) between the PPIs reported in the manually curated HPRD (Keshava Prasad et al., 2009), Reactome (Matthews et al., 2009), and BIND (Bader et al., 2003) databases. This suggests that, the number of PPIs available in the literature is much larger than the ones existing in the manually curated databases.

PPI networks can be represented as graphs

A PPI network can be represented as a graph, where the proteins are represented as nodes and an interaction between a pair of proteins is represented with an edge connecting them. Figure 1.4 shows a sample PPI network created by using the VisANT[7] online visualization and analysis tool for biological interaction data (Hu et al., 2004). VisANT is supported by the Predictome database (Mellor et al.,

[6]http://www.hprd.org/ (Accessed on October 5, 2009.)
[7]http://visant.bu.edu/

2002), which stores interactions that are identified using experimental techniques (e.g. yeast two-hybrid system, coimmunoprecipitation, mass spectrometry) or compuational techniques (e.g. gene fusion, chromosomal proximity, gene co-evolution). The Predictome databases also integrates experimentally determined interactions from curated databases such as MINT (Zanzoni et al., 2002), BIND (Bader et al., 2003), and HPRD (Keshava Prasad et al., 2009). The network in Figure 1.4 shows the human protein interactions determined by the yeast two hybrid method. The network consists of 7314 nodes and 19443 edges.

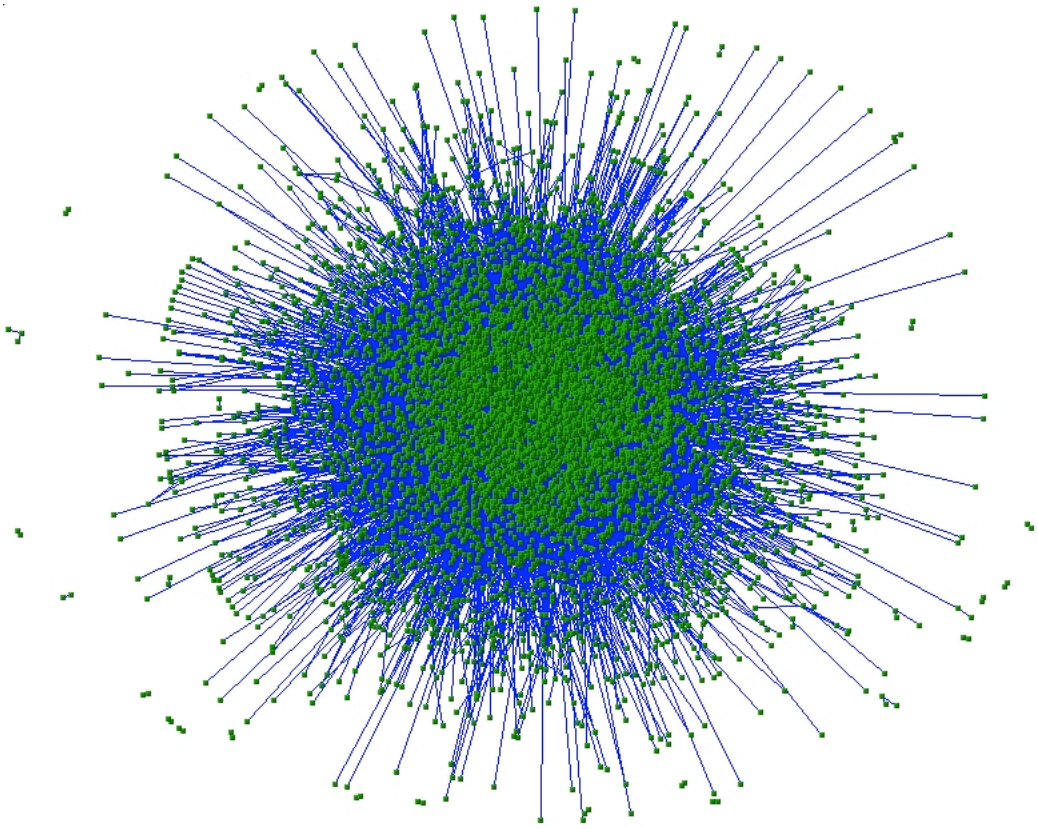


Figure 1.4: Network of human protein interactions, which were derived by using the yeast two hybrid method. The network is created with VisANT (http://visant.bu.edu/) (Hu et al., 2004). The picture is included with the permission of the authors.

The graph representation allows the analysis of PPIs from a graph theory and complex networks perspective, which can give biologists a variety of new insights.

For example, the function of a protein may be predicted by looking at the proteins with which it interacts, as it is generally assumed that neighboring proteins in the network have common functions (Schwikowski et al., 2000; Zhang, 2009). Similarly, subgraphs that are densely connected within themselves but sparsely connected with the rest of the network might form molecular modules that function as a unit in certain biological processes (Spirin & Mirny, 2003; Zhang, 2009).

Topological features of PPI networks can also facilitate our understanding of biological systems. Recent studies have shown that PPI networks share similar topological properties such as being small-world and scale-free (Chen & Sharp, 2004; Hoffmann & Valencia, 2005; Jeong et al., 2001), with each other and with various non-biological complex systems such as the WWW (Huberman & Adamic, 1999), the Internet (Yook et al., 2002), and social networks (Barabási et al., 2002; Watts & Strogatz, 1998). A scale-free network is characterized by having a power-law degree distribution, $P(k) \sim k^{-\gamma}$, where $P(k)$ is the probability that a randomly selected node will have a degree (i.e. number of connections) of k (Albert & Barabási, 2002). In scale-free networks most nodes make only a few connections, while a small set of nodes (known as hubs) have very large number of links. This is different from random networks, which follow Poisson distribution, where majority of the nodes have degrees close to the average degree of the network. Small-world networks have a relatively short distance between any two nodes, where distance is defined as the number of edges along the shortest path connecting them, and a clustering coefficient that is significantly higher than that of a random network with the same number of nodes. The clustering coefficient of a node describes how well connected a node's neighbors are and is defined as the number of connections between this node's neighbors divided by the number of possible connections between them (Watts & Strogatz,

1998). The clustering coefficient of a network is the average of the clustering coefficients of the nodes in the network (Watts & Strogatz, 1998). The small-world phenomenon was first studied as a concept in sociology. The most popular example of small world networks is the "six-degree-of-separation" concept uncovered by the social psychologist Stanley Milgram in 1967, where he concluded that there is a path of acquaintances with a typical length of six between most pairs of people in the United States (Milgram, 1967).

1.2.2 Biomedical Information Extraction

This section provides a brief background on biomedical information extraction. A current survey on information extraction in general can be found in (Sarawagi, 2008) and surveys on biomedical information extraction and text mining can be found in (Cohen & Hersh, 2005; Zweigenbaum et al., 2007).

The goal of information extraction (IE) is to automatically extract the explicitly stated factual information in structured format from unstructured text. Conceptually, IE consists of two main components, named entity recognition (NER) and relation extraction. NER is the task of identifying the mentions of specific types of entities. Once the named entities have been identified, the next step in IE is relation extraction that involves the extraction of predefined types of relations among these entities. NER and relation extraction are usually targeted to specific domains. The named entities and the relevant relations depend on the domain of focus. For example, names of *persons*, *organizations*, and *locations* are types of named entities and *company employee* and *company acquisition* are types of relations considered in the newspaper articles domain. Consider the sentence *"Medical Corp said it named Victor Vaguine as president and chief executive officer"*, which is taken from a newspaper article. A relation that can be extracted from this sentence is

the employee_of(PERSON, ORGANIZATION) relation, where *"Victor Vaguine"* is a person who is an employee of the *"Medical Corp"* organization.

Our focus is on the biomedical domain, which has a very specialized language with complex and continually changing terminology. What makes the IE task even more challenging is the abundance of sentences with complex structures. Gene and protein names, diseases, drugs, metabolites, cellular components, cell and tissue types are examples of named entity types encountered in the biomedical domain. While the earliest systems for biomedical NER were usually based on rule-based approaches (e.g. (Fukuda et al., 1998)), as annotated corpora became available machine-learning based methods have recently gained popularity (e.g. (GuoDong & Jian, 2004; Zhao, 2004; Mcdonald & Pereira, 2005; Tsai et al., 2006; Hsu et al., 2008)). State-of-the-art gene and protein NER systems achieve a practically applicable level of performance (e.g. 87% F-score performance was obtained at the second BioCreative shared task on gene mention tagging (Smith et al., 2008)). Genia Tagger (Tsuruoka et al., 2005), ABNER (Settles, 2005), and BANNER (Leaman & Gonzalez, 2008) are some of the publicly available biomedical NER tools.

In this thesis we focus on extracting relations and either use the gold standard names for the entities when available or use one of the available NER tools to identify the named entities. While referring to genes and proteins, we usually adapt the commonly applied GENETAG-style named entity annotation, which does not differentiate between genes and proteins (Tanabe et al., 2005). In other words, the terms gene and protein are usually used interchangeably to refer to the genes and gene products.

Types of relations that have been targeted in the biomedical domain include gene-disease associations (Gonzalez et al., 2007; Chen et al., 2006; Adamic et al., 2002; Al-

Mubaid & Singh, 2005), protein localizations (Craven & Kumlien, 1999), gene-drug interactions (Rindflesch et al., 2000), and disease-treatment relationships (Rosario & Hearst, 2004). The relation extraction problem that has drawn the most attention in the biomedical IE field is the recognition of protein-protein interactions (PPI) (Jelier et al., 2005; Blaschke et al., 1999; Ono et al., 2001; Temkin & Gilder, 2003; Daraselia et al., 2004; Fundel et al., 2007; Mitsumori et al., 2006; Bunescu & Mooney, 2007). The PPI extraction problem is typically formulated as extracting the binary relationships between the proteins from a given biomedical text. The goal is to identify the pairs of proteins that are stated to interact with each other in the text. Most PPI extraction systems operate on a sentence-level to extract the interactions. The underlying assumption is that the majority of the relations are contained within a single sentence. Analysis of the Genia event corpus (Kim et al., 2008, 2009) supports this assumption, since only 5% of the relations in the corpus span multiple sentences (Björne et al., 2009).

Figure 1.5 displays a sample biomedical abstract (Sato et al., 2005) with all protein names shown in blue. Consider the sentence *"ZIPK specifically interacted with STAT3, and did not bind to STAT1, STAT4, STAT5a, STAT5b or STAT6."*. There are seven proteins in this sentence, which means there are $\binom{7}{2} = 21$ difference protein pairs. The sentence states an interaction only between one pair (i.e., *ZIPK-STAT3* pair). The last sentence in the abstract *"Taken together, our data suggest that ZIPK interacts with STAT3 within the nucleus to regulate the transcriptional activity of STAT3 via phosphorylation of Ser727."*, also provides an interesting example. The speculation keyword *"suggest"* renders the statement *"ZIPK interacts with STAT3 within the nucleus to regulate the transcriptional activity of STAT3 via phosphorylation of Ser727"* speculative, conveying that the authors are not completely certain

Physical and functional interactions between STAT3 and ZIP kinase.

Signal transducer and activator of transcription 3 (STAT3) is a latent cytoplasmic transcription factor that can be activated by cytokines and growth factors. It plays important roles in cell growth, apoptosis and cell transformation, and is constitutively active in a variety of tumor cells. In this study, we provide evidence that zipper-interacting protein kinase (ZIPK) interacts physically with STAT3. ZIPK specifically interacted with STAT3, and did not bind to STAT1, STAT4, STAT5a, STAT5b or STAT6. ZIPK phosphorylated STAT3 on serine 727 (Ser727) and enhanced STAT3 transcriptional activity. Small interfering RNA-mediated reduction of ZIPK expression decreased leukemia inhibitory factor (LIF)- and IL-6-induced STAT3-dependent transcription. Furthermore, LIF- and IL-6-mediated STAT3 activation stimulated ZIPK activity. Taken together, our data suggest that ZIPK interacts with STAT3 within the nucleus to regulate the transcriptional activity of STAT3 via phosphorylation of Ser727.

Figure 1.5: A sample biomedical abstract with all protein names shown in blue.

about the inferred conclusion. Even though speculations are frequently occurring language phenomena that modify the factuality of the information contained in text, they have been neglected by most information extraction systems. While speculative information might still be useful, it is important that it is distinguished from factual information. As protein interaction extraction and speculation detection are the focus of Chapters II and III, respectively, we will discuss the related work in more detail in these corresponding chapters.

1.2.3 Literature-Based Discovery

Relationship extraction targets extracting the explicitly stated relationships between entities from text. Literature-based discovery (LBD) on the other hand, aims to go beyond that, and use the extracted information to infer new (implicit) relationships. These new relationships can be proposed as potential scientific hypotheses, which can be verified by further (experimental) studies. There is no an immediately available ground truth for potentially currently unknown knowledge. This makes evaluating LBD systems difficult. One strategy to test LBD systems is trying to

replicate some of the already known discoveries. Another approach is manually reviewing the literature to find evidence supporting the generated hypotheses.

The idea of discovering new relations from bibliographic databases of scientific literature was introduced by Swanson, who proposed a LBD model that is based on connecting concepts using logical inference (Swanson, 1986). The intuition behind this model, which is commonly referred to as *Swanson's ABC* model, is that *"if A is related to B, and B is related to C, then it is likely that A is related to C"*. Swanson used this model to discover various new hypotheses by manually linking concepts between journal articles. For example, literature related to fish oil, provided the information that fish oil lowers blood viscosity, inhibits platelet aggregation, and causes vascular reactivity. Raynauds disease related literature contained the information that patients with Raynauds disease have impaired vascular reactivity, high blood viscosity, and high platelet aggregation. By connecting the information from these two disjoint literatures, Swanson hypothesized that fish oil may have beneficial effects in patients with Raynauds disease (Swanson, 1986), two years before clinical trials confirmed its correctness. In another study, Swanson discovered 11 indirect links between migraine and magnesium (Swanson, 1988). These connections were later verified experimentally (Ramadan et al., 1989; Ferrari, 1992). Swanson, together with Smalheiser, subsequently contributed several other discoveries including the relationships between *magnesium deficiency – neurologic disease* (Smalheiser & Swanson, 1994), *Estrogen – Alzheimers Disease* (Smalheiser & Swanson, 1996b), *Indomethacin – Alzheimers Disease* (Smalheiser & Swanson, 1996a) and *Calcium Independent Phospholipase A2 – Schizophrenia* (Smalheiser & Swanson, 1998).

Swanson's original discoveries were based on exhaustively reading the titles and abstracts of articles. Since then, several studies have tried to automate the process

(an overview is presented in (Weeber et al., 2005)). For example, the Arrowsmith system was developed by making use of Swanson's search strategies in his earlier work (Swanson & Smalheiser, 1997). The Arrowsmith system, as well as the many others that followed make use of *Swanson's ABC* model (Gordon & Dumais, 1998; Weeber et al., 2000; Fuller et al., 2004; Hristovski et al., 2005; Srinivasan, 2004; Lindsay & Gordon, 1999; Yetisgen-Yildiz & Pratt, 2006; Wren, 2004; Swanson et al., 2006; Baker & Hemminger, 2010). The discovery process begins with an *A term* (e.g. *migraine*). A correlation-mining approach, which is typically based on term co-occurrence, is used to identify *B terms* (e.g. *calsium channel blockers*, and *spreading cortical depression*) that are correlated with the *A term*. Some of the different correlation mining approaches that have been used in LBD systems are Association Rules (Hristovski et al., 2005), TF-IDF (Lindsay & Gordon, 1999; Srinivasan, 2004), Z-score (Yetisgen-Yildiz & Pratt, 2006), and Mutual Information (Wren, 2004). After identifying the *B terms*, the same correlation mining technique is used to detect the *C terms* that are correlated with the *B terms*. The *C terms* (e.g. *magnesium*) are the potential new discoveries that are related to the *A term*. In general, a large number of *C terms* is produced. So, a ranking approach (e.g. *B term* count (Yetisgen-Yildiz & Pratt, 2006) and literature cohesiveness (Swanson et al., 2006)) is used to order the discovered *C terms*.

While most LBD studies target the biomedical domain, there have been a few that focus on other domains. Gordon et al. (2002) apply *Swanson's ABC* model on the World Wide Web to discover novel applications for existing problem solutions. For example, they use "genetic algorithms" as their *A term* and discover many potential fields of application such as "virtual reality", "computer graphics", and "fluid dynamics". Cory (1997) applied LBD on online humanities databases to discover

hidden analogies.

1.3 Guide to Remaining Chapters

Chapters II and III describe our work towards meeting the first goal of this thesis, i.e., developing natural language processing and machine learning based methods to extract information about proteins, genes, and their interactions from biomedical text. Our work related to the second goal of this thesis, which is using the extracted information to generate new scientific hypothesis, is presented in Chapters III and IV. Most of what follows is published work. Below is a summary of the remaining chapters with references to the relevant publications.

- **Chapter II:** We introduce a kernel based relation extraction method to identify the interacting protein pairs in sentences. Our approach is based on making use of the shortest path between a protein pair in the dependency parse tree of the corresponding sentence. Our motivating assumption is that this path is a good representation of the semantic relationship between the protein pair. We propose two separate kernel functions based on cosine similarity and edit distance among the dependency parse tree paths connecting the protein names. Using these kernel functions, we investigate the performances of the supervised machine learning algorithms Support Vector Machines and k-nearest-neighbor, and their semi-supervised counterparts, transductive SVMs and harmonic functions. We report significant improvement over the previous results in the literature. Chapter II is based on the work published as (Erkan, Özgür, & Radev, 2007a, 2007b). Using the dependency path edit kernel we also contributed to the BioCreaive Meta-Server by annotating abstracts as containing protein interactions or not (Leitner et al., 2008)[8].

[8] http://bcms.bioinfo.cnio.es/

- **Chapter III:** While speculative information might still be useful for scientists, it is important that it is distinguished from factual information. For example, identifying whether a protein interaction that is extracted from an article has been reported with a speculative language rather than being reported as a fact is an important context information regarding that extracted information. Most previous studies on speculation detection focus on identifying speculative sentences. However, in many cases, not the entire sentence, but fragments of a sentence are speculative. We propose an approach based on solving two sub-problems to identify speculative fragments of sentences. The first sub-problem is identifying the speculation keywords in the sentences and the second one is resolving their linguistic scopes. We tackle the first sub-problem as a supervised classification task, where we classify the potential keywords as real speculation keywords or not. We investigate using a diverse set of linguistic features that represent the contexts of the keywords with the Support Vector Machines classifier. To determine the scopes of the keywords, we develop a rule-based method using the part of speech tags of the keywords and the syntactic parse trees of the sentences. Chapter III is published as (Özgür & Radev, 2009).
- **Chapter IV:** We propose a literature-based discovery (LBD) approach to generate new scientific hypothesis using the known gene-gene relationships automatically extracted from the literature. Unlike most previous LBD methods that are based on *Swanson's ABC* model and depend on co-occurrence information among the entities, our approach integrates text mining with network analysis in a novel way. We start with a set of genes (seed genes) known to be related to a concept and extract the interactions of these genes from the literature using the natural language processing and machine learning based method

introduced in Chapter II. Analyzing the concept-specific literature-mined network using graph centrality metrics enables us to infer novel genes that are likely to be related to the concept of interest. In this chapter we present the application of this approach to identify gene-disease associations. We use full text articles from PubMed Central Open Access[9] and a set of 15 genes known to be related to prostate cancer and show the effectiveness of our LBD method in predicting new genes related to the disease. Chapter IV is based on the work published as (Özgür, Vu, Erkan, & Radev, 2008).

- **Chapter V:** We adapt the LBD method that we introduced in Chapter IV to discover genes potentially important for vaccine development research. We use only one seed gene, i.e., human interferon-gamma (IFNG), as a gene known to be critical for vaccine induced protective immunity and analyze all the article abstracts available in PubMed. We build two gene interaction networks. The first network is the generic IFNG network, which is the network of interactions of IFNG and its neighbors. We use the concept in which we are interested, i.e., the term "vaccine", to create the second network. The second network is a sub-graph of the first one, which only consists of the gene interactions extracted from sentence that contain the term "vaccine" or its variants. Analyzing these two networks from graph centrality perspective and comparing them enabled us to identify several genes that are good candidates for further IFNG and vaccine development studies. We also investigated incorporating ontology support to our LBD method. Integrating the Vaccine Ontology[10] led to the discovery of vaccine related genes, which we were not able to discover without ontology support. Chapter V is based on the work published as (Özgür, Xiang, Radev,

[9] http://www.ncbi.nlm.nih.gov/pmc/about/openftlist.html
[10] http://www.violinet.org/vaccineontology/

& He, 2010; Özgür, Radev, & He, 2010).

- **Chapter VI:** We conclude by summarizing the main contributions and the future directions of our research.

- **Appendix:** We describe two tools, i.e. GIN-IE (Gene INteraction - Information Extraction) and GIN-NA (Gene INteraction - Network Analysis) that are by-products of this research. State-of-the-art machine learning based approaches for relationship extraction usually achieve high F-Measure performances. However, high F-Measure performance of a system does not necessarily reflect its usability to the end users. Most real-life applications require high precision, even if it comes at the expense of recall. We developed a high-precision interaction extraction system (GIN-IE) based on rules defined on the dependency parse trees of the sentences. GIN-IE has been integrated with the Michigan Molecular Interactions (MiMI) database[11] and made available to the end users (e.g. biomedical scientists) (Tarcea et al., 2009). GIN-NA is designed to support literature-based discovery of genes related to a concept. Given a gene or a list of genes known to be related to a concept, GIN-NA retrieves their interaction network from MiMI and provides an analysis of this network as well as its most central genes.

[11] http://mimi.ncibi.org/

CHAPTER II

Dependency Parsing and Machine Learning for Extracting Protein Interactions from Biomedical Text

2.1 Introduction

Protein-protein interactions (PPIs) play critical roles in vital biological processes such as metabolic and signaling pathways, cell cycle control, and DNA replication and transcription (Phizicky & Fields, 1995). PPI information is crucial for understanding these processes. The manual construction of databases such as MINT (Zanzoni et al., 2002), BIND (Bader et al., 2003), and HPRD (Keshava Prasad et al., 2009) that store PPI information in structured and standard formats, is very time-consuming and labor-intensive. As a consequence, most PPI information is only available in the text of published articles. Therefore, the automatic extraction of PPI information from free texts has become an important research area in Natural Language Processing for Biology (BioNLP).

We introduce an information extraction approach to identify sentences in text that indicate an interaction relation between two proteins. Our method is different than most of the previous studies (see Section 2.2) on this problem in two aspects: First, we generate the dependency parses of the sentences that we analyze, making use of the dependency relationships among the words. This enables us to make more syntax-aware inferences about the roles of the proteins in a sentence compared

to the classical pattern-matching information extraction methods. We propose two kernel functions based on the dependency parse trees of the sentences. Second, we investigate semi-supervised machine learning methods on top of the dependency features we generate. Although there have been a number of learning-based studies in this domain, our methods are the first semi-supervised efforts to our knowledge. The high cost of labeling free text for this problem makes semi-supervised methods particularly valuable.

We focus on two semi-supervised learning methods: transductive SVMs (TSVM) (Joachims, 1999b), and harmonic functions (Zhu et al., 2003). We also compare these two methods with their supervised counterparts, namely SVMs and k-nearest neighbor (kNN) algorithm. Because of the nature of these algorithms, we propose two similarity functions (*kernels* in SVM terminology) among the instances of the learning problem. The instances in this problem are natural language sentences with protein names in them, and the similarity functions are defined on the positions of the protein names in the corresponding parse trees. Our motivating assumption is that the *path* between two protein names in a dependency tree is a good description of the semantic relation between them in the corresponding sentence. We propose two kernel functions; one based on the cosine similarity and the other based on the edit distance among such paths. This chapter is based on the work published as (Erkan, Özgür, & Radev, 2007a, 2007b).

2.2 Related Work

There have been many approaches to extract protein interactions from free text. The simplest approach is using the co-occurrence statistics of the proteins in text (Jelier et al., 2005). The underlying assumption is that whenever two (or more) entities

are mentioned together in text, there is a semantic relationship between them. However, this does not necessarily mean that the two entities interact. As a consequence, this approach can provide high recall, but usually suffers from low precision.

Another common approach is based on matching pre-specified patterns and rules based on the sequence of words and/or their parts of speech in the sentences (Blaschke et al., 1999; Ono et al., 2001; Blaschke & Valencia, 2002). However, complex cases that are not covered by the pre-defined patterns and rules cannot be extracted by these methods. Huang et al. (2004) proposed a method where patterns are discovered automatically from a set of sentences by dynamic programming. Bunescu et al. (2005) have studied the performance of rule learning algorithms. They propose two methods for protein interaction extraction. One is based on the rule learning method Rapier and the other on longest common subsequences. They show that these methods outperform hand-written rules.

Another class of approaches is using more syntax-aware natural language processing (NLP) techniques. Both full and partial (shallow) parsing strategies have been applied in the literature. In partial parsing the sentence structure is decomposed partially and local dependencies between certain phrasal components are extracted. An example of the application of this method is relational parsing for the *inhibition* relation (Pustejovsky et al., 2002). In full parsing, however, the full sentence structure is taken into account. Temkin and Gilder (2003) used a full parser with a lexical analyzer and a context free grammar (CFG) to extract protein-protein interaction from text. Another study that uses full-sentence parsing to extract human protein interactions is (Daraselia et al., 2004). Alternatively, Yakushiji et al. (2005) propose a system based on head-driven phrase structure grammar (HPSG). In their system protein interaction expressions are presented as predicate argument structure

patterns from the HPSG parser. These parsing approaches consider only syntactic properties of the sentences and do not take into account semantic properties. Thus, although they are complicated and require many resources, their performance is not satisfactory.

Machine learning techniques for extracting protein interaction information have gained interest in the recent years. The PreBIND system uses SVM to identify the existence of protein interactions in abstracts and uses this type of information to enhance manual expert reviewing for the BIND database (Donaldson et al., 2003). Words and word bigrams are used as binary features. This system is also tested with the Naive Bayes classifier, but SVM is reported to perform better. Mitsumori et al. (2006) also use SVM to extract protein-protein interactions. They use bag-of-words features, specifically the words around the protein names. These systems do not use any syntactic or semantic information. Sugiyama et al. (2003) extract features from the sentences based on the verbs and nouns in the sentences such as the verbal forms, and the part of speech tags of the 20 words surrounding the verb (10 before and 10 after it). Further features are used to indicate whether a noun is found, as well as the part of speech tags for the 20 words surrounding the noun, and whether the noun contains numerical characters, non-alpha characters, or uppercase letters. They construct k-nearest neighbor, decision tree, neural network, and SVM classifiers by using these features. They report that the SVM classifier performs the best. They use part-of-speech information, but do not consider any dependency or semantic information.

For relation extraction in the newswire domain syntactic parse trees augmented with semantic labels have been used in a generative model (Miller et al., 2000). A fairly new class of algorithms that have also been used for relation extraction are

kernel-based methods. Kernel functions based on the shallow, syntactic, and dependency parses of the sentences have been proposed (Collins & Duffy, 2001; Zelenko et al., 2003; Culotta & Sorensen, 2004; Bunescu & Mooney, 2005a; Moschitti, 2006). Although machine learning methods with features extracted from the syntactic or dependency parse trees of the sentences have been successfully applied for relation extraction in the newswire domain, this approach is relatively new in the biomedical domain.

2.3 Sentence Similarity Based on Dependency Parsing

In order to apply the semi-supervised harmonic functions and its supervised counterpart kNN, and the kernel based TSVM and SVM methods, we need to define a similarity measure between two sentences. For this purpose, we use the dependency parse trees of the sentences. Unlike a syntactic parse (which describes the syntactic constituent structure of a sentence), the dependency parse of a sentence captures the semantic predicate-argument relationships among its words. The nodes of a dependency parse tree represent the words of a sentence and the edges represent the types of the dependencies among the words such as subject, object and modifier (Figure 2.1). We define the similarity between two sentences based on the paths between two proteins in the dependency parse trees of the sentences.

In this study we assume that the protein names have already been annotated and focus instead on the task of extracting protein-protein interaction sentences for a given protein pair. We parse the sentences with the Stanford Parser[1] (de Marneffe et al., 2006). From the dependency parse trees of each sentence we extract the shortest path between a protein pair.

Figure 2.1 shows the dependency tree we got for the sentence "*The results demon-*

[1] http://nlp.stanford.edu/software/lex-parser.shtml

strated that KaiC interacts rhythmically with KaiA, KaiB, and SasA." This example

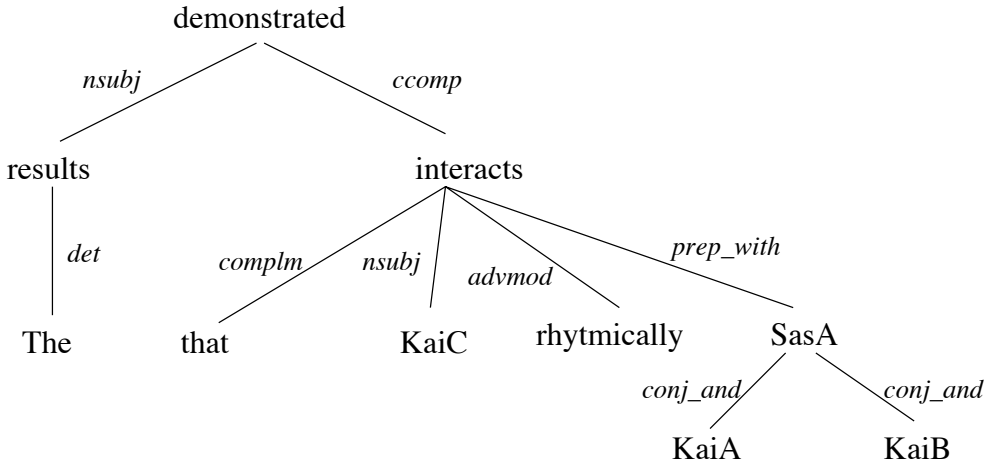


Figure 2.1: The dependency tree of the sentence "*The results demonstrated that KaiC interacts rhythmically with KaiA, KaiB, and SasA*."

sentence illustrates that the dependency path between a protein pair captures the relevant information regarding the relationship between the proteins better compared to using the words in the unparsed sentence. Consider the protein pair KaiC and SasA. The words in the sentence between these proteins are *interacts, rhythmically, with, KaiA, KaiB,* and *and*. Among these words *rhythmically, KaiA, and* and *KaiB* are not directly related to the interaction relationship between KaiC and SasA. On the other hand, the words in the dependency path between this protein pair (i.e., *nsubj, interacts,* and *prep_with*) give sufficient information to identify their relationship.

In this sentence we have four proteins (KaiC, KaiA, KaiB, and SasA). So there are six pairs of proteins for which the sentence may or may not be describing an interaction. The following are the paths between the six protein pairs.

1. KaiC – nsubj – interacts – prep_with – SasA
2. KaiC – nsubj – interacts – prep_with – SasA – conj_and – KaiA
3. KaiC – nsubj – interacts – prep_with – SasA – conj_and – KaiB
4. SasA – conj_and – KaiA

5. SasA – conj_and – KaiB
6. KaiA – conj_and – SasA – conj_and – KaiB

In this example there is a single path between each protein pair. However, there may be more than one paths between a protein pair, if one or both appear multiple times in the sentence. In such cases, we select the shortest paths between the protein pairs.

If a sentence contains n different proteins, there are $\binom{n}{2}$ different pairs of proteins. We use machine learning approaches to classify each sentence as an interaction sentence or not for a protein pair. A sentence may be an interaction sentence for one protein pair, while not for another protein pair. For instance, our example sentence is a positive interaction sentence for the KaiC and SasA protein pair. However, it is a negative interaction sentence for the KaiA and SasA protein pair, i.e., it does not describe an interaction between this pair of proteins. Thus, before parsing a sentence, we make multiple copies of it, one for each protein pair. To reduce data sparseness, we rename the proteins in the pair as *PROTX1* and *PROTX2*, and all the other proteins in the sentence as *PROTX0*. So, for our example sentence we have the following instances in the training set:

1. *PROTX1* – nsubj – interacts – prep_with – *PROTX2*
2. *PROTX1* – nsubj – interacts – prep_with – *PROTX0* – conj_and – *PROTX2*
3. *PROTX1* – nsubj – interacts – prep_with – *PROTX0* – conj_and – *PROTX2*
4. *PROTX1* – conj_and – *PROTX2*
5. *PROTX1* – conj_and – *PROTX2*
6. *PROTX1* – conj_and – *PROTX0* – conj_and – *PROTX2*

The first three instances are positive as they describe an interaction between *PROTX1* and *PROTX2*. The last three are negative, as they do not describe an

interaction between *PROTX1* and *PROTX2*.

We propose two kernel functions to use with the machine learning algorithms. The first one is based on the cosine similarity and the second one is based on the edit distance between the dependency tree path representations of the instances. Our underlying assumption is that, the more similar two dependency tree paths are, the more likely they belong to the same class; that is, either both describe or both do not describe an interaction for the corresponding protein pair.

2.3.1 Dependency Path Cosine Kernel

Suppose p_i and p_j are the paths between *PROTX1* and *PROTX2* in instance x_i and instance x_j, respectively. We represent p_i and p_j as vectors of term frequencies in the vector-space model. The cosine similarity measure is the cosine of the angle between these two vectors and is calculated as follows:

$$cos\_sim(p_i, p_j) = cos(\mathbf{p_i}, \mathbf{p_j}) = \frac{\mathbf{p_i} \bullet \mathbf{p_j}}{\|\mathbf{p_i}\|\|\mathbf{p_j}\|} \tag{2.1}$$

that is, it is the dot product of $\mathbf{p_i}$ and $\mathbf{p_j}$ divided by the lengths of $\mathbf{p_i}$ and $\mathbf{p_j}$. The cosine similarity measure takes values in the range $[0, 1]$. If all the terms in p_i and p_j are common, then it takes the maximum value of 1. If none of the terms are common, then it takes the minimum value of 0.

2.3.2 Dependency Path Edit Kernel

A shortcoming of cosine similarity is that it only takes into account the common terms, but does not consider their order in the path. For this reason, we also use a similarity measure based on edit distance (also called Levenshtein distance). Edit distance between two strings is the minimum number of operations that have to be performed to transform the first string to the second. In the original character-based edit distance there are three types of operations. These are insertion, deletion,

or substitution of a single character. We modify the character-based edit distance into a word-based one, where the operations are defined as insertion, deletion, or substitution of a single word.

The edit distance between path 1 and path 2 of our example sentence is 2. We need to perform two insertion operations (i.e., insert *PROTX0* and *conj_and*) to path 1 to transform it to path 2.

1. *PROTX1* – nsubj – interacts – prep_with – **insert (*PROTX0*)** – **insert (*conj_and*)** – *PROTX2*
2. *PROTX1* – nsubj – interacts – prep_with – *PROTX0* – conj_and – *PROTX2*

We normalize edit distance by dividing it by the length (number of words) of the longer path, so that it takes values in the range $[0, 1]$. We convert the distance measure into a similarity measure (kernel function) as follows.

$$edit\_sim(p_i, p_j) = e^{-\gamma(edit\_distance(p_i, p_j))} \tag{2.2}$$

Bunescu and Mooney (2005a) propose a similar method for relation extraction in general (i.e., in the newswire domain). To extract the relationship between two entities, they design a kernel function that uses the shortest path in the dependency tree between them. They show that their approach outperforms the dependency tree kernel of Culotta and Sorensen (2004), which is based on the subtree that contains the two entities.

Here, we adapt the idea of using the shortest dependency tree paths to the task of identifying protein-protein interaction sentences and propose the path cosine and path edit kernel functions. The kernel function proposed by Bunescu and Mooney (2005a) is based on the number of overlapping words between two paths. When

two paths have different lengths, they assume the similarity between them is zero. On the other hand, our cosine similarity and edit distance based kernel functions can also account for deletions and insertions of words. The shortest path kernel in (Bunescu & Mooney, 2005a) only encodes the information about the direction of the dependencies among the words. The two kernel functions that we define also take into account the dependency relationship types among the words, which could carry important information about the semantic relationship between the entities. Consider *path 1* of the example sentence. The dependency direction between "interacts" and "PROTX1" is *"interacts → PROTX1"*, in other words "interacts" is the governor and "PROTX1" is the dependent. The dependency relationship type is "nsubj", that is "PROTX1" is the *noun subject* of "interacts".

2.4 Supervised and Semi-Supervised Machine Learning Approaches

2.4.1 kNN and Harmonic Functions

When a similarity measure is defined among the instances of a learning problem, a simple and natural choice is to use a nearest neighbor based approach that classifies each instance by looking at the labels of the instances that are most similar to it. Perhaps the simplest and most popular similarity-based learning algorithm is the k-nearest neighbor classification method (kNN).

Suppose L is the set of labeled instances, and U is the set of unlabeled instances in a learning problem. Given an instance $x \in U$, the k nearest neighbors among the labeled instances (i.e., $N_k^L(x)$) are found. The category labels of these neighbors are used to estimate the category label of x. In the traditional approach, the most common category label among the k nearest neighbors is assigned to x. Weighted kNN is a refinement to the traditional approach, where the contribution of each of the k nearest neighbors is weighted according to its similarity to x. The protein

interaction extraction problem that we consider in this study is a binary classification problem. The task is to assign one of the two categories 0 (the instance does not describe an interaction) or 1 (the instance describes an interaction) to the unlabeled instances. Each instance is a dependency path between the proteins in the pair and the similarity function can be one of the functions we have defined in Section 2.3. The weighted kNN equation for a binary classification problem can be written as follows (Erkan, 2007):

$$y(x) = \sum_{z \in N_k^L(x)} \frac{sim(x,z)y(z)}{\sum_{z' \in N_k^L(x)} sim(x,z')} \tag{2.3}$$

$y(z) \in \{0,1\}$ is the label of instance z and $y(x)$ is a real number in the $[0,1]$ interval. The class label of x can be assigned by setting a threshold in this interval. For example, if the threshold is set as 0.5, x is assigned to class 1 if $y(x) > 0.5$ and it is assigned to class 0, otherwise.

Erkan (2007) has shown that a semi-supervised version of the weighted kNN algorithm can be formulated as follows by taking into account both the labeled and unlabeled instances when computing the nearest neighbors of x:

$$y(x) = \sum_{z \in N_k^{L \cup U}(x)} \frac{sim(x,z)y(z)}{\sum_{z' \in N_k^{L \cup U}(x)} sim(x,z')} \tag{2.4}$$

This can be represented as an undirected graph, where each instance $z' \in L \cup U$ is a node and each of the k nearest neighbors of z' are connected to z' with an edge. For each $z \in L$, $y(z)$ is set to 0 or 1 depending on the label of z. For each $x \in U$, $y(x)$ is equal to the average of the $y(z')$ values of its neighbors, where $z' \in N_k^{L \cup U}(x)$. Such a function is called a *harmonic* function and has been shown to exist and have a unique solution (Doyle & Snell, 1984). Harmonic functions were first introduced as a semi-supervised learning method by Zhu et al. (2003). They were shown to be effective in

text clustering and classification (Erkan, 2007). Here we present the first application of the harmonic functions to the problem of recognizing protein interactions in text.

2.4.2 SVM and Transductive SVM

Support vector machines (SVM) is a supervised machine learning approach designed for solving two-class pattern recognition problems. The aim is to find the decision surface that separates the positive and negative labeled training examples of a class with maximum margin (Burges, 1998).

Transductive support vector machines (TSVM) are an extension of SVM, where unlabeled data is used in addition to labeled data. The aim now is to assign labels to the unlabeled data and find a decision surface that separates the positive and negative instances of the original labeled data and the (now labeled) unlabeled data with maximum margin. Intuitively, the unlabeled data pushes the decision boundary away from the dense regions. However, unlike SVM, the optimization problem now is NP-hard (Zhu, 2005). Pointers to studies for approximation algorithms can be found in (Zhu, 2005).

In Section 2.3 we defined the similarity between two instances based on the cosine similarity and the edit distance based similarity between the paths in the instances. Here, we use these path similarity measures as kernels for SVM and TSVM and modify the SVM^{light} package (Joachims, 1999b) by plugging in our two kernel functions.

A well-defined kernel function should be symmetric positive definite. While cosine kernel is well-defined, Cortes et al. (2004) proved that edit kernel is not always positive definite. However, it is possible to make the kernel matrix positive definite by adjusting the γ parameter, which is a positive real number. Li and Jiang (2005) applied the edit kernel to predict initiation sites in eucaryotic mRNAs and obtained improved results compared to polynomial kernel.

2.5 Experimental Results

2.5.1 Data Sets and Evaluation Metrics

One of the problems in the field of protein-protein interaction extraction is that different studies generally use different data sets and evaluation metrics. Thus, it is difficult to compare their results. Bunescu et al. (2005) manually developed the *AIMED* corpus[2] for protein-protein interaction and protein name recognition. They tagged 199 Medline abstracts, obtained from the Database of Interacting Proteins (DIP) (Xenarios et al., 2001) and known to contain protein interactions. This corpus is becoming a standard, as it has been used in the recent studies by (Bunescu et al., 2005; Bunescu & Mooney, 2005b, 2007; Mitsumori et al., 2006; Yakushiji et al., 2005).

In our study we used the *AIMED* corpus and the *CB* (Christine Brun) corpus that is provided as a resource by BioCreAtIvE II (Critical Assessment for Information Extraction in Biology) challenge evaluation[3]. We pre-processed the CB corpus by first annotating the protein names in the corpus automatically and then, refining the annotation manually. As discussed in Section 2.3, we pre-processed both of the data sets as follows. We replicated each sentence for each different protein pair. For n different proteins in a sentence, $\binom{n}{2}$ new sentences are created, as there are that many different pairs of proteins. In each newly created sentence we marked the protein pair considered for interaction as *PROTX1* and *PROTX2*, and all the remaining proteins in the sentence as *PROTX0*. If a sentence describes an interaction between *PROTX1* and *PROTX2*, it is labeled as positive, otherwise it is labeled as negative. The summary of the data sets after pre-processing is displayed in Table 2.1[4]. Since

[2] ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/
[3] http://biocreative.sourceforge.net/biocreative_2.html
[4] The pre-processed data sets are available at http://clair.si.umich.edu/clair/biocreative/datasets/

previous studies that use AIMED corpus perform 10-fold cross-validation. We also performed 10-fold cross-validation in both data sets and report the average results over the runs.

Data Set	Sentences	+ Sentences	- Sentences
AIMED	4026	951	3075
CB	4056	2202	1854

Table 2.1: Data Sets

We use precision, recall, and F-score as our metrics to evaluate the performances of the methods. Precision (π) and recall (ρ) are defined as follows:

$$\pi = \frac{TP}{TP + FP}; \qquad \rho = \frac{TP}{TP + FN} \tag{2.5}$$

Here, TP (True Positives) is the number of sentences classified correctly as positive; FP (False Positives) is the number of negative sentences that are classified as positive incorrectly by the classifier; and FN (False Negatives) is the number of positive sentences that are classified as negative incorrectly by the classifier.
F-score is the harmonic mean of recall and precision.

$$F\text{-}score = \frac{2\pi\rho}{\pi + \rho} \tag{2.6}$$

2.5.2 Results and Discussion

We evaluate and compare the performances of the semi-supervised machine learning approaches (TSVM and harmonic functions) with their supervised counterparts (SVM and kNN) for the task of protein-protein interaction extraction. As discussed in Section 2.3, we use cosine similarity and edit distance based similarity as similarity functions in harmonic functions and kNN, and as kernel functions in TSVM and

SVM. Our instances consist of the shortest paths between the protein pairs in the dependency parse trees of the sentences. In our experiments, we tuned the γ parameter of the edit distance based path similarity function to 4.5 with cross-validation. The results in Table 2.2 and Table 2.3 are obtained with 10-fold cross-validation. We report the average results over the runs.

Table 2.2 shows the results obtained for the AIMED data set. Edit distance based path similarity function performs considerably better than the cosine similarity function with harmonic functions and kNN and usually slightly better with SVM and TSVM. We achieve our best F-score performance of 59.96% with TSVM with edit kernel. While SVM with edit kernel achieves the highest precision of 77.52%, it performs slightly worse than SVM with cosine kernel in terms of F-score measure. TSVM performs slightly better than SVM, both of which perform better than harmonic functions. kNN is the worst performing algorithm for this data set.

In Table 2.2, we also show the results obtained previously in the literature by using the same data set. Yakushiji et al. (2005) use an HPSG parser to produce predicate argument structures. They utilize these structures to automatically construct protein interaction extraction rules. Mitsumori et al. (2006) use SVM with bag of words features to extract protein interaction sentences. Here, we show their best result obtained by using the three words to the left and to the right of the proteins. The most closely related study to ours is the shortest path kernel (SPK) method proposed by Bunescu and Mooney (2005a) (see Section 2.3). They apply this method to the domain of protein-protein interaction extraction in (Bunescu & Mooney, 2007). Here, they also test the methods Extraction Using Longest Common Subsequences (ELCS) (Bunescu et al., 2005) and Subsequence Kernel (SSK) (Bunescu & Mooney, 2005b). We cannot compare our results to theirs directly, because they report their

results as a precision-recall graph. However, the best F-score in their graph seems to be around 0.50 and definitely lower than the best F-scores we have achieved ($\approx$ 0.60). Bunescu and Mooney (2007) also use SVM as their learning method in their SPK approach. Our improved performance with SVM and the shortest path edit and cosine kernel functions might be due to the fact that, unlike the SPK method, these functions use the dependency relationship types among the words on the paths and can also handle paths of different lengths. Besides the overlapping words, path edit kernel also takes into account the word order. Our results show that, SVM, TSVM, and harmonic functions achieve better F-score and recall performances than the previous studies by Yakushiji et al. (2005), Mitsumori et al. (2006), and the SSK and ELCS approaches of Bunescu and Mooney (2007). SVM and TSVM also achieve higher precision scores. Since, Mitsumori et al. (2006) also use SVM in their study, our improved results with SVM confirms our motivation of using dependency paths as features instead of the surface representations of the sentences.

Table 2.3 shows the results we got with the CB data set. The F-score performance with the edit distance based similarity function is always better than that of cosine similarity function for this data set. The difference in performances is considerable for harmonic functions and kNN. Our best F-score is achieved with TSVM with edit kernel (85.22%). TSVM performs slightly better than SVM. When cosine similarity function is used, kNN performs better than harmonic functions. However, when edit distance based similarity is used, harmonic functions achieve better performance. SVM and TSVM perform better than harmonic functions. But, the gap in performance is low when edit distance based similarity is used with harmonic functions.

Semi-supervised approaches are usually more effective when there is less labeled

Method	Precision	Recall	F-Score
SVM-edit	77.52	43.51	55.61
SVM-cos	61.99	54.99	58.09
TSVM-edit	59.59	60.68	59.96
TSVM-cos	58.37	61.19	59.62
Harmonic-edit	44.17	74.20	55.29
Harmonic-cos	36.02	67.65	46.97
kNN-edit	68.77	42.17	52.20
kNN-cos	40.37	49.49	44.36
(Yakushiji et al., 2005)	33.70	33.10	33.40
(Mitsumori et al., 2006)	54.20	42.60	47.70

Table 2.2: Experimental Results – AIMED Data Set

Method	Precision	Recall	F-Score
SVM-edit	85.15	84.79	84.96
SVM-cos	87.83	81.45	84.49
TSVM-edit	85.62	84.89	85.22
TSVM-cos	85.67	84.31	84.96
Harmonic-edit	86.69	80.15	83.26
Harmonic-cos	72.28	70.91	71.56
kNN-edit	72.89	86.95	79.28
kNN-cos	65.42	89.49	75.54

Table 2.3: Experimental Results – CB Data Set

data than unlabeled data, which is usually the case in real applications. To see the effect of semi-supervised approaches we perform experiments by varying the amount of labeled training sentences in the range $[10, 3000]$. For each labeled training set size, sentences are selected randomly among all the sentences, and the remaining sentences are used as the unlabeled test set. The results that we report are the averages over 10 such random runs for each labeled training set size. We report the results for the algorithms when edit distance based similarity is used, as it mostly performs better than cosine similarity.

Figure 2.2 shows the results obtained over the AIMED data set. Semi-supervised approaches TSVM and harmonic functions perform considerably better than their supervised counterparts SVM and kNN when we have small number of labeled training data. It is interesting to note that, although SVM is one of the best performing

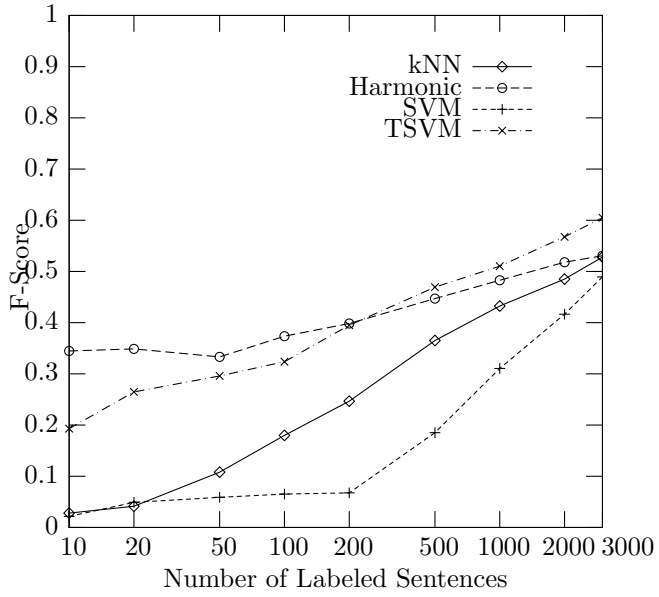


Figure 2.2: The F-score on the AIMED dataset with varying sizes of training data

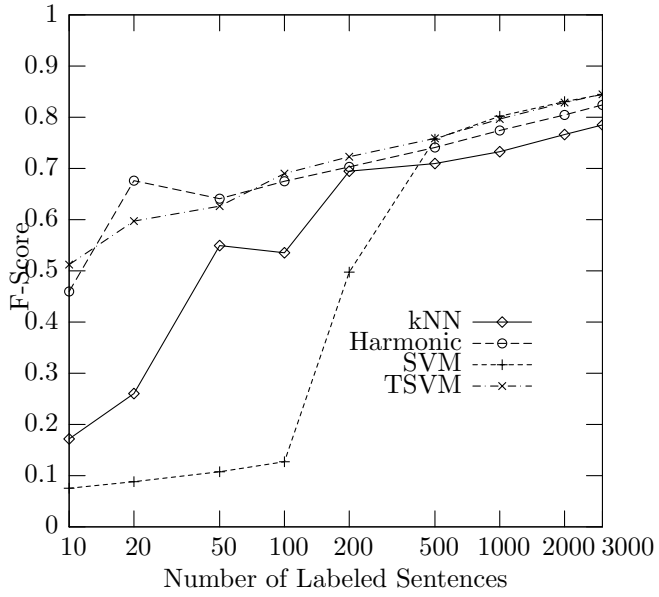


Figure 2.3: The F-score on the CB dataset with varying sizes of training data

algorithms with more training data, it is the worst performing algorithm with small amount of labeled training sentences. Its performance starts to increase when number of training data is larger than 200. Eventually, its performance gets close to that of the other algorithms. Harmonic functions is the best performing algorithm when we have less than 200 labeled training data. TSVM achieves better performance when there are more than 500 labeled training sentences.

Figure 2.3 shows the results obtained over the CB data set. When we have less than 500 labeled sentences, harmonic functions and TSVM perform significantly better than kNN, while SVM is the worst performing algorithm. When we have more than 500 labeled training sentences, kNN is the worst performing algorithm, while the performance of SVM increases and gets similar to that of TSVM and slightly better than that of harmonic functions.

2.6 Conclusion

We introduced a relation extraction approach based on dependency parsing and machine learning to identify protein interaction sentences in biomedical text. Unlike syntactic parsing, dependency parsing captures the semantic predicate argument relationships between the entities in addition to the syntactic relationships. We extracted the shortest paths between protein pairs in the dependency parse trees of the sentences and proposed two kernel functions for these paths based on cosine similarity and edit distance. Supervised machine learning approaches have been applied to this domain. However, they rely only on labeled training data, which is difficult to gather. To our knowledge, this is the first effort in this domain to apply semi-supervised algorithms, which make use of both labeled and unlabeled data. We evaluated and compared the performances of two semi-supervised machine learn-

ing approaches (harmonic functions and TSVM), with their supervised counterparts (kNN and SVM). We showed that, in the task of protein interaction extraction path edit kernel usually performs better than path cosine kernel since it takes into account not only common words, but also word order. Our 10-fold cross validation results showed that, TSVM performs slightly better than SVM, both of which perform better than harmonic functions. The worst performing algorithm is kNN. We compared our results with previous results published with the AIMED data set. We achieved the best F-score performance with TSVM with the path edit kernel (59.96%) which is significantly higher than the previously reported results for the same data set.

In most real-world applications there are much more unlabeled data than labeled data. Semi-supervised approaches are usually more effective in these cases, because they make use of both the labeled and unlabeled instances when making decisions. To test this hypothesis for the application of extracting protein interaction sentences from text, we performed experiments by varying the number of labeled training sentences. Our results show that, semi-supervised algorithms perform considerably better than their supervised counterparts, when there are small number of labeled training sentences. An interesting result is that, in such cases SVM performs significantly worse than the other algorithms. Harmonic functions achieve the best performance when there are only a few labeled training sentences. As number of labeled training sentences increases the performance gap between supervised and semi-supervised algorithms decreases.

CHAPTER III

Identifying Speculative Sentence Fragments in Scientific Text

3.1 Introduction

Speculation, also known as hedging, is a frequently used language phenomenon in scientific articles, especially in experimental studies, which are common in the biomedical domain. When researchers are not completely certain about the inferred conclusions, they use speculative language to convey this uncertainty. Consider the following example sentences from abstracts of articles in the biomedical domain. The abstracts are available at the U.S. National Library of Medicine PubMed web page[1]. The PubMed Identifier (PMID) of the corresponding article is given in parenthesis.

1. *We showed that the Roaz protein bound specifically to O/E-1 by using the yeast two-hybrid system. (PMID: 9151733)*
2. *These data suggest that p56lck is physically associated with Fc gamma RIIIA (CD16) and functions to mediate signaling events related to the control of NK cellular cytotoxicity. (PMID: 8405050)*

The first sentence is definite, whereas the second one contains speculative information, which is conveyed by the use of the word *"suggest"*. While speculative

[1] http://www.ncbi.nlm.nih.gov/pubmed/

information might still be useful for biomedical scientists, it is important that it is distinguished from the factual information.

Recognizing speculations in scientific text has gained interest in the recent years. Previous studies focus on identifying speculative sentences (Light et al., 2004; Medlock & Briscoe, 2007; Szarvas, 2008; Kilicoglu & Bergler, 2008). However, in many cases, not the entire sentence, but fragments of a sentence are speculative. Consider the following example sentences.

1. *The mature mitochondrial forms of the erythroid and housekeeping ALAS isozymes are predicted to have molecular weights of 59.5 kd and 64.6 kd, respectively. (PMID: 2050125)*
2. *Like RAD9, RAD9B associates with HUS1, RAD1, and RAD17, suggesting that it is a RAD9 paralog that engages in similar biochemical reactions. (PMID: 14611806)*

Both sentences are speculative, since they contain speculative information, which is signaled by the use of the word *"predicted"* in the first sentence and the word *"suggesting"* in the second sentence. The scope of the speculation keyword *"predicted"* in the first sentence spans the entire sentence. Therefore, classifying the sentence as speculative does not cause information loss. However, the scope of the speculation keyword *"suggesting"* in the second sentence applies only to the second clause of the sentence. In other words, only the statement *"RAD9B is a RAD9 paralog that engages in similar biochemical reactions"* is speculative. The statement *"Like RAD9, RAD9B associates with HUS1, RAD1, and RAD17"* conveys factual information. Therefore, classifying the entire sentence as speculative will result in information loss.

In this study, we aim to go beyond recognizing speculative sentences and tackle the problem of identifying speculative fragments of sentences. We propose an approach which is based on solving two sub-problems: (1) detecting the real speculation keywords, (2) resolving their linguistic scopes in the sentences. As the previous examples demonstrated, speculations are signaled with speculation keywords (e.g. *might, suggest, likely, hypothesize, could, predict, and etc.*). However, these keywords are not always used in a speculative context. In other words, they are not always real speculation keywords. Unlike previous approaches which classify sentences as speculative or not, we formulate the problem as classifying the keywords as real speculation keywords or not. We extract a diverse set of features such as linguistic features that represent the context of the keyword and positional features of the sentence in which the keyword occurs. We use these features with Support Vector Machines (SVM) to learn models to classify whether the occurrence of a keyword is in a speculative context or not. After detecting the real speculation keywords, we use the syntactic structures of the sentences to identify their linguistic scopes. This chapter was first published as (Özgür & Radev, 2009).

3.2 Related Work

Although hedging in scientific articles has been studied from a linguistics perspective since the 1990s (e.g. (Hyland, 1998)), it has only gained interest from a natural language processing perspective in the recent years.

The problem of identifying speculative sentences in biomedical articles has been introduced by Light et al. (2004). The authors discussed the possible application areas of recognizing speculative language and investigated whether the notion of speculative sentences can be characterized to enable manual annotation. The authors

developed two automated systems to classify sentences as speculative or not. The first method is based on substring matching. A sentence is classified as speculative if it contains one of the 14 predefined strings (*suggest, potential, likely, may, at least, in part, possibl, further investigation, unlikely, putative, insights, point toward, promise, propose*). The second method is based on using SVM with bag-of-words features. The substring matching method performed slightly better than the SVM with bag-of-words features approach.

Medlock and Briscoe (2007) extended the work of Light et al. (2004) by refining their annotation guidelines and creating a publicly available data set (FlyBase data set) for speculative sentence classification. They proposed a weakly supervised machine learning approach to classify sentences as speculative or not with the aim of minimizing the need for manually labeled training data. Their approach achieved 76% precision/recall break-even point (BEP) performance on the FlyBase data set, compared to the BEP of 60% obtained by Light et al. (2004)'s substring matching approach on the same data set. Szarvas (2008) extended the weakly supervised machine learning methodology of Medlock and Briscoe (2007) by applying feature selection to reduce the number of candidate keywords, by using limited manual supervision to filter the features, and by extending the feature representation with bigrams and trigrams. In addition, by following the annotation guidelines of Medlock and Briscoe (2007), Szarvas (2008) made available the BMC Bioinformatics data set, by annotating four full text papers from the open access BMC Bioinformatics website. They achieved a BEP performance of 85.29% and an F-measure of 85.08% on the FlyBase data set. The F-measure performance achieved on the BMC Bioinformatics data set was 74.93% when the FlyBase data set was used for training. Kilicoglu and Bergler (2008) compiled a list of speculation keywords from the examples in (Hyland, 1998)

and extended this list by using WordNet (Fellbaum, 1998) and UMLS SPECIALIST Lexicon (McCray et al., 1994). They used manually crafted syntactic patterns to identify speculative sentences and achieved a BEP and an F-measure of 85% on the FlyBase data set and a BEP and an F-measure of 82% on the BMC Bioinformatics data set.

Unlike pervious studies, which treat the problem of identifying speculative language as a sentence classification task, we tackle the more challenging problem of identifying the portions of sentences which are speculative. In other words, we allow a sentence to include both speculative and non-speculative parts. We introduce and evaluate a diverse set of features that represent the context of a keyword and use these features in a supervised machine learning setting to classify the keywords as real speculation keywords or not. Then, we develop a rule-based method to determine their linguistic scopes by considering the keyword-specific features and the syntactic structures of the sentences. To the best of our knowledge, the BioScope corpus (Vincze et al., 2008) is the only available data set that has been annotated for speculative sentence fragments and we report the first results on this corpus.

3.3 Corpus

The BioScope corpus[2] has been annotated at the token level for speculation keywords and at the sentence level for their linguistic scopes (Vincze et al., 2008). The corpus consists of three sub-corpora: medical free texts (radiology reports), biomedical article abstracts, and biomedical full text articles. In this study we focus on identifying speculations in scientific text. Therefore, we use the biomedical article abstracts and the biomedical full text articles in our experiments. The statistics (number of documents, number of sentences, and number of occurrences of spec-

[2]Available at: http://www.inf.u-szeged.hu/rgai/bioscope

ulation keywords) for these two sub-corpora are given in Table 3.1. The scientific

Data Set	Documents	Sentences	Hedge Keywords
Abstracts	1273	11871	2694
Full Papers	9	2670	682

Table 3.1: Summary of the biomedical scientific articles sub-corpora of the BioScope corpus

abstracts in the BioScope corpus were included from the Genia corpus (Collier et al., 1999). The full text papers consist of five articles from the FlyBase data set and four articles from the open access BMC Bioinformatics website. The sentences in the FlyBase and BMC Bioinformatics data sets were annotated as speculative or not and made available by Medlock and Briscoe (2007) and Szarvas (2008), respectively and have been used by previous studies in identifying speculative sentences (Medlock & Briscoe, 2007; Kilicoglu & Bergler, 2008; Szarvas, 2008). Vincze et al. (2008) annotated these full text papers and the Genia abstracts for speculation keywords and their scopes and included them to the BioScope corpus. The keywords were annotated with a minimalist strategy. In other words, the minimal unit that expresses speculation was annotated as a keyword. A keyword can be a single word (e.g. suggest, predict, might) or a phrase (complex keyword), if none of the words constituting the phrase expresses a speculation by itself. For example the phrase *"no evidence of'"* in the sentence *"Direct sequencing of the viral genomes and reinfection kinetics showed no evidence of wild-type reversion even after prolonged infection with the Tat- virus."* is an example of a complex keyword, since the words forming the phrase can only express speculation together.

In contrast to the minimalist strategy followed when annotating the keywords, the annotation of scopes of the keywords was performed by assigning the scope to the largest syntactic unit possible by including all the elements between the keyword and the target word to the scope (in order to avoid scopes without a keyword) and

by including the modifiers of the target word to the scope (Vincze et al., 2008). The reader can refer to (Vincze et al., 2008) for the details of the corpus and the annotation guidelines.

The inter-annotator agreement rate was measured as the F-measure of the annotations of the first annotator by considering the annotations of the second one as the gold standard. The agreement rate for speculation keyword annotation is reported as 92.05% for the abstracts and 90.81% for the full text articles and the agreement rate for speculation scope resolution is reported as 94.04% for the abstracts and 89.67% for the full text articles (Vincze et al., 2008). These rates can be considered as the upper bounds for the automated methods proposed in this study.

3.4 Identifying Speculation Keywords

Words and phrases such as *"might", "suggest", "likely", "no evidence of", and "remains to be elucidated"* that can render statements speculative are called speculation keywords. Speculation keywords are not always used in speculative context. For instance, consider the following sentences:

1. *Thus, it appears that the T-cell-specific activation of the proenkephalin promoter is mediated by NF-kappa B. (PMID: 91117203)*
2. *Differentiation assays using water soluble phorbol esters reveal that differentiation becomes irreversible soon after AP-1 appears. (PMID: 92088960)*

The keyword *"appears"* in the first sentence renders it speculative. However, in the second sentence, *"appears"* is not used in a speculative context.

The first sub-problem that we need to solve in order to identify speculative sentence fragments is identifying the real speculation keywords in a sentence (i.e. the keywords which convey speculative meaning in the sentence). We formulate the

problem as a supervised classification task. We extract the list of keywords from the training data which has been labeled for speculation keywords. We match this list of keywords in the unlabeled (test data) and train a model to classify each occurrence of a keyword in the unlabeled test set as a real speculation keyword or not. The challenge of the task can be demonstrated by the following statistics from the Genia Abstracts of the BioScope corpus. There are 1273 abstracts in the corpus. There are 138 unique speculation keywords and the total number of their occurrence in the abstracts is 6125. In only 2694 (less than 50%) of their occurrences they are used in speculative context (i.e., are real speculation keywords).

In this study we focus on identifying the features that represent the context of a speculation keyword and use SVM with linear kernel (we used the SVM^{light} package (Joachims, 1999a)) as our classification algorithm. The following sub-section describes the set of features that we propose.

3.4.1 Feature Extraction

We introduce a set of diverse types of features including keyword specific features such as the stem and the part-of-speech (POS) of the keyword, and keyword context features such as the words surrounding the keyword, the dependency relation types originating at the keyword, the other keywords that occur in the same sentence as the keyword, and positional features such as the section of the paper in which the keyword occurs. While designing the features, we were inspired by studies on other natural language processing problems such as Word Sense Disambiguation (WSD) and summarization. For example, machine learning methods with features based on part-of-speech tags, word stems, surrounding and co-occurring words, and dependency relationships have been successfully used in WSD (Montoyo et al., 2005; Ng & Lee, 1996; Dligach & Palmer, 2008) and positional features such as the position

of a sentence in the document have been used in text summarization (e.g. (Radev et al., 2004)).

Keyword Features

Statistics from the BioScope corpus suggest that different keywords have different likelihoods of being used in a speculative context (Vincze et al., 2008). For example, the keyword *"suggest"* has been used in a speculative context in all its occurrences in the abstracts and in the full papers. On the other hand, *"appear"* is a real speculation keyword in 86% of its occurrences in the abstracts and in 83% of its occurrences in the full papers, whereas *"can"* is a real speculation keyword in 12% of its occurrences in the abstracts and in 16% of its occurrences in the full papers. POS of a keyword might also play a role in determining whether it is a real speculation keyword or not. For example, consider the keyword *"can"*. It is more likely to have been used in a speculative context when it is a modal verb, than when it is a noun. Based on these observations, we hypothesize that features specific to a keyword such as the keyword itself, the stem of the keyword, and the POS of the keyword might be useful in discriminating the speculative versus non-speculative use of it. We use Porter's Stemming Algorithm (Porter, 1980) to obtain the stems of the keywords and Stanford Parser (de Marneffe et al., 2006) to get the POS of the keywords. If a keywords consists of multiple words, we use the concatenation of the POS of the words constituting the keyword as a feature. For example, the extracted POS feature for the keywords *"no evidence"* and *"no proof"* is *"DT.NN"*.

Dependency Relation Features

Besides the occurrence of a speculation keyword, the syntactic structure of the sentence also plays an important role in characterizing speculations. Kilicoglu and

Bergler (2008) showed that manually identified syntactic patterns are effective in classifying sentences as speculative or not. They identified that, while some keywords do not indicate hedging when used alone, they might act as good indicators of hedging when used with a clausal complement or with an infinitival clause. For example, the *"appears"* keyword in the example sentences, which are given in the beginning of Section 3.4, is not a real speculation keyword in the second example *"...soon after AP-1 appears."*, whereas it is a real speculation keyword in the first example, where it is used with a *that* clausal complement *"...it appears that..."*. Similarly, *"appears"* is used in a speculative context in the following sentence, where it is used with an infinitival clause: *"Synergistic transactivation of the BMRF1 promoter by the Z/c-myb combination appears to involve direct binding by the Z protein."*.

Another observation is that, some keywords act as real speculation keywords only when used with a negation. For example, words such as *"know", "evidence", and "proof"* express certainty when used alone, but express a speculation when used with a negation (e.g., *"not known", "no evidence", "no proof"*).

Auxiliaries in verbal elements might also give clues for the speculative meaning of the main verbs. Consider the example sentence: *"Our findings may indicate the presence of a reactivated virus hosted in these cells."*. The modal auxiliary *"may"* acts as a clue for the speculative context of the main verb *"indicate"*.

We defined boolean features to represent the syntactic structures of the contexts of the keywords. We used the Stanford Dependency Parser (de Marneffe et al., 2006) to parse the sentences that contain a candidate speculation keyword and extracted the following features from the dependency parse trees.

Clausal Complement: A Boolean feature which is set to 1, if the keyword has a child which is connected to it with a clausal complement or infinitival clause

dependency type.

Negation: A Boolean feature which is set to 1, if the keyword (1) has a child which is connected to it with a negation dependency type (e.g. "not known": "not" is a child of "known", and the Stanford Dependency Type connecting them is "neg") or (2) the determiner "no" is a child of the keyword (e.g., "no evidence": "no" is a child of "evidence" and the Stanford Dependency Type connecting them is "det").

Auxiliary: A Boolean feature which is set to 1, if the keyword has a child which is connected to it with an auxiliary dependency type (e.g. "may indicate": "may" is a child of "indicate", and the Stanford Dependency Type connecting them is "aux").

If a keyword consists of multiple-words, we examine the children of the word which is the ancestor of the other words constituting the keyword. For example, "no evidence" is a multi-word keyword, where "evidence" is the parent of "no". Therefore, we extract the dependency parse tree features for the word "evidence".

Surrounding Words

Recent studies showed that using machine learning with variants of the "bag-of-words" feature representation is effective in classifying sentences as speculative vs. non-speculative (Light et al., 2004; Medlock & Briscoe, 2007; Szarvas, 2008). Therefore, we also decided to include bag-of-words features that represent the context of the speculation keyword. We extracted the words surrounding the keyword and performed experiments both with and without stemming, and with window sizes of one, two, and three. Consider the sentence: *"Our findings may indicate the presence of a reactivated virus hosted in these cells."*. The bag-of-words features for

the keyword "indicate", when a window size of three and no stemming is used are: *"our", "findings", "may", "indicate", "the", "presence", "of"*. In other words, the feature set consists of the keyword, the three words to the left of the keyword, and the three words to the right of the keyword.

Positional Features

Different parts of a scientific article might have different characteristics in terms of the usage of speculative language. For example, Hyland (1998) analyzed a data set of molecular biology articles and reported that the distribution of speculations is similar between abstracts and full text articles, whereas the Results and Discussion sections tend to contain more speculative statements compared to the other sections (e.g. Materials and Methods or Introduction and Background sections). The analysis of Light et al. (2004) showed that the last sentence of an abstract is more likely to be speculative than non-speculative.

For the scientific abstracts data set, we defined the following boolean features to represent the position of the sentence the keyword occurs in. Our intuition is that titles and the first sentences in the abstract tend to be non-speculative, whereas the last sentence of the abstract tends to be speculative.

Title: A Boolean feature which is set to 1, if the keyword occurs in the title.

First Sentence: A Boolean feature which is set to 1, if the keyword occurs in the first sentence of the abstract.

Last Sentence: A Boolean feature which is set to 1, if the keyword occurs in the last sentence of the abstract.

For the scientific full text articles data set, we defined the following features that represent the position of the sentence in which the keyword occurs. Our assumption

is that the "Results and Discussion" and the "Conclusion" sections tend to contain more speculative statements than the "Materials and Methods" and "Introduction and Background" sections. We also assume that figure and table legends are not likely to contain speculative statements.

Title: A Boolean feature which is set to 1, if the keyword occurs in the title of the article, or in the title of a section or sub-section.

First Sentence: A Boolean feature which is set to 1, if the keyword occurs in the first sentence of the abstract.

Last Sentence: A Boolean feature which is set to 1, if the keyword occurs in the last sentence of the abstract.

Background: A Boolean feature which is set to 1, if the keyword occurs in the Background or Introduction section.

Results: A Boolean feature which is set to 1, if the keyword occurs in the Results or in the Discussion section.

Methods: A Boolean feature which is set to 1, if the keyword occurs in the Materials and Methods section.

Conclusion: A Boolean feature which is set to 1, if the keyword occurs in the Conclusion section.

Legend: A Boolean feature which is set to 1, if the keyword occurs in a table or figure legend.

Co-occurring Keywords

Speculation keywords usually co-occur in the sentences. Consider the sentence: *"We, therefore, wished to determine whether T3SO4 could mimic the action of thyroid hormone in vitro."*. Here, *"whether"* and *"could"* are speculation keywords

and their co-occurence might be a clue for their speculative context. Therefore, we decided to include the co-occurring keywords to the feature set of a keyword.

3.5 Resolving the Scope of a Speculation

After identifying the real speculation keywords, the next step is determining their scopes in the sentences, so that the speculative sentence fragments can be detected. Manual analysis of sample sentences from the BioScope corpus and their parse trees suggests that the scope of a keyword can be characterized by its part-of-speech and the syntactic structure of the sentence in which it occurs. Consider the example sentence whose parse tree is shown in Figure 3.1. The sentence contains three speculation keywords, "or" and two occurrences of "might". The scope of the conjunction "or", extends to the "VP" whose children it coordinates. In other words, the scope of "or" is "[might be one of the earliest crucial steps in the lysis of normal and dex-resistant CEM cells, *or* might serve as a marker for the process]". Here, "or" conveys a speculative meaning, since we are not certain which of the two sub-clauses (sub-clause 1: [might be one of the earliest crucial steps in the lysis of normal and dex-resistant CEM cells] *or* sub-clause 2: [might serve as a marker for the process]) is correct. The scope of both occurrences of the modal verb "might" is the parent "VP". In other words, the scope of the first occurrence of "might" is "[*might* be one of the earliest crucial steps in the lysis of normal and dex-resistant CEM cells]" and the scope of the second occurrence of "might" is "[*might* serve as a marker for the process]". By examining the keywords, sample sentences and their syntactic parse trees we developed the following rule-based approach to resolve the scopes of speculation keywords. The examples given in this section are based on the syntactic structure of the Penn Tree Bank. But, the rules are generic (e.g. "the scope of a

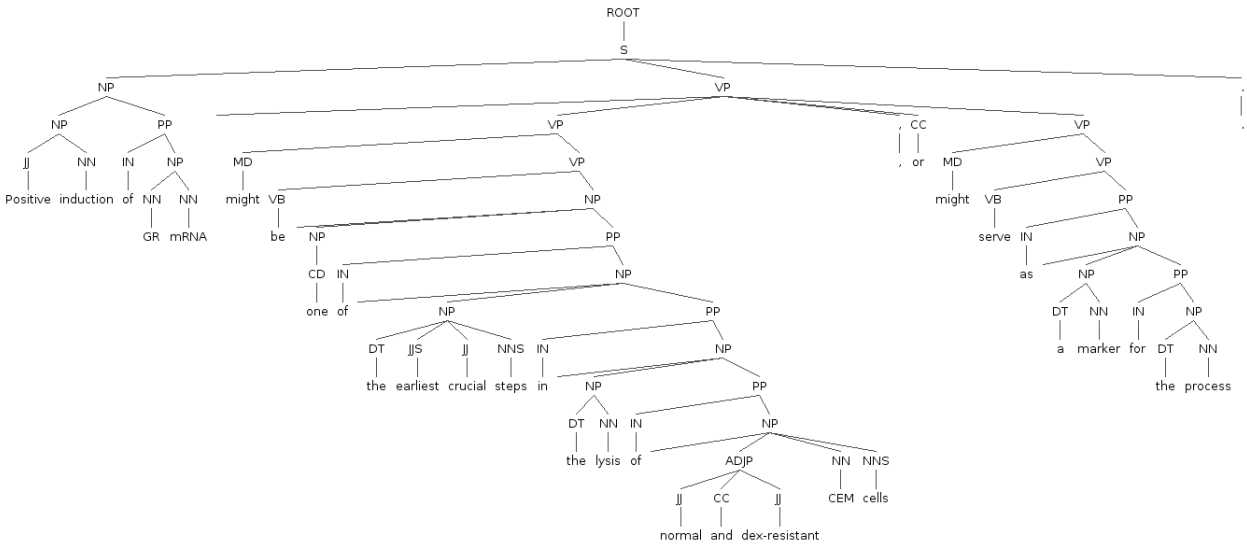


Figure 3.1: The syntactic parse tree of the sentence "*Positive induction of GR mRNA might be one of the earliest crucial steps in the lysis of normal and dex-resistant CEM cells, or might serve as a marker for the process.*"

verb followed by an infinitival clause, extends to the whole sentence").

The scope of a conjunction or a determiner (e.g. or, and/or, vs) is the syntactic phrase to which it is attached. For example, the scope of "or" in Figure 3.1 is the "VP" immediately dominating the "CC".

The scope of a modal verb (e.g. may, might, could) is the "VP" to which it is attached. For example, the scope of "might" in Figure 3.1 is the "VP" immediately dominating the "MD".

The scope of an adjective or an adverb starts with the keyword and ends with the last token of the highest level "NP" which dominates the adjective or the adverb. Consider the sentence "The endocrine events that are rapidly expressed (seconds) are due to a [possible interaction with cellular membrane]." The scope of the speculation keyword "possible" is enclosed in rectangular brackets. The sub-tree that this scope maps to is: "(NP (NP (DT a) (JJ possible) (NN interaction)) (PP (IN with) (NP (JJ cellular) (NN membrane))))". If there does not exist a "NP" dominating the adverb

or adjective keyword, the scope extends to the whole sentence. For example the scope of the speculation adverb "probably" in the sentence "[The remaining portion of the ZFB motif was probably lost in TPases of insect Transib transposons]" is the whole sentence.

The scope of a verb followed by an infinitival clause extends to the whole sentence. For example, the scope of the verb "appears" followed by the "to" infinitival clause is the whole sentence in "[The block of pupariation appears to involve signaling through the adenosine receptor (AdoR)]".

The scope of a verb in passive voice extends to the whole sentence such as the scope of "suggested" in "[The existence of such an independent mechanism has also been suggested in mammals]".

If none of the above rules apply, the scope of a keyword starts with the keyword and ends at the end of the sentence (or clause). An example is the scope of "suggested" in "This [suggested that there is insufficient data currently available to determine a reliable ratio for human]".

3.6 Evaluation

We evaluated our approach on two different types of scientific text from the biomedical domain, namely the scientific abstracts sub-corpus and the full text articles sub-corpus of the BioScope corpus (see Section 3.3). We used stratified 10-fold cross-validation to evaluate the performance on the abstracts. In each fold, 90% of the abstracts are used for training and 10% are used to test. To facilitate comparison with future studies the PubMed Identifiers of the abstracts that we used as a test set in each fold are provided[3]. The full text papers sub-corpus consists of nine articles. We used leave-one-out cross-validation to evaluate the performance on the full text

[3]http://clair.si.umich.edu/clair/bioscope/

papers. In each iteration eight articles are used for training and one article is used to test. We report the average results over the runs for each data set.

3.6.1 Evaluation of Identifying Speculation Keywords

To classify whether the occurrence of a keyword is in speculative context or not, we built linear SVM models by using various combinations of the features introduced in Section 3.4.1. Tables 3.2 and 3.3 summarize the results obtained for the abstracts and the full text papers, respectively. *BOW N* is the bag-of-words features obtained from the words surrounding the keyword. N is the window size. We experimented both with the stemmed and non-stemmed versions of this feature type. The non-stemmed versions performed slightly better than the stemmed versions. The reason might be due to the different likelihoods of being used in a speculative context of different inflected forms of words. For example, consider the words "appears" and "appearance". They have the same stems, but "appearance" is less likely to be a real speculation keyword than "appears". Another observation is that, decreasing the window size led to improvement in performance. This suggests that the words right before and right after the candidate speculation keyword are more effective in distinguishing its speculative vs. non-speculative context compared to a wider local context. Wider local context might create sparse data and degrade performance. Consider the example, "it appears that TP53 interacts with AR". The keyword "appears", and BOW1 ("it" and "that") are more relevant for the speculative context of the keyword than "TP53", "interacts", and "with". Therefore, for the rest of the experiments we used the *BOW 1* version, i.e., the non-stemmed surrounding bag-of-words with window size of 1. *KW* stands for the keyword specific features, i.e., the keyword, its stem, and its part-of-speech. *DEP* stands for the dependency relation features. *POS* stands for the positional features and *CO-KW* stands for the co-

occurring keywords feature.

Our results are not directly comparable with the prior studies about identifying speculative sentences (see Section 3.2), since we attempted to solve a different problem, which is identifying speculative parts of sentences. Only the substring matching approach that was introduced in (Light et al., 2004) could be adapted as a keyword classification task, since the substrings are keywords themselves and we used this approach as a baseline in the keyword classification sub-problem. We compare the performances of our models with two baseline methods, which are based on the substring matching approach. Light et al. (2004) have shown that the substring matching method with a predefined set of 14 strings performs slightly better than an SVM model with bag-of-words features in classifying sentences as speculative vs. non-speculative (see Section 3.2). In baseline 1, we use the 14 strings identified in (Light et al., 2004) and classify all the keywords in the test set that match any of them as real speculation keywords. Baseline 2 is similar to baseline 1, with the difference that rather than using the set of strings in (Light et al., 2004), we extract the set of keywords from the training set and classify all the words (or phrases) in the test set that match any of the keywords in the list as real speculation keywords.

Baseline 1 achieves high precision, but low recall. Whereas, baseline 2 achieves high recall in the expense of low precision. All the SVM models in Tables 3.2 and 3.3 achieve more balanced precision and recall values, with F-measure values significantly higher than the baseline methods. We start with a model that uses only the keyword-specific features (KW). This type of feature alone achieved a significantly better performance than the baseline methods (90.61% F-measure for the abstracts and 80.57% F-measure for the full text papers), suggesting that the keyword-specific features are important in determining its speculative context. We extended the

feature set by including the dependency relation (DEP), surrounding words (BOW 1), positional (POS), and co-occurring keywords (CO-KW) features. Each new type of included feature improved the performance of the model for the abstracts. The best F-measure (91.69%) is achieved by using all the proposed types of features. This performance is close to the upper bound, which is the human inter-annotator agreement F-measure of 92.05%.

Including the co-occurring keywords to the feature set for full text articles slightly improved precision, but deceased recall, which led to lower F-measure. The best F-measure (82.82%) for the full text articles is achieved by using all the feature types except the co-occurring keywords. The achieved performance is significantly higher than the baseline methods, but lower than the human inter-annotator agreement F-measure of 90.81%. The lower performance for the full text papers might be due to the small size of the data set (9 full text papers compared to 1273 abstracts).

Method	Recall	Precision	F-Measure
Baseline 1	52.84	92.71	67.25
Baseline 2	97.54	43.66	60.30
BOW 3 - stemmed	81.47	92.36	86.51
BOW 2 - stemmed	81.56	93.29	86.97
BOW 1 - stemmed	83.08	93.83	88.05
BOW 3	82.58	92.04	86.98
BOW 2	82.77	92.74	87.41
BOW 1	83.27	93.67	88.10
KW: kw, kw-stem, kw-pos	88.62	92.77	90.61
KW, DEP	88.77	92.67	90.64
KW, DEP, BOW 1	88.46	94.71	91.43
KW, DEP, BOW 1, POS	88.16	95.21	91.50
KW, DEP, BOW 1, POS, CO-KW	88.22	95.56	91.69

Table 3.2: Results for the Scientific Abstracts

3.6.2 Evaluation of Resolving the Scope of a Speculation

We compared the proposed rule-based approach for scope resolution with two baseline methods. Previous studies classify sentences as speculative or not, therefore

Method	Recall	Precision	F-Measure
Baseline 1	33.77	86.75	47.13
Baseline 2	88.22	52.57	64.70
BOW 3 - stemmed	70.79	83.88	76.58
BOW 2 - stemmed	72.31	85.49	78.11
BOW 1 - stemmed	73.49	84.35	78.41
BOW 3	70.54	82.56	75.88
BOW 2	71.52	85.93	77.94
BOW 1	73.72	86.27	79.43
KW: kw, kw-stem, kw-pos	75.21	87.08	80.57
KW, DEP	75.02	89.49	81.53
KW, DEP, BOW 1	76.15	89.54	82.27
KW, DEP, BOW 1, POS	76.17	90.81	82.82
KW, DEP, BOW 1, POS, CO-KW	75.76	90.82	82.58

Table 3.3: Results for the Scientific Full Text Papers

implicitly assigning the scope of a speculation to the whole sentence (Light et al., 2004; Medlock & Briscoe, 2007; Szarvas, 2008; Kilicoglu & Bergler, 2008). Baseline 1 follows this approach and assigns the scope of a speculation keyword to the whole sentence. Szarvas (2008) suggest assigning the scope of a keyword from its occurrence to the end of the sentence. They state that this approach works accurately for clinical free texts, but no any results are reported (Szarvas, 2008). Baseline 2 follows the approach proposed in (Szarvas, 2008) and assigns the scope of a keyword to the fragment of the sentence that starts with the keyword and ends at the end of the sentence. Table 3.4 summarizes the accuracy results obtained for the abstracts and the full text papers.

The poor performance of baseline 1, emphasizes the importance of detecting the portions of sentences that are speculative, since less than 5% of the sentences that contain speculation keywords are entirely speculative. Classifying the entire sentences as speculative or not leads to loss in information for more than 95% of the sentences. The rule-based method significantly outperformed the two baseline methods, indicating that the part-of-speech of the keywords and the syntactic parses of the sentences are effective in characterizing the speculation scopes.

Method	Accuracy-Abstracts	Accuracy-Full text
Baseline 1	4.82	4.29
Baseline 2	67.60	42.82
Rule-based method	79.89	61.13

Table 3.4: Scope resolution results

3.7 Conclusion

We presented an approach to identify speculative sentence fragments in scientific articles. Our approach is based on solving two sub-problems. The first one is identifying the keywords which are used in speculative context and the second one is determining the scopes of these keywords in the sentences. We evaluated our approach for two types of scientific texts, namely abstracts and full text papers from the BioScope corpus.

We formulated the first sub-problem as a supervised classification task, where the aim is to learn models to classify the candidate speculation keywords as real speculation keywords or not. We focused on identifying different types of linguistic features that capture the contexts of the keywords. We achieved a performance which is significantly better than the baseline methods and close to the upper bound, which is the human inter-annotator agreement F-measure.

We hypothesized that the scope of a speculation keyword can be characterized by its part-of-speech and the syntactic structure of the sentence and developed rules to map the scope of a keyword to the nodes in the syntactic parse tree. Our results show that the rule-based method is effective in resolving the scopes of the speculation keywords. The considerably lower performance of the baseline of assigning the scope of a speculation keyword to the whole sentence indicates the importance of detecting speculative sentence portions rather than classifying the entire sentences as speculative or not.

CHAPTER IV

Centrality-Based Literature Mining for Discovering Gene-Disease Associations

4.1 Introduction

In Chapter II we proposed a dependency tree kernel-based method to extract protein interactions from text. An exciting usage of the information extracted from the scientific literature is trying to uncover hidden links. In this chapter we propose a literature-based discovery (LBD) method to infer gene-disease associations by analyzing the topology of a gene interaction network extracted from the biomedical scientific literature. In Chapter V we will present the general framework of this LBD approach and adapt it to solve a different biological problem, namely to discover genes which are potentially important for vaccine development. Unlike most previous LBD methods that are based on *Swanson's ABC* model and depend on co-occurrence information among the entities (see Chapter I), our approach integrates natural language processing based text mining with network analysis in a novel way.

One of the major goals of the post-genome era is to understand the role of genetics in human health and diseases (Lander et al., 2001; Venter et al., 2001). While fewer than 100 gene-disease associatons were known before the project started in 1990, currently more than 1400 gene-disease relationships have been identified[1]. Determining

[1] http://www.genome.gov/11006929

gene-disease associations will enhance developing new techniques for prevention, diagnosis, and treatment of the diseases.

There are curated databases that store gene-disease association information. One of the most well-known such databases is (OMIM, 2007), which provides summaries of publications about gene-disease relationships. However, it usually takes time before new discoveries are included in the curated databases. Given that the amount of biomedical literature regarding the identification of disease genes is increasing rapidly, one of the challenges that scientists in this domain face is that most of the relevant information remains hidden in the unstructured text of the published papers.

Another challenge is that the identification of new disease genes requires laborious experiments. For example, the genetic linkage analysis method is successfully used to determine the genomic regions that are associated with a disease. However, these regions often contain hundreds of genes and experimentally identifying the actual disease genes out of the large amount of candidate genes is very time-consuming and costly. Therefore, predicting good candidate genes before experimental analysis is crucial.

To address these challenges, we propose an approach based on integrating text mining and network analysis methods to automatically extract known disease genes and to predict unknown disease genes, which can be good candidates for experimental study. We started by collecting an initial set of genes (seed genes) known to be related to a disease from curated databases such as OMIM. We then used the information extraction approach based on dependency parsing and support vector machines (SVM), which we introduced in Chapter II, to extract the interactions of the seed genes and their neighbors (the genes that the seed genes interact with) from the biomedical literature. We generated the dependency parses of the sentences that

contain at least two seed or neighbor genes and extracted the paths between all pairs of genes from the dependency parse trees. We used SVM with dependency path edit kernel to classify the sentences as describing an interaction between a gene pair or not.

After extracting the interactions, we constructed a disease-specific gene interaction network, where the nodes are the seed genes and their neighbors, and two genes are linked, if we have extracted an interaction between them. Next, we ranked the genes in the network by degree, eigenvector, betweenness, and closeness network centrality metrics. Our main hypothesis is that the central genes in the disease-specific network are likely to be related to the disease. To our knowledge, this is the first effort of building a gene interaction network by automatic literature mining and applying network centrality to predict gene-disease associations on that network. This chapter is based on the work published as (Özgür et al., 2008).

4.2 Related Work

In this section we discuss closely related work on protein interaction networks and identifying gene-disease associations. Related work on literature-based discovery was discussed in Chapter I and related work on protein interaction extraction from text was presented in Chapter II.

Most of the previous studies that use text mining to extract gene-disease associations from the biomedical literature are based on the co-occurrence frequencies of genes and diseases. For example, Adamic et al. (2002) present a method based on determining whether the frequency of occurrence of a gene in articles that mention a certain disease is statistically significantly higher than the expected frequency of occurrence computed by the Binomial distribution. They evaluated their approach

for breast cancer and confirmed the relevance of 7 out of the 10 highest ranked genes to breast cancer by using a human edited breast cancer gene database[2]. Another relevant study is conducted by Al-Mubaid and Singh (2005). Given a disease name, the set of documents that contain the disease name (positive document set) and a randomly-selected document set (negative document set) are extracted. Co-occurrence and term frequency based concepts from information theory are used to determine the genes that are significantly associated with the disease. The authors found six genes significantly associated with Alzheimer's disease and confirmed the correctness of their results through articles from PubMed. Ade et al. (2007) developed the Gene2MeSH[3] web system that links genes to MeSH (Medical Subject Heading) terms based on the significance of their associations in PubMed abstracts.

Determining the genes that cause a disease usually requires laborious experiments over a large number of candidate genes. Therefore, another challenge in the domain is predicting and prioritizing candidate disease genes, which can further be validated by detailed experiments. Most proposed data mining approaches make use of available curated databases and predict gene-disease associations by using keyword similarity to known disease genes and phenotypes. For example, GeneSeeker (van Driel et al., 2002) is a web-based system that integrates positional and expression/phenotypic data from nine different human and mouse databases and provides a quick overview of interesting candidate genes. The authors evaluated their approach for ten syndromes. On average, the system reduced a list of 163 candidate genes to a list of 22 genes, which still contained the correct disease gene. Freudenberg and Propping (2002) proposed a method based on clustering diseases based on their phenotypic similarity, which is computed by considering the similarity of the disease index terms

[2]http://tyrosine.biomedcomp.com
[3]http://gene2mesh.ncibi.org/

in the OMIM database. Candidate genes for a disease in a cluster are predicted by selecting functionally similar genes to the genes associated with the other diseases in the cluster. The authors performed a leave-one-out cross-validation of 878 diseases using $10,672$ genes. They reported that in roughly one-third of the diseases, the correct disease gene was within the top scoring 321 genes, and in the two-third of the diseases, the correct disease gene was within the top scoring $1,600$ genes. The G2D system (Perez-Iratxeta et al., 2002, 2005) uses a fuzzy logic and text mining approach based on co-occurrence of relevant keywords in biomedical abstracts to associate pathological conditions with gene ontology (GO) terms (Ashburner et al., 2000). Prediction of candidate genes is performed by searching for genes homologous to the GO-annotated and disease-associated genes. The authors evaluated their system with 100 known disease-associated genes and found that the correct disease gene was among the 8 top-scoring genes with 25% chance, and among the 30 top-scoring genes with 50% chance.

Another line of research that is related to ours is building and analyzing protein-protein interaction (PPI) networks (see Chapter I for an overview on PPI networks). PPIs can be represented as complex networks, where the nodes are the proteins and the edges represent the interactions between the pairs of proteins they connect. This representation makes it possible to analyze PPI networks from a graph theory and complex networks perspective, which can give biologists a variety of new insights. Most graph-theoretic studies of PPI networks extract the interactions from curated databases (Jeong et al., 2001; Wuchty et al., 2003; Spirin & Mirny, 2003; Schwikowski et al., 2000). There are also recent studies that analyze protein interaction networks constructed by mining the literature (Chen & Sharp, 2004; Hoffmann & Valencia, 2005). It has been shown that the interaction networks constructed in either way,

share similar topological properties such as being small-world and scale-free, with each other and with various non-biological complex systems such as the WWW, the Internet, and social networks (Chen & Sharp, 2004; Hoffmann & Valencia, 2005; Jeong et al., 2001).

Graph-theoretic analysis of protein interaction networks have been successfully applied in many biological domains. For example, protein interaction networks have been used for evolutionary comparisons among organisms (Wuchty et al., 2003), for identifying functional modules and network motifs (Spirin & Mirny, 2003), and for predicting functional annotations based on network connectivity (Schwikowski et al., 2000). Schwikowski et al. (2000) used a majority-rule method that assigns to a protein the function that occurs most commonly among its neighbors and reported an accuracy of 70% for the yeast protein interaction network.

Recently, protein interaction networks have also been used to predict gene-disease associations (Chen et al., 2006; Gonzalez et al., 2007). Chen et al. (2006), used an initial gene list (seed genes) for Alzheimer's from the OMIM database, and built an interaction network by extracting the interactions of these genes from the Online Predicted Human Interaction Database (OPHID) (Brown & Jurisica, 2005). They define a heuristic scoring function for the genes based on their connectedness in the graph. When building the network, only the interactions among the seed genes and the interactions of seed genes with their neighbors are considered. The interactions among the neighbors is not taken into account. Thus, this approach is biased in favor of the seed genes. 19 of the top scoring genes are seed and only one is a non-seed (inferred) gene. Gonzalez et al. (2007) start with a list of seed genes obtained from the automatically mined CBioC database and create an interaction network by extracting the interactions of the seed genes from the CBioC database (Baral et al.,

2005) and curated databases such as BIND (Bader et al., 2003) and MINT (Zanzoni et al., 2002). Like Chen et al. (2006), they do not take into account the interactions among the non-seed genes. To eliminate the bias in favor of the seed genes, they refine the scoring function by considering just the interactions with seed genes and including a measure for the impact of each gene on the connectivity of the network. 45% of their top scoring 20 genes are non-seed and 66.67% of these non-seed genes are correctly inferred genes, i.e., reported in OMIM or in the literature as being related to the disease.

Our approach is different than most previous approaches in two aspects. First, we create a protein interaction network by automatic literature mining using the dependency path edit kernel method introduced in Chapter II. Second, we use degree, eigenvector, betweenness, and closeness centrality to rank the gene-disease associations. Centrality measures, which define the relative importance of a node in the graph, have originally been developed and used in nonbiological domains. For example, the web pages in the popular search engine Google are ranked by using the Pagerank algorithm, which is based on eigenvector centrality (Page et al., 1999). Recently, eigenvector centrality has also been used in document summarization to identify the most important sentences (Erkan & Radev, 2004) as well as to identify the most influential members of the US Senate (Fader et al., 2007). A number of recent studies have successfully applied centrality measures in biological domains. For example, Jeong et al. (2001) used degree centrality to predict lethal mutations in the yeast protein interaction network. They showed that the network is tolerant to random errors, whereas errors related to the most central proteins cause lethality. Similarly, Joy et al. (2005) and Hahn and Kern (2005) have found that there is an association between the betweenness centrality and the essentiality of a gene, where

an essential gene is a gene that causes the organism to die when it malfunctions. Goh et al. (2007) showed that central genes based on degree are also essential.

4.3 Methods

The high level system description for predicting gene-disease associations is shown in Figure 4.1. The approach is described in more detail in the following subsections.

4.3.1 Corpus

To construct the literature-mined protein interaction network we used 48,245 articles from PubMed Central (PMC) Open Access[4], which is an open access digital archive of biomedical and life science journals. Unlike PubMed, articles in PMC are full-text.

We pre-processed the corpus by segmenting the articles into sentences with Mx-Terminator (Reynar & Ratnaparkhi, 1997). Protein and gene names are annotated with Genia Tagger (Tsuruoka et al., 2005), whose developers report an F-score performance of 71.37% for biological named entity recognition[5].

4.3.2 Initial List of Seed Genes

To build an interaction network for a disease and to infer gene-disease associations from the network properties, we started with an initial list of seed genes known to be related to the disease.

We evaluated our system for prostate cancer. We compiled 15 prostate cancer seed genes from the Morbid Map component of Online Mendelian Inheritance in Man database (OMIM, 2007). OMIM Morbid Map shows the cytogenetic map location of disease-associated genes described in OMIM. Table 4.1 lists the seed genes for prostate cancer.

[4]http://www.pubmedcentral.nih.gov/about/openftlist.html
[5]http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

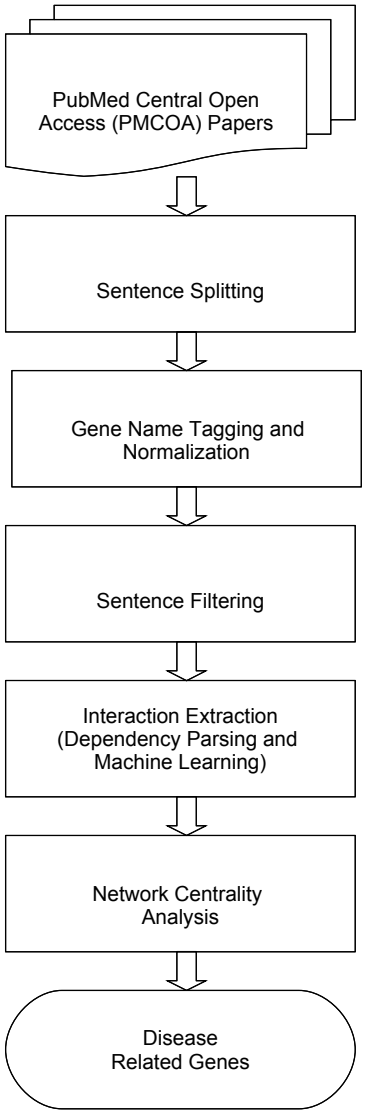


Figure 4.1: Description of the literature-based discovery system for identifying gene-disease associations.

Gene	Description
AR	androgen receptor
BRCA2	breast cancer 2, early onset
MSR1	macrophage scavenger receptor 1
EPHB2	EPH receptor B2
KLF6	Kruppel-like factor 6
MAD1L1	MAD1 mitotic arrest deficient-like 1 (yeast)
HIP1	huntingtin interacting protein 1
CD82	CD82 molecule
ELAC2	elaC homolog 2 (E. coli)
MXI1	MAX interactor 1
PTEN	phosphatase and tensin homolog
RNASEL	ribonuclease L (2',5'-oligoisoadenylate synthetase-dependent)
HPC1	hereditary prostate cancer 1
CHEK2	CHK2 checkpoint homolog (S. pombe)
PCAP	predisposing for prostate cancer

Table 4.1: The prostate cancer seed genes retrieved from OMIM Morbid Map.

4.3.3 Gene Name Normalization

Identifying the gene and protein names in text is usually not sufficient to uniquely identify the corresponding entities. This is due to the fact that, most gene and protein names have several different synonyms and spelling variations. In order to link the gene/protein names to their corresponding entries in the interaction databases such as UniProt (Bairoch et al., 2005) and MiMI (Tarcea et al., 2009) or to build an interaction network using the interactions extracted from the literature, the gene and protein names have to be normalized (mapped) to a canonical name. For example, the PTEN gene might appear in text as MMAC1, TEP1, PTEN1, or phosphatase and tensin homolog. Similarly, the TP53 gene can occur in text as TP53, p53, LFS1, or tumor protein p53. If the gene names that correspond to the same gene are not normalized, each different synonym and spelling variant will be represented as a separate node in a gene-interaction network extracted from the literature as shown in Figure 4.2. With five different synonyms for PTEN and four different synonyms for TP53, 20 different edges can be obtained although they actually represent the same

edge (interaction). Therefore, we used a dictionary-based approach to normalize the gene names tagged by Genia Tagger so that each gene is represented by a single node in the interaction network. We used the HUGO Gene Nomenclature Committee (HGNC) database[6] (Wain et al., 2004) as the dictionary for gene names and their synonyms[7]. We matched the tagged gene names against the approved symbol, approved name, previous symbols, previous names, aliases, and name aliases fields of the database. We unified each tagged gene name to its corresponding approved gene symbol.

4.3.4 Extracting the Gene Interaction Network from the Literature

We used the initial list of seed genes to build a disease-specific gene interaction network mined automatically from the literature. Before applying our text mining approach to extract gene interactions, we selected the potential interaction sentences from the PMC Open Access corpus. A list of interaction words, which consists of 45 noun and 53 verb roots was compiled from the literature. We extended the list to contain all the inflected forms of the words and spelling variations such as *coactivate/co-activate* and *localize/localise*. Our assumption is that a sentence that describes an interaction between a pair of genes should contain at least two genes and an interaction word (e.g. binds, bound, interacts, activates, inhibits, and phosphorylates). We expanded the seed gene list, by including all the genes that appear in the same sentence with a seed gene. We filtered out the sentences that do not contain an interaction word and at least two genes from the expanded gene list.

To build the gene interaction network, we used the path edit kernel with SVM, which was introduced in Chapter II, to automatically extract the protein interactions from the literature. We trained the system by combining the AIMED and CB data

[6] http://www.genenames.org/index.html

[7] As of September, 2007 the database contains 24,680 approved gene records

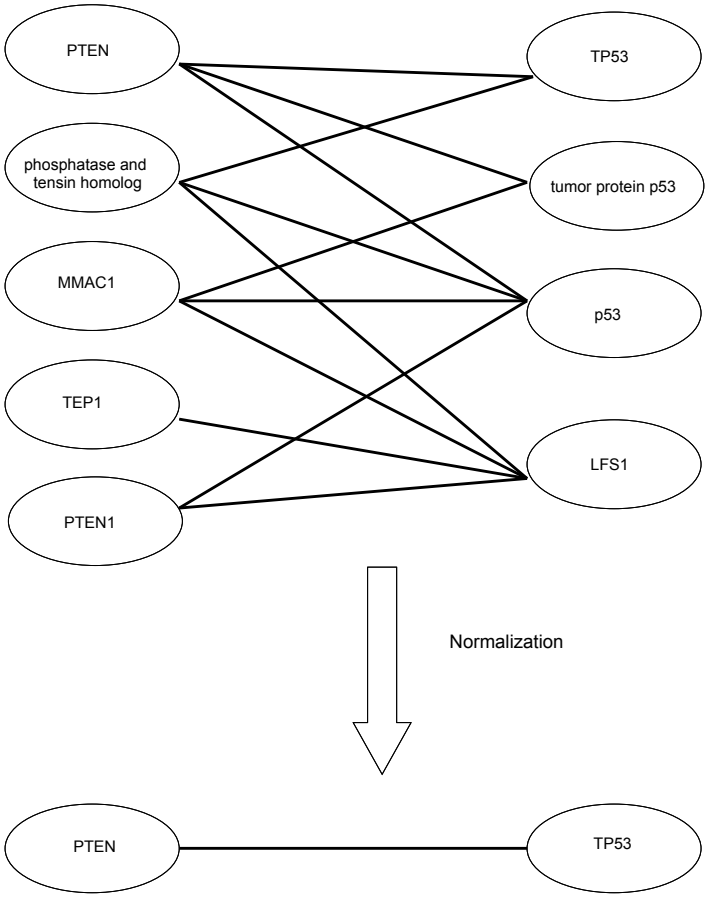


Figure 4.2: Gene name normalization example.

sets. The trained system is used to classify the new sentences as describing an interaction between a gene pair or not.

4.3.5 Network Centrality for Inferring Gene-Disease Associations

Centrality of a node in a graph defines how important a node in the graph is. The importance of a node can be defined in different ways.

Degree Centrality

A graph can be represented by an adjacency matrix A, where $A_{ij} = 1$, if there is an edge between nodes i and j; and $A_{ij} = 0$, if there does not exist an edge between nodes i and j. Degree centrality is the simplest network centrality measure. It only takes into account the degree of a node, which is the number of nodes that a given node is connected to (Freeman, 1979). The degree k_i of node i is calculated as follows.

$$k_i = \sum_{j=1}^{n} A_{ij} \tag{4.1}$$

Degree centrality measures the extent of influence that a node has on the network. The more neighbors a node has, the more important it is.

Eigenvector Centrality

In degree centrality each neighbor contributes equally to the centrality of a node. However, in many real-world situations not all the relationships (connections) between nodes in a network are equally important in determining the centrality of a node. This notion is defined as "prestige" in social networks. Intuitively, the prestige of a person does not only depend on the number of acquaintances he has, but also how prestigious his acquaintances are. A node in a network is more central if it is connected to many central nodes. The centrality x_i of node i is proportional to the sum of the centralities of its neighbors (Newman, 2003):

$$x_i = \lambda^{-1} \sum_{j=1}^{n} A_{ij} x_j \tag{4.2}$$

Let's represent the centralities of the nodes as a vector $\mathbf{x} = (x_1, x_2, ..., x_n)$ and rewrite Equation 4.2 in matrix form.

$$\lambda \mathbf{x} = \mathbf{A}\mathbf{x} \tag{4.3}$$

Here, $\mathbf{x}$ is an eigenvector of the adjacency matrix $\mathbf{A}$ with eigenvalue λ. By Perron-Frobenius theorem, there is only one eigenvector $\mathbf{x}$ with all centrality values non-negative and this is the unique eigenvector that corresponds to the largest eigenvalue λ (Newman, 2003). Eigenvector centrality assigns each node a centrality that not only depends on the quantity of its connections, but also on their qualities.

Closeness Centrality

Closeness centrality of a node measures the centrality of a node based on how close it is to other nodes in the network. The smaller the total distance of a node to other nodes, the higher its closeness is. We calculate the closeness centrality measure x_i for node i by inverting the sum of the shortest distances from it to other nodes in the network (Freeman, 1979) (Equation 4.4).

$$x_i = \left[\sum_{j=1}^{n} d_{ij}\right]^{-1} \tag{4.4}$$

Here, d_{ij} is the geodesic distance (i.e., the length of the shortest path) between node i and node j.

Betweenness Centrality

Betweenness centrality of a node is the number of shortest paths between other nodes that run through the node in interest (Freeman, 1977). For a node i, this measure is computed by taking the sum of the number of shortest paths between pairs of nodes that pass through node i divided by the total number of shortest paths between pairs of nodes (Equation 4.5).

$$x_i = \sum_{j<k}^{n} \frac{g_{jk}(i)}{g_{jk}} \tag{4.5}$$

Here, $g_{jk}(i)$ is the number of shortest paths between nodes j and k that pass through node i. g_{jk} is the total number of shortest paths between nodes j and k.

Betweenness centrality characterizes the control of a node over the information flow of the network. A node is considered central if it appears on many paths that connect pairs of nodes (i.e. it acts as a bridge between pairs of nodes in the network).

4.4 Results and Discussion

4.4.1 Properties of the Prostate Cancer Network

The prostate cancer related gene interaction network consists of 226 nodes (distinct genes) and 1,187 edges (interactions among these genes). The diameter of the network (the longest of the shortest paths between the pairs of genes in the interaction network) is 6 and the average shortest path length (the average of the shortest paths between all genes in the network) is 2.57. The clustering coefficient (Watts & Strogatz, 1998) is 0.4497, which is significantly higher than the clustering coefficient of a random graph with the same number of nodes (0.0487). The prostate cancer network is a small-world network, characterized by having a small average shortest path length and a clustering coefficient that is significantly higher than that of a random network with the same number of nodes. In addition, the network is a scale-free network, which is characterized by having a power-law degree distribution, $P(k) \sim k^{-\gamma}$, where $P(k)$ is the probability that a randomly selected node will have a degree (i.e. number of connections) of k (Albert & Barabási, 2002). The exponent (γ) of the power-law degree distribution of the network is 2.24. The scale-free and small-world characteristics of the network confirm the results of previous PPI

network studies (Chen & Sharp, 2004; Hoffmann & Valencia, 2005; Jeong et al., 2001).

4.4.2 Centrality and Gene-Disease Associations

We used the Prostate Gene DataBase (PGDB) (Li et al., 2003), which is a curated database of genes related to prostate cancer, for the initial evaluation of the methods. In the next sub-section we analyze the most central 20 genes in more detail.

Table 4.2 shows the precisions of the methods for the top ranked n genes, i.e., the percentage of the top ranked "n" genes that are marked by PGDB as being related to prostate cancer. The entire network (226 genes) is the neighborhood of the seed genes and 17.70% of the 226 genes are related to prostate cancer. As the centrality score of the genes decreases (i.e., as "n" increases), the percentage of the genes related to prostate cancer decreases and the performances of the four methods converge to each other. For genes with high centrality, eigenvector, degree, and betweenness metrics achieve similar performances, whereas closeness centrality performs worse than them.

For baseline evaluation, we created a co-occurrence network by linking two genes if they appear in the same sentence and at least one of them is a seed gene. We ranked the genes by the number of connections they make with the seed genes.

Top n	Degree	Eigenvector	Betweenness	Closeness	Baseline
10	80.00	80.00	90.00	70.00	50.00
20	75.00	80.00	70.00	55.00	45.00
30	60.00	63.33	63.33	56.67	43.33
40	55.00	57.50	52.50	47.50	32.50
50	46.00	50.00	48.00	42.00	28.00
75	33.33	36.00	34.67	33.33	34.67
100	26.00	28.00	26.00	27.00	27.00
125	23.20	25.60	23.20	23.30	22.40
150	20.67	22.00	20.00	20.00	18.67
175	18.29	20.57	18.29	18.29	17.14
200	17.50	19.00	18.50	17.00	15.00
226	17.70	17.70	17.70	17.70	13.27

Table 4.2: Percentage of top n genes associated with prostate cancer based on the PGDB database

Betweenness centrality achieves the highest precision (90%) for the top 10 genes. The precision of degree and eigenvector centrality measures is 80%, and the precision of closeness centrality is 70%. The baseline approach performs considerably worse (50% precision).

When we consider the top 20 genes, the highest precision is achieved by eigenvector centrality (80%). Degree centrality follows eigenvector centrality with 75% precision, whereas the precision of betweenness centrality drops to 70% and the precision of closeness centrality drops to 55%. Degree, eigenvector, and betweenness centrality perform significantly better than the baseline method (p-value < 0.05, Fisher's Exact Test (Fisher, 1970)).

To analyze the error tolerance of the gene-disease identification approach, we performed experiments by randomly removing edges from the gene interaction network. When up to 25% of the edges were removed randomly from the network, there was no decrease in the precisions of the centrality metrics for the top 20 genes. An insignificant decrease in the precisions of the metrics was observed when 40% of the edges were removed. The precision of degree centrality dropped by 13.3% (from 75% to 65%), eigenvector centrality by 6.25%, betweenness centrality by 7.14%, and closeness centrality by 9.1%. This shows that the proposed approach is robust against random errors.

4.4.3 Detailed Analysis of the Most Central Genes

For each centrality method, we performed a detailed evaluation for the top 20 ranked genes by finding evidence of their association to the disease from various resources as presented in Table 4.3. The descriptions of the genes are presented in Table 4.4. Seed genes are known to be related to the disease. To verify the newly found (inferred) genes, we first used the PGDB database. If a gene is not marked

by PGDB as being related to prostate cancer, we searched for published articles in the literature that state that the gene is related to prostate cancer and also checked whether the gene appears in the KEGG pathway for prostate cancer[8], which is a manually drawn pathway map of the currently known molecular interaction and reaction network for prostate cancer (Kanehisa & Goto, 2000; Kanehisa et al., 2006, 2010).

12 of the genes in Table 4.3 are confirmed to be related to prostate cancer by using the PGDB database. The centrality methods were able to find four genes, which are not included in PGDB, but were confirmed to be related to prostate cancer by searching for evidence in the literature and in the KEGG pathway for prostate cancer. Two genes (MDM2 and INS) are part of the KEGG pathway for prostate cancer. For these genes we also found articles in the literature that support their association to prostate cancer. For example, Wang et al. (2003) and Zhang et al. (2003) state that "MDM2 has a role in prostate cancer growth via p53-dependent and p53-independent mechanisms". For the INS (insulin) gene, Ho et al. (2003) state that "Polymorphism of the insulin gene is associated with increased prostate cancer risk". Supportive evidence for the association of NR3C1 to prostate cancer is presented by Wei et al. (2007), who show that it is differentially expressed in androgen-independent prostate cancer. For the gene MAPK1, Sarfaraz et al. (2006) state that "apoptosis induced by cannabinoid receptor CB1 and CB2 agonists leads to activation of ERK1/2 leading to G1 cell cycle arrest in prostate cancer cells". Here ERK2 is a synonym of MAPK1. Another article that provides supportive evidence for the MAPK1-prostate cancer association includes the statement "lysophosphatidic acid (LPA), the receptor LPA(1), ERK2 and p38alpha are important regulators for

[8]http://www.genome.ad.jp/kegg/pathway/hsa/hsa05215.html

prostate cancer cell invasion and thus could play a significant role in the development of metastasis" (Hao et al., 2007). For the remaining 7 genes in the table, we found neither positive nor negative evidence for their association to prostate cancer.

Gene	Degree	Eigenvector	Closeness	Betweenness	Evidence
TP53	+	+	+	+	PGDB
BRCA1	+	+	+	+	PGDB
EREG	+	+	+	+	None
AKT1	+	+	+	+	PGDB
MAPK1	+	+	+	+	Literature (Sarfaraz et al., 2006; Hao et al., 2007)
TNF	+	+	+	+	PGDB
CCND1	+	+	+	+	PGDB
MYC	+	+	+	+	PGDB
APC	+	+	−	−	PGDB
CDKN1B	+	+	+	−	PGDB
MAPK8	+	+	+	+	PGDB
NR3C1	−	+	+	−	Literature (Wei et al., 2007)
VEGFA	+	+	+	−	PGDB
MDM2	+	+	+	−	KEGG & Literature (Wang et al., 2003; Zhang et al., 2003)
POLD1	−	−	+	+	None
SNORA62	−	−	+	+	None
CNTN2	−	−	−	+	None
PPA1	−	−	−	+	None
TMEM37	−	−	+	−	None
FZR1	−	−	+	−	PGDB
SSSCA1	−	−	+	−	None
BCL2	+	−	−	−	PGDB
INS	+	−	−	−	KEGG & Literature (Ho et al., 2003)

Table 4.3: Genes inferred by degree, eigenvector, closeness, and betweenness centralities. "+" indicates that the given gene is found by the centrality method with score ranking within the top 20 and "−" indicates that the gene is not among the top 20 genes inferred by the method. Evidences for each gene-disease relationship are confirmed by using PGDB, KEGG pathway for prostate cancer, and published articles (literature).

Table 4.5 lists the definitions used in Table 4.6, which shows the summary of the results for the top 20 genes.

Using degree centrality, among its top 20 ranking genes, 5 genes of the original 15 seed genes are found (AR, BRCA2, CD82, PTEN, and CHEK2). The remaining 15 genes (75% of the top 20 genes) are inferred genes in which we were able to confirm the association of 14 genes (93.33% of the inferred genes) to prostate cancer, except for 1 gene: EREG. For this exceptional gene, we did not find negative nor positive evidence, which implies that the gene may still potentially be a prostate cancer gene.

The result of eigenvector centrality is as successful as degree centrality method

Gene	Description
TP53	tumor protein p53 (Li-Fraumeni syndrome)
BRCA1	breast cancer 1, early onset
EREG	epiregulin
AKT1	v-akt murine thymoma viral oncogene homolog 1
MAPK1	mitogen-activated protein kinase 1
TNF	tumor necrosis factor (TNF superfamily, member 2)
CCND1	cyclin D1
MYC	v-myc myelocytomatosis viral oncogene homolog (avian)
APC	adenomatosis polyposis coli
CDKN1B	cyclin-dependent kinase inhibitor 1B (p27, Kip1)
MAPK8	mitogen-activated protein kinase 8
NR3C1	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)
VEGFA	vascular endothelial growth factor A
MDM2	mouse double minute 2, human homolog of; p53-binding protein
POLD1	polymerase (DNA directed), delta 1, catalytic subunit 125kDa
SNORA62	small nucleolar RNA, H/ACA box 62
CNTN2	contactin 2 (axonal)
PPA1	pyrophosphatase (inorganic) 1
TMEM37	transmembrane protein 37
FZR1	fizzy/cell division cycle 20 related 1 (Drosophila)
SSSCA1	Sjogren's syndrome/scleroderma autoantigen 1
BCL2	B-cell CLL/lymphoma
INS	insulin

Table 4.4: Gene names normalized by Hugo and their description

Seed gene:	A gene, which is one of the prostate cancer genes retrieved from OMIM Morbid Map (i.e., one of the genes in Table 4.1).
Inferred gene:	A non-seed gene.
% of inferred genes:	(# of inferred genes / 20) * 100
Confirmed inferred gene:	An inferred gene found to be related to prostate cancer based on PGDB, KEGG pathway for prostate cancer, and published articles.
% of confirmed inferred genes:	(# of confirmed inferred genes / # of inferred genes) * 100
% of confirmed genes:	((# of confirmed inferred genes + # of seed genes) / 20) * 100

Table 4.5: Definitions used in the evaluation of the top 20 genes

	Degree	Eigenvector	Betweenness	Closeness
# of seed genes	5	6	7	2
# of inferred genes	15	14	13	18
% of inferred genes	75	70	65	90
# of confirmed inferred genes	14	13	8	13
% of confirmed inferred genes	93.33	92.86	61.54	72.22
% of confirmed genes	95	95	75	75

Table 4.6: Summary of the results for the top 20 genes

with 95% of the top ranked 20 genes having supportive evidence. Eigenvector centrality found 6 seed genes (AR, BRCA2, CD82, MXI1, PTEN, and CHEK2) and 14 inferred genes. Out of the 14 inferred genes, 13 are confirmed (92.86% of the inferred genes) and the same gene EREG is not.

Using closeness centrality, we found 2 seed genes (AR and BRCA2) and inferred 18 new genes. 13 of the inferred genes (72.22% of the inferred genes) have evidence which indicate that they are related to prostate cancer and 5 inferred genes (EREG, POLD1, SNORA62, TMEM37, and SSSCA1) do not have such affirmative evidence.

Betweenness centrality found the most seed genes among the four centrality methods. In its result, we have 7 seed genes (AR, BRCA2, CD82, MXI1, PTEN, CHEK2, and KLF6) and 13 inferred genes, of which 8 inferred genes (61.54% of the inferred genes) are verified to have relation to the disease. The five inferred genes that we were not able to confirm are EREG, POLD1, SNORA62, CNTN2, and PPA1.

We observed that degree and eigenvector centrality methods generate highly accurate results; 95% of the top ranked 20 genes are actually related to prostate cancer. They are significantly better than the baseline method in which only 65% of the top 20 genes are prostate cancer genes. We used Fisher's Exact Test (Fisher, 1970) to measure the significance level of the differences in performances between the centrality methods and the baseline method. Degree and eigenvector centrality perform significantly better (p-value < 0.05) than the baseline approach in terms of the percentage of the confirmed genes and confirmed inferred genes. These methods are good candidates for use in practice for mining existing genes related to a particular disease. On the other hand, although closeness and betweenness centrality methods are not statistically significantly better than the baseline method in finding known prostate cancer genes, compared to degree and eigenvector centrality they introduce

more genes that are not currently identified as related to the disease of interest. These methods can be used to generate new hypothesis on gene-disease research, which are candidates for experimental validation. In our experiments, even though we were not able to find evidence of whether gene EREG is related to prostate cancer or not; the fact that all four centrality methods suggest this gene gives more confidence to EREG-prostate cancer relation. We believe that EREG is a strong candidate for prostate cancer gene research.

Our approach of building a disease-specific PPI network by literature mining (including the interactions among the non-seed genes), and applying network centrality measures achieved a higher proportion of non-seed (inferred) genes and a higher accuracy of the inferred genes compared to the previous studies (Chen et al., 2006; Gonzalez et al., 2007) (see Section 4.2). For example, with closeness centrality the proportion of inferred genes is 90% and 72.22% of these inferred genes are correct; with degree centrality the proportion of inferred genes is 75% and 93.33% of these genes are correct.

4.5 Conclusion

We have presented a new approach to predict gene-disease associations based on integrating text mining and network analysis. We collected an initial list of seed genes known to be related to a disease and constructed a disease-specific gene interaction network by extracting the interactions among the seed genes and their neighbors automatically from the biomedical literature by using support vector machines with dependency path edit kernel. Next, we used degree, eigenvector, closeness, and betweenness centrality metrics to rank the genes in the network according to their relevance to the disease. We hypothesized that the genes that are central in the

constructed disease-specific network are likely to be associated with the disease.

We evaluated our approach for prostate cancer and showed that degree and eigenvector centrality metrics achieve highly accurate results (95% of the top 20 genes are actually related to the disease), whereas closeness and betweenness centrality metrics introduce genes that are currently unknown to be related to the disease. We were able to extract genes, which are not marked as being related to prostate cancer by the curated Prostate Gene DataBase (PGDB) even though there are recent articles that confirm the association of these genes with the disease. The proposed approach can be used to extract known gene-disease associations from the literature, as well as to infer unknown gene-disease associations which are good candidates for experimental analysis.

CHAPTER V

Literature-Based Discovery of Vaccine Mediated Gene Interaction Networks

5.1 Introduction

In Chapter IV we introduced a literature-based discovery (LBD) method to infer gene-disease associations, and demonstrated its effectiveness in identifying prostate cancer related genes. In this chapter we present the general framework of this LBD approach and adapt it to find genes that are important for vaccine development.

Figure 5.1 shows the general framework of the proposed LBD approach which integrates literature mining with network centrality analysis. Given a concept of interest and a set of known concept-related genes (seed genes), the goal is to predict novel concept-related genes. First, a gene interaction network is built by automatically extracting the interactions of the seed genes and their neighbors from the literature. Then, network centrality metrics are used to rank the genes in the network. Our underlying hypothesis is that the central genes in this concept-specific network of interactions are also likely to be related to the concept.

In Chapter IV our concept of interest was "prostate cancer" and we started with a set of 15 seed genes known to be associated with prostate cancer. We processed a collection of 48245 full text articles from PubMed Central (PMC) Open Access[1]

[1] http://www.ncbi.nlm.nih.gov/pmc/about/openftlist.html

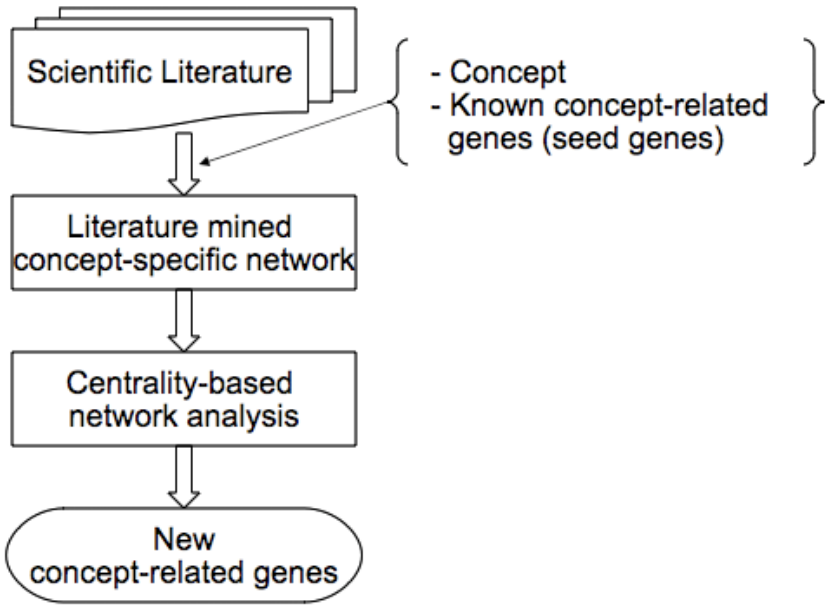


Figure 5.1: General framework of the literature-based discovery approach.

to build the prostate cancer-specific gene interaction network. We were able to identify genes that are not marked as being related to prostate cancer by the curated OMIM (OMIM, 2007) or PGDB (Li et al., 2003) databases even though there are recent articles that confirm their association to the disease.

In this chapter, our concept of interest is "vaccine". To the best of our knowledge, we present the first literature-based discovery study in the vaccine informatics domain. We use only one gene, interferon gamma (synonyms: IFNG, IFN-γ), as a seed gene and analyze all the abstracts indexed in PubMed[2] (over 19 million) to discover novel vaccine-related genes. We build two gene interaction networks by extracting the interactions of IFN-γ and its neighbors from abstracts in PubMed using the method presented in Chapter II. The first network is the generic IFNG network, which is the network of interactions of IFNG and its neighbors. The strategy used to build the prostate cancer gene interaction network in Chapter IV is used to build this network. The second network is the vaccine-specific subgraph of the first network

[2]http://www.ncbi.nlm.nih.gov/pubmed/

(IFNG-vaccine network), which is built using only the interactions that are extracted from vaccine relevant sentences. We use the concept term "vaccine" and its variants to identify these sentence. Analysis and comparison of these two types of networks using network centrality methods provides new insights and hypotheses worth future investigations. The results support our hypothesis that the central genes in the two IFN-γ networks are related to the functions of IFN-γ and part of the gene list are important for vaccine development. Many predicted genes and gene networks are good candidates for further IFN-γ and vaccine development studies.

We also investigate incorporating concept ontology support to our LBD method. We create a third network by using terms from tthe Vaccine Ontology (VO)[3] (He et al., 2009) besides the concept term "vaccine" and its variants. This network (IFNG-vaccine-VO network) is a sub-graph of the generic IFNG network and contains the IFNG-vaccine network. Our results indicate that VO support facilitates the literature-based discovery of vaccine-associated genes. This chapter is based on the work published as (Özgür et al., 2010a, 2010b).

5.2 Biological Motivation

In 1965 Wheelock et al. first reported Interferon-gamma (IFN-γ)-like virus inhibitor, induced in supernatant fluid of cultures of fresh human leukocytes following incubation with phytohemagglutinin (Billiau & Matthys, 2009). In early 1970s, IFN-γ was further studied, and its name was eventually designated. IFN-γ is the only type II IFN family member. It is secreted by activated immune cells - primarily T and NK cells, but also B-cells, NKT cells and professional antigen presenting cells. IFN-γ has been widely studied and found critical in anti-infectious host defense, inflammatory conditions, cancer, and auto-immune diseases (Billiau & Matthys, 2009;

[3] http://www.violinet.org/vaccineontology/

Wieder et al., 2008). The most striking phenotype from mice lacking either IFN-γ or its receptor has increased susceptibility to bacterial and viral pathogens (Schroder et al., 2004). IFN-γ is also critical for tumor immuno-surveillance as assessed using spontaneous, transplantable and chemical carcinogen-induced experimental tumors. Additionally, IFN-γ is found important in leukocyte homing, cellular adhesion, immunoglobulin class switching, T helper cell polarity, antigen presentation, cell cycle arrest and apoptosis, neutrophil trafficking and NK cell activation (Billiau & Matthys, 2009; Gough et al., 2008; Takayanagi et al., 2005).

The induction of IFN-γ response is critical for successful development of vaccines against various viruses and intracellular bacteria, for example, human immunodeficiency virus (HIV) (Streeck et al., 2009), Mycobacterium tuberculosis (Fletcher, 2007), Leishmania spp. (Mansueto et al., 2007), and Brucella spp. (He et al., 2001, 2002). The IFN-γ analysis is widely used for the quantification and characterization of the HIV-specific CD8+ T cell responses (Streeck et al., 2009). It is a marker used as a representative function of cytotoxic T cells to quantify the HIV-specific cellular immune response. IFN-γ is required for protection against mycobacterial infection (Wallis et al., 2009). M. tuberculosis-stimulated whole-blood production of IFN-γ, although imperfect, is the best available correlate of protective immunity to M. tuberculosis in humans (Fletcher, 2007). In humans, complete IFN-γR deficiency is associated with frequent infection and ultimately death from the attenuated M. tuberculosis BCG vaccine (Jouanguy et al., 1996). The inability to secrete IFN-γ or the development of auto-antibodies neutralizing endogenous IFN-γ resulted in the death of a patient by overwhelming mycobacterium infection (Doffinger et al., 2004).

Today IFN-γ is ranked as one of the most important endogenous regulators of immune responses. Thousands of relevant papers have been published. However, a

comprehensive understanding of how it works and what other factors it interacts with is still largely unclear. Although IFN-γ is essential for protective immunity, animal and human studies have found that IFN-γ alone is not sufficient for the prevention of tuberculosis disease (Fletcher, 2007). Our goal is to analyze the network of IFN-γ with other genes through literature mining and investigate what other genes or gene interaction networks are needed to stimulate protective immunity. Since IFN-γ is one of the most important immune factors and critical for vaccine development, we hypothesized that genes central in the networks built around IFN-γ might be important for vaccine development as well.

5.3 Methods

The details of the literature-based discovery approach to predict new concept-related genes were presented in Chapter IV in the context of identifying gene-disase associations. In this section, we summarize the main steps of applying this approach to discover genes important for vaccine research (Figure 5.2).

5.3.1 Literature corpus

In Chapter IV we used 48,245 full text articles from PubMed Central (PMC) Open Access to extract the prostate cancer gene interaction network. In this chapter we use all article abstracts available in PubMed to construct the literature-mined IFN-γ gene interaction network. The sentences of the abstracts are obtained from the BioNLP database in the National Center for Integrative Biomedical Informatics (NCIBI)[4], which were generated using the MxTerminator sentence boundary detection tool (Reynar & Ratnaparkhi, 1997).

We tagged the gene names using Genia Tagger (Tsuruoka et al., 2005) and nor-

[4] http://ncibi.org/

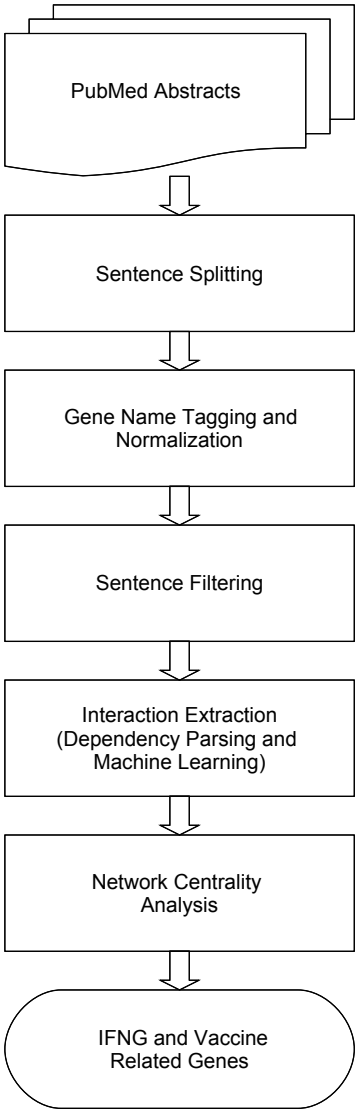


Figure 5.2: Description of the literature-based discovery system for identifying IFN-γ and vaccine related genes.

malized them using the HUGO Gene Nomenclature Committee (HGNC) database[5] (Wain et al., 2004). Each tagged gene name was unified with its corresponding approved gene symbol[6]. In the HGNC database, the official gene symbol for the IFN-γ gene is listed as IFNG, and the description is listed as "interferon, gamma". The database does not include any synonyms for the gene. However, IFN-γ is frequently mentioned in text with the names "interferon-gamma", "interferon gamma", "IFN-gamma", and "IFNgamma". Therefore, we included these names to the HGNC dictionary as synonyms for IFN-γ.

5.3.2 Gene interaction extraction from the literature

To extract the IFN-γ (IFNG) gene-interaction network from the literature, we used path edit kernel with SVM (see Chapter II), which is the same method that we used to extract the prostate cancer gene interaction network in Chapter IV. The system was trained by combining the AIMED and CB data sets.

Before classifying the sentences in the literature corpus as describing an interaction between a gene pair or not, the potential interaction sentences were selected from the abstracts in PubMed that have "human" in the MeSH heading. We extended the list of interaction keywords described in Chapter IV to include 826 interaction keywords such as binds, bound, interacts, activates, inhibits, and phosphorylates[7]. Our assumption is that a sentence that describes an interaction between a pair of genes should contain an interaction keyword and at least two distinct normalized gene names. The sentences that do not meet this requirement were filtered out.

The IFNG gene-interaction network was built in two steps. In the first step, the genes that interact with IFNG (i.e., the neighbors of IFNG) were extracted.

[5] http://www.genenames.org/index.html
[6] As of October, 2009 the database contains 28,240 approved gene records
[7] The list of interaction keywords is available at: http://clair.si.umich.edu/clair/ifngnet/interaction_keywords.txt

The number of sentences that contain IFNG or one of its synonyms (case-insensitive match) and are from abstracts that have "human" in the MeSH headings is 73,024. We filtered out those sentences that don't have at least one interaction keyword and at least two distinct normalized gene names, one of which is IFNG. As a result, 26,876 sentences were analyzed with our interaction extraction module for identification of the genes that interact with IFNG. The interaction extraction module extracted 1059 neighbors of IFNG.

In the second step, the interactions among the neighbors of IFNG were extracted. There are over 9 million sentences that are from abstracts which have "human" in the MeSH headings and contain at least one of the IFNG neighbors or their synonyms. Out of these, the sentences for further processing by the interaction extraction module are those that have at least one interaction keyword, and at least two distinct normalized gene names, which were identified as neighbors of IFNG in the first step. In total, 422,566 sentences met these criteria and were further processed by the interaction extraction module.

5.3.3 Network centrality analysis

We build the *IFNG network* by representing IFNG and its neighbors as nodes and connecting two genes with an edge if we have extracted an interaction between them from the literature. The gene names in the network are normalized and represented with their official HGNC symbols. We also create a vaccine-specific subgraph of this network, i.e., the *IFNG-vaccine network*. This network contains only the interactions that have been extracted from sentences that contain the term "vaccin", which is the root form of the vaccine related terms such as vaccine, vaccines, vaccination, and vaccinated. Therefore, the edges in this subgraph are all vaccine specific. Analysis of this IFNG-vaccine network helps us understand the genes and interactions that

play important roles in both the vaccine and IFNG network. We analyze the two literature-mined IFNG interaction networks using the degree, eigenvector, betweenness, and closeness centrality methods that were discussed in Chapter IV. Since IFNG is one of the most important immune factors and critical for vaccine development, we hypothesized that genes central in the generic IFNG and IFNG-vaccine networks might be important for vaccine development. The results presented in the next section support the hypothesis.

5.3.4 Gene annotation enrichment analysis

The web-based DAVID bioinformatics program was used to perform the gene annotation enrichment analysis (Huang et al., 2009).

5.4 Comparison of the IFNG and IFNG-vaccine Networks

5.4.1 Topological properties of the networks

Our method detected 1060 nodes (genes including IFNG and its neighbors) linked by $26,313$ edges (interactions) (Figure 5.3). Since all the genes in the IFNG network are connected to IFNG, the diameter of the network is 2 and the average shortest path length is 1.95. The clustering coefficient of the network is 0.4933, which is an order of magnitude higher than the clustering coefficient of a random network with the same number of nodes (0.0473). The IFNG network is a small-world network, characterized by having a small average shortest path length and a clustering coefficient that is significantly higher than that of a random network with the same number of nodes. The IFNG network is a scale-free network with a power-law degree distribution, where the exponent γ is 2.15. The graph of the IFNG network is shown in Figure 5.4.

The IFNG and vaccine-associated network (IFNG-vaccine network) is a much smaller subset of the generic IFNG network. This small subnetwork contains 102

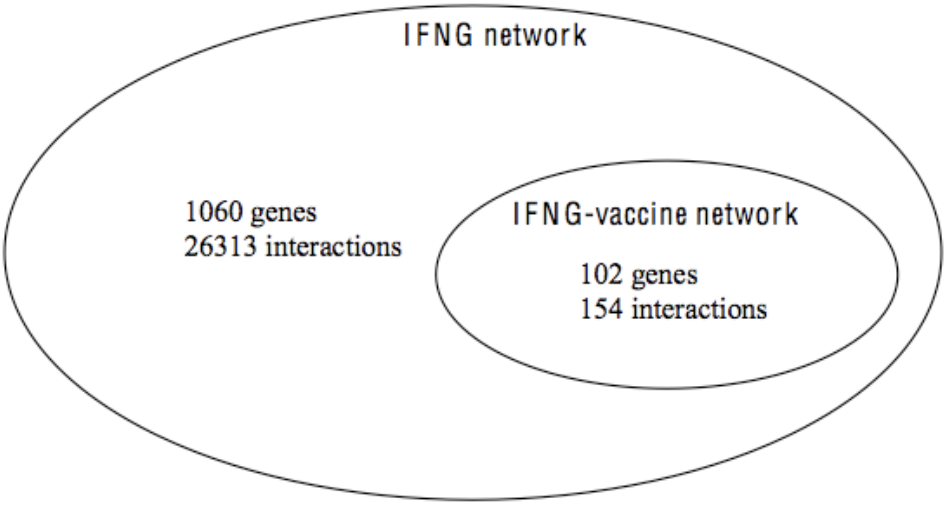


Figure 5.3: Summary of the IFNG network and its vaccine-specific subnetwork

genes and 154 interactions (Figure 5.3). Since the IFNG-vaccine network is built by removing the edges that are not associated with "vaccine" from the IFNG network, some of the genes that were connected in the IFNG network are not connected in the IFNG-vaccine network. Therefore, the IFNG-vaccine network contains 84 genes that are interconnected and 18 genes that are separated from this largest connected component of 84 genes (Figure 5.5). Also, the diameter of the IFNG-vaccine network and the average shortest path length are larger than those of the IFNG network. The diameter of the IFNG-vaccine network is 9 and the average shortest path length is 3.55. The IFNG-vaccine network still possesses the small-world property with a relatively small average shortest path length and a clustering coefficient (0.2218) that is significantly higher than the clustering coefficient of a random network with the same number of nodes (0.0388). The network is scale-free with a power-law degree distribution with exponent 2.37. The small-world and scale-free characteristics of the generic IFNG and the IFNG-vaccine networks are consistent with the topological properties of the prostate cancer network presented in Chapter IV as well as with

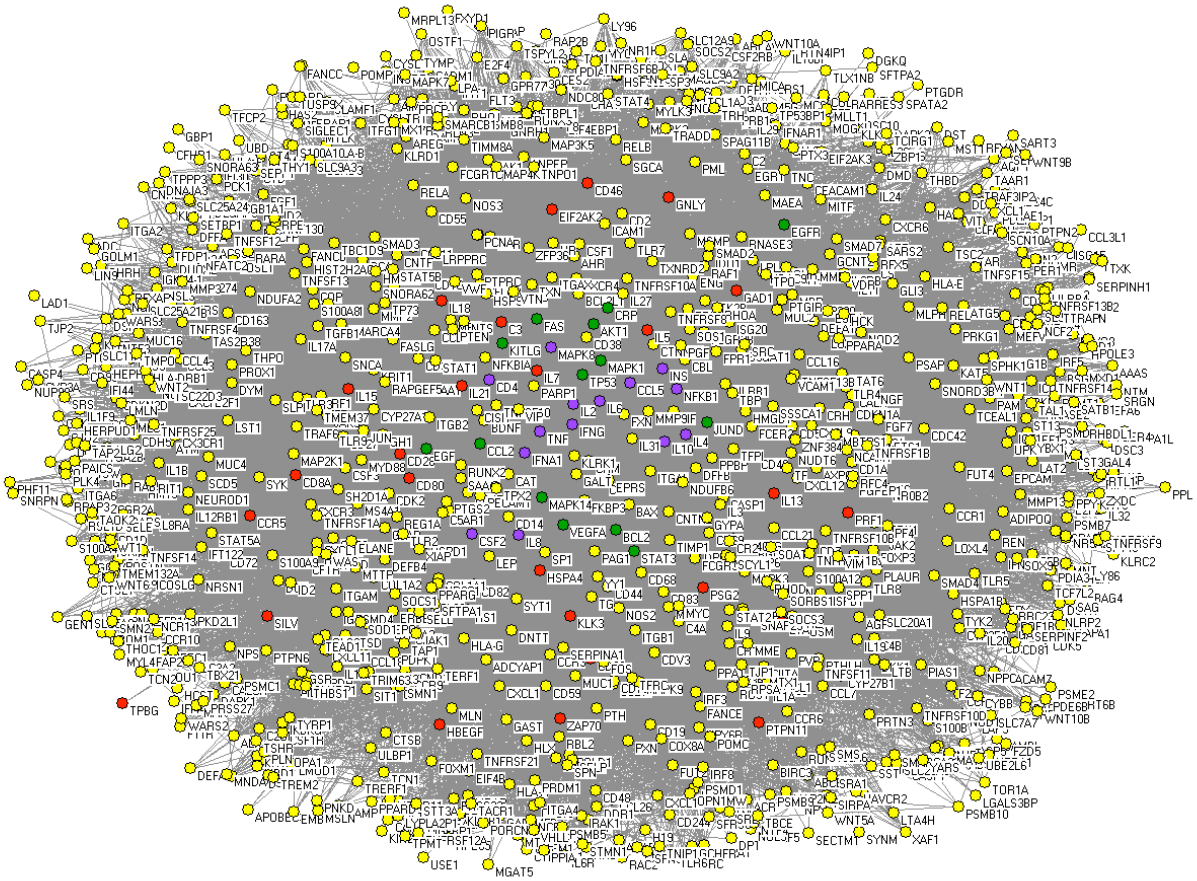


Figure 5.4: The graph of the generic IFNG network extracted from the literature. The network consists of 1060 nodes (genes) and 26,313 edges (interactions). The purple nodes are the genes that are central in both the generic and the IFNG-vaccine networks. The green nodes are the genes that are central in only the generic IFNG network and the red nodes are the genes that are central in only the IFNG-vaccine network. The rest of the nodes are shown in yellow.

previously studied biological networks (Jeong et al., 2001; Chen & Sharp, 2004; Hoffmann & Valencia, 2005) and non-biological networks such as the Internet (Yook et al., 2002) and social networks (Watts & Strogatz, 1998).

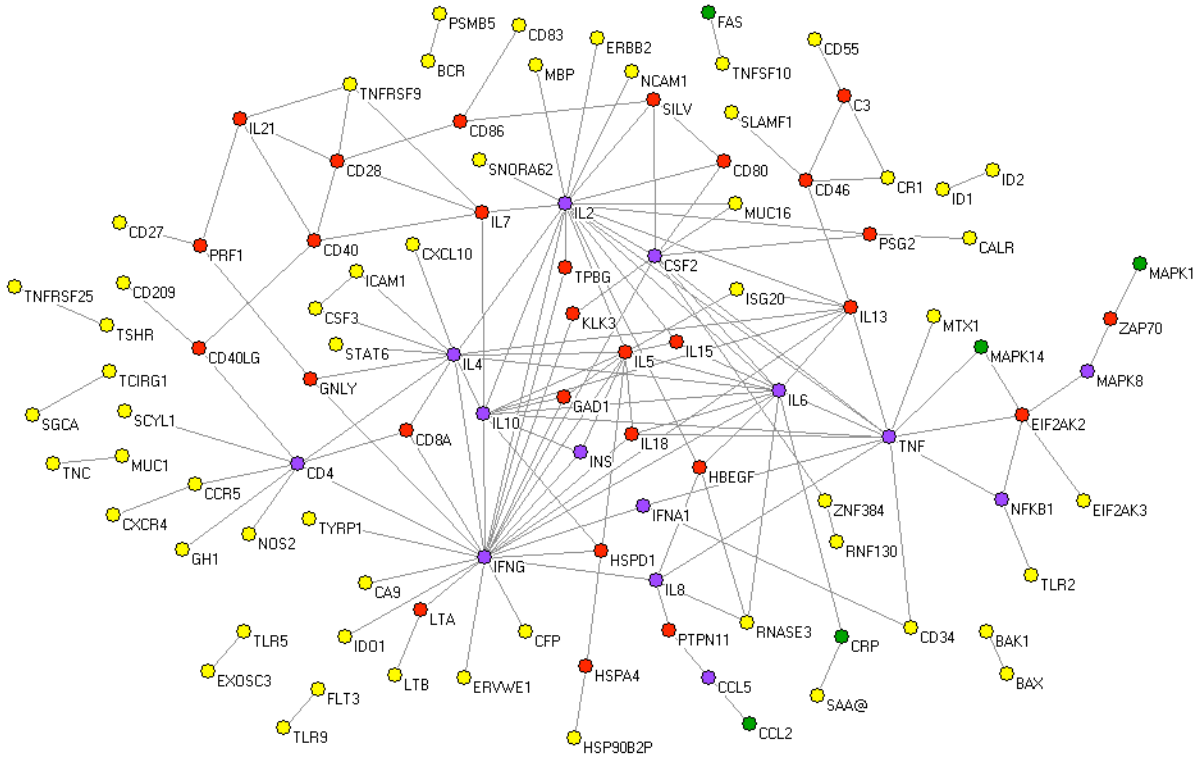


Figure 5.5: The graph of the IFNG-vaccine network extracted from the literature. The network consists of 102 nodes (genes) and 154 edges (interactions). All the edges in the network are associated with the term "vaccine" and its variants. The purple nodes are the genes that are central in both the generic and the IFNG-vaccine networks. The red nodes are the genes that are central only in the IFNG-vaccine network. The green nodes are the genes that are central only in the generic IFNG network. The rest of the nodes are shown in yellow.

5.4.2 Lists of genes are predicted and sorted by centrality analyses

All the genes in the two networks (generic IFNG network and IFNG-vaccine network) are sorted based on centrality analyses. The files that list the rankings of all the genes in the generic IFNG network and the genes in the IFNG-vaccine network are available at the following URLs, respectively.

- http://www.hindawi.com/journals/jbb/2010/426479.f2.pdf
- http://www.hindawi.com/journals/jbb/2010/426479.f3.pdf

IFNG is not included in these rankings, since it is trivially ranked highest by all the centrality measures in both networks due to the fact that the networks are specific to IFNG. The most central genes (the genes ranked among the top 25 by at least one of the centrality measures) are analyzed in more detail in Table 5.1. These genes (a total of 56 genes) are predicted to be associated with IFNG and relevant for vaccine development. Literature evidence was manually curated for the IFNG association (IFNG-Ref column in Table 5.1) and the vaccine development relatedness (Vaccine-Ref column in Table 5.1) of these genes.

It is interesting that in the generic IFNG network, all centrality measures find the same 23 genes among the top 25, although the ranking might change slightly (Table 5.1). For example IL10 is ranked 5^{th} by degree and closeness centralities, but 4^{th} by eigenvector and betweenness centralities. Since all the genes in the generic IFNG network are connected to IFNG, the distance (shortest path length) between a pair of genes is at most two. In other words, the distance between a pair of genes is one if they are directly connected to each other and it is two if they are not directly connected to each other (i.e., they are connected through IFNG). Therefore, in this network, the more genes a gene is connected to (higher degree centrality), the less distant it is to the other genes (higher closeness centrality). So, the degree and closeness centralities produce the same rankings for the generic IFNG network. For the IFNG-vaccine network, the top 25 genes sorted based on centrality analyses overlapped with the sorted results from the generic IFNG network.

Three different levels of prediction are available based on the comparison between the generic IFNG network and the more specific IFNG-vaccine network:

	Generic IFNG Network					IFNG-vaccine Network				
Gene	D	E	B	C	IFNG-Ref	D	E	B	C	Vaccine-Ref
TNF	1	1	1	1	3132506	2	3	2	7	16446013
NFKB1	2	2	2	2	9888423	-	23	-	-	16971487
IL6	3	3	3	3	1719090	3	4	7	3	10225849
IL8	4	5	6	4	8473010	10	13	10	9	11378044
IL10	5	4	4	5	8102388	6	8	11	2	10930151
IL4	6	6	5	6	2136895	4	2	4	4	8519092
MAPK1	7	9	9	7	15307176	-	-	-	-	19428911
IL2	8	7	8	8	6429853	1	1	1	1	8459207
VEGFA *	9	10	10	9	12816689	-	-	-	-	17502972
TP53 *	10	8	7	10	16391798	-	-	-	-	18846387
BCL2 *	11	13	13	11	11064392	-	-	-	-	19389797
AKT1 *	12	11	12	12	11135576	-	-	-	-	19107122
MAPK8	13	14	14	13	18950753	-	-	15	-	19428911
INS	14	12	11	14	8383325	-	-	-	16	19203100
MAPK14	15	15	18	15	10700460	-	-	-	-	19428911
CSF2	16	18	17	16	11665752	7	6	6	6	19459853
FAS	17	17	16	17	10895367	-	-	-	-	15979942
CCL2	18	19	19	18	9407497	-	-	-	-	19833737
IFNA1	19	16	15	19	11449378	-	-	-	13	19667099
EGFR *	20	20	23	20	17362940	-	-	-	-	19178753
JUND *	21	21	22	21	10070035	-	-	-	-	19124729
KITLG *	22	24	-	22	7540064	-	-	-	-	-
CCL5	23	23	21	23	8921438	-	24	-	-	15827150
CD4	24	22	20	24	15173593	9	5	3	12	17298856
EGF *	25	25	-	25	18160214	-	-	-	-	16357522
CRP	-	-	24	-	10675363	-	-	-	-	16395099
STAT3 *	-	-	25	-	7488223	-	-	-	-	-
IL5	-	-	-	-	9432015	5	7	20	8	11138639
IL13	-	-	-	-	12670721	8	9	5	5	12232042
IL7	-	-	-	-	7594482	11	14	12	17	17496983
EIF2AK2	-	-	-	-	11342638	12	10	8	-	19596385
CD28	-	-	-	-	7634349	13	12	-	-	12594842
HSPD1	-	-	-	-	12407015	14	19	16	14	12218165
SILV	-	-	-	-	11839572	15	20	17	23	11459172
IL21	-	-	-	-	14657853	16	17	-	-	16785513
IL18	-	-	-	-	8666798	17	-	-	10	19467215
HBEGF	-	-	-	-	9062364	18	25	-	21	10729731
CD46	-	-	-	-	15307176	19	11	9	-	11757799
CD40	-	-	-	-	7554483	20	16	-	-	11403919
PSG2	-	-	-	-	2516715	21	-	22	-	11155821
GAD1	-	-	-	-	9703171	22	-	-	18	12421990
IL15	-	-	-	-	9834271	23	-	-	22	16785513
C3	-	-	-	-	1337336	24	15	-	-	19477524
PRF1	-	-	-	-	19651871	25	22	19	-	15214037
ZAP70	-	-	-	-	11034358	-	18	23	-	-
CD40LG	-	-	-	-	10769003	-	21	18	-	11403919
GNLY	-	-	-	-	17382591	-	-	13	19	10644038
PTPN11	-	-	-	-	12270932	-	-	14	-	-
CD86	-	-	-	-	9836505	-	-	21	-	12594842
CCR5	-	-	-	-	9616137	-	-	24	-	16672545
HSPA4	-	-	-	-	18442794	-	-	25	-	11779704
TPBG	-	-	-	-	16630022	-	-	-	11	16630022
KLK3	-	-	-	-	16000955	-	-	-	15	19171173
CD8A	-	-	-	-	1904117	-	-	-	20	18425263
CD80	-	-	-	-	7537534	-	-	-	24	10498243
LTA	-	-	-	-	3102976	-	-	-	25	15908422

Table 5.1: Predicted 56 genes related to IFN-γ and vaccine networks. The genes that are ranked among the top 25 by the centrality measures (D: Degree; E: Eigenvector; B: Betweenness; C: Closeness) in the generic IFNG network or the IFNG-vaccine network. The genes are represented with their official HGNC symbols. Literature evidences for the relatedness of the genes to IFNG (IFNG- Ref) and to vaccine development (Vaccine-Ref) are manually curated. "-" indicates that the gene is not ranked among the top 25 by the corresponding centrality measure in the corresponding network or no literature evidence was found.

(i) Genes ranked high in both networks

Thirteen genes were ranked among the top 25 in both networks by at least one of the centrality measures. Among these 13 genes, 8 genes are central by all centrality measures in both networks: TNF, IL6, IL8, IL10, IL4, IL2, CSF2, and CD4. These genes are well studied in both generic IFNG research and vaccine specific research. The ranking may change in both networks. For example, IL2 was ranked top 1 in the IFNG-vaccine network, while it was ranked top 7-8 in the generic IFNG network based on different centrality scores. This is probably due to the fact that the role of IL2 in vaccine research has widely been recognized and studied in more depth in the vaccine context.

Among the 13 genes in this group, five genes (NFKB1, MAPK8, INS, IFNA1, and CCL5) were ranked high in the IFNG network by all measures but only high in the IFNG-vaccine network by certain centrality measures. For example, MAPK8 (mitogen-activated protein kinase 8; Aliases: JNK, JNK1, SAPK1) was ranked high by all centrality metrics in the IFNG network, whereas it was ranked high by only the betweenness centrality metric in the IFNG-vaccine network (Table 5.1). The high betweenness score was reflected by the fact that MAPK8 connects the two genes (ZAP70 and MAPK1) to the rest of the network (Figure 5.5). In the generic IFNG network, 322 other genes are directly connected to MAPK8 (Figure 5.6). Many of these genes (e.g., NFKB1, IL4, and CD40) also exist in the IFNG-vaccine network (Figure 5.5) although they do not directly interact with MAPK8. However, the majority of these 322 genes (e.g., TLR4 and IL1B) are not in the IFNG-vaccine network. It is reasonable to suggest that many of these genes that were found in the IFNG-MAPK8 network (Figure 5.6) but not in the IFNG-vaccine network (Figure 5.5) may also be important for vaccine specific network through an interaction

with MARK8. Therefore, the comparison between these two networks may lead to hypothesis of new genes involved in vaccine specific immune network, some of which deserve further experimental verifications.

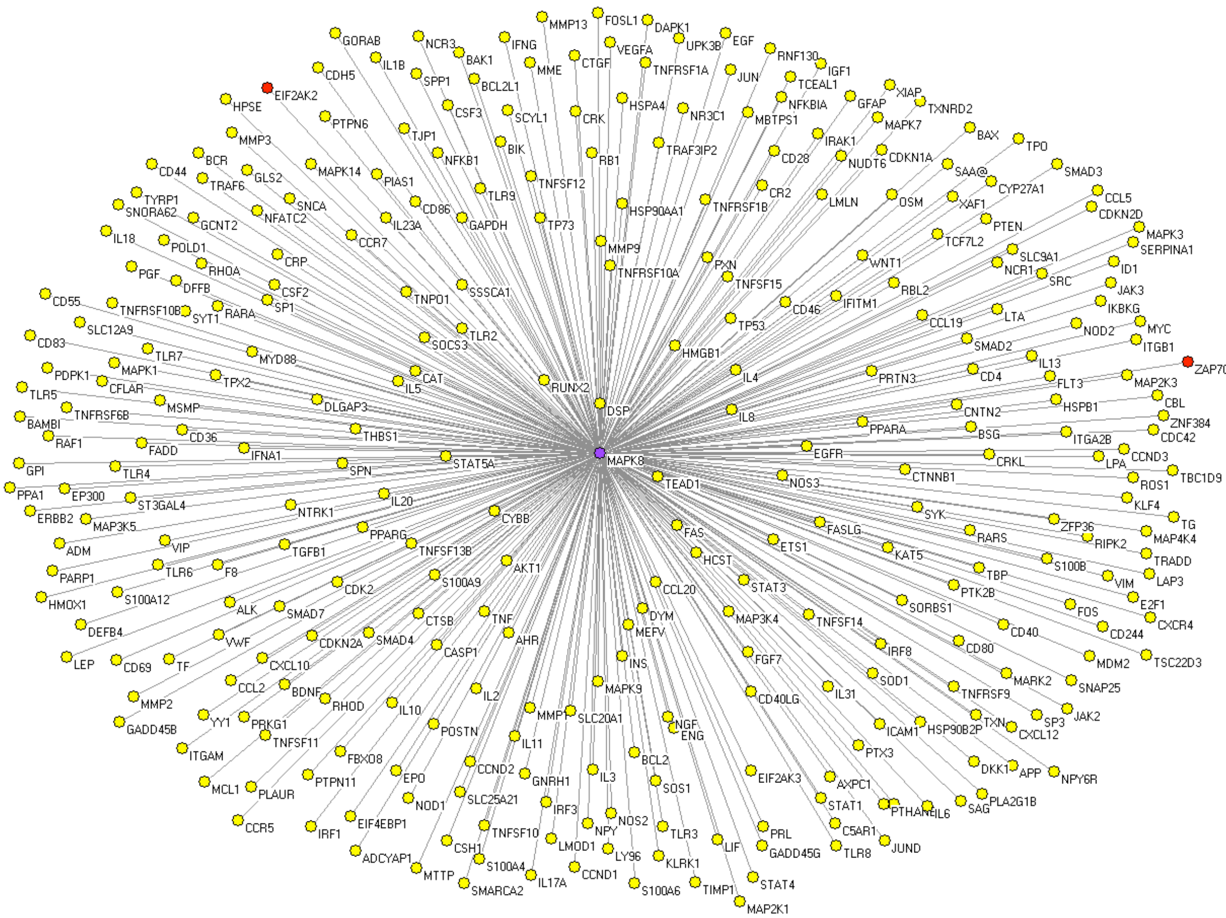


Figure 5.6: Interactions of MAPK8 with other genes in the generic IFNG network (the IFNG-MAPK8 network). MAPK8 is shown in purple. The two genes that MAPK8 also interacts in the IFNG-vaccine network are shown in red.

(ii) Genes ranked high in the generic IFNG network but not in the IFNG-vaccine network

In total 14 genes are included in this group. Nine out of these 14 genes were not found in the IFNG-vaccine network. These genes are labeled with "*" in Table 5.1. These genes have not been well studied in the vaccine context. However, since these

genes are strongly associated with IFNG, it is likely that each of these genes may also play an important role in vaccine-induced protective immune network. For example, as one of the 14 genes, the serine/threonine kinase AKT1 is a key regulator of cell proliferation and death. AKT1 regulates lymphocyte apoptosis and Th1 cytokine propensity (Bommhardt et al., 2004). IFNG is a representative cytokine in Th1 response that is crucial for induction of vaccine-induced protection. Therefore, it is reasonable to hypothesize that AKT1 plays an important role in regulated vaccine-induced protective immune responses.

Among the 14 genes in this group, five genes (MAPK1, MAPK14, FAS, CCL2, and CRP) were found in the IFNG-vaccine network but not ranked high based on any centrality analysis. For example, FAS is a critical gene in regulation of programmed cell death through the FAS pathway. FAS (TNF receptor superfamily, member 6; Aliases: CD95, APO-1) has been found to play an important role in promoting an appropriate effector response following vaccinations against Helicobacter pylori (Avitzur et al., 2005), hepatitis C virus (Langhans et al., 2005), and cancer (Shi et al., 2005). Since FAS is well studied and ranked top in the generic IFNG network, more knowledge about its interactions with other genes shown form the generic IFNG network provides valuable basis for further analysis of FAS-related, vaccine-specific interaction network.

(iii) Genes ranked high in the IFNG-vaccine network but not in the generic IFNG network

In total, 29 genes that were ranked among the top 25 in the IFNG-vaccine network based on at least one of the centrality scores are not ranked among the top 25 in the generic IFNG network (Table 5.1). These genes may be more vaccine-specific and play relatively less important roles in many other IFNG-regulated immune systems

(e.g., cell cycle). It is also possible that some of these genes are very important for other IFNG-related immune functions. In that case, the data for these genes obtained from vaccine research may provide supportive results for expanded studies. One important set of these 29 genes cover many interleukins including IL5, IL7, IL13, IL15, IL18, and IL21. For example, interleukin-18 (IL18) is a newly discovered cytokine with profound effects on T-cell activation. IL18 can possibly be used as a strong vaccine adjuvant (Dinarello, 1999). The new knowledge obtained from IL18 in vaccine research may be applied to other IFNG-related immune systems.

5.4.3 Gene annotation enrichment shows various immune responses regulated by IFN-γ

The 56 genes ranked among the top 25 by at least one of the centrality methods in one or both networks were used for gene enrichment analysis using DAVID (Huang et al., 2009). These genes were classified in various immune mechanisms such as response to extracellular stimulus, lymphocyte activation, and regulation of apoptosis (Table 5.2). These gene annotation enrichment results are correlated with current knowledge about IFN-γ (Billiau & Matthys, 2009; Gough et al., 2008; Takayanagi et al., 2005). It further demonstrates the capability of our literature-based discovery approach in correctly extracting genes related to IFN-γ.

5.5 Vaccine Ontology Support

We were able to generate many new observations and hypotheses by comparing the generic IFNG network and its vaccine-specific subnetwork (IFNG-vaccine). It is possible to further improve the literature-based network discovery by applying biomedical ontologies. A biomedical ontology represents the consensus-based controlled vocabularies of terms and relations which are logically formulated to pro-

Category	Term	Count	P-Value	FDR
GOTERM_BP_ALL	GO:0050896 ~ response to stimulus	43	2.99E-22	5.71E-19
GOTERM_BP_ALL	GO:0007154 ~ cell communication	39	5.74E-13	1.10E-09
GOTERM_BP_ALL	GO:0007165 ~ signal transduction	35	9.70E-11	1.86E-07
GOTERM_BP_ALL	GO:0006950 ~ response to stress	29	7.14E-20	1.37E-16
GOTERM_BP_ALL	GO:0030154 ~ cell differentiation	28	6.94E-13	1.33E-09
GOTERM_BP_ALL	GO:0006952 ~ defense response	26	7.12E-23	1.36E-19
GOTERM_BP_ALL	GO:0006955 ~ immune response	26	8.88E-18	1.70E-14
GOTERM_BP_ALL	GO:0008283 ~ cell proliferation	23	9.37E-16	1.70E-12
GOTERM_BP_ALL	GO:0008219 ~ cell death	23	2.28E-15	4.46E-12
GOTERM_BP_ALL	GO:0006915 ~ apoptosis	22	9.38E-15	1.78E-11
GOTERM_BP_ALL	GO:0007242 ~ intracellular signaling cascade	19	4.85E-07	9.27E-04
GOTERM_BP_ALL	GO:0001775 ~ cell activation	18	5.86E-19	1.12E-15
GOTERM_BP_ALL	GO:0006954 ~ inflammatory response	17	8.26E-16	1.49E-12
GOTERM_BP_ALL	GO:0046649 ~ lymphocyte activation	14	1.77E-14	3.38E-11
GOTERM_BP_ALL	GO:0006468 ~ protein amino acid phosphorylation	14	2.60E-07	4.98E-04
GOTERM_BP_ALL	GO:0006807 ~ nitrogen compound metabolic process	13	4.47E-08	8.56E-05
GOTERM_BP_ALL	GO:0042110 ~ T cell activation	12	9.02E-14	1.73E-10
GOTERM_BP_ALL	GO:0048534 ~ hemopoietic or lymphoid organ development	12	4.57E-11	8.74E-08
GOTERM_CC_ALL	GO:0005576 ~ extracellular region	29	5.33E-18	8.27E-15
GOTERM_MF_ALL	GO:0005125 ~ cytokine activity	19	5.30E-21	9.51E-18
GOTERM_MF_ALL	GO:0008083 ~ growth factor activity	12	6.77E-12	1.21E-08
KEGG_PATHWAY	hsa04060: Cytokine-cytokine receptor interaction	23	7.90E-16	9.77E-13
KEGG_PATHWAY	hsa04620: Toll-like receptor signaling pathway	13	3.12E-10	3.91E-07
KEGG_PATHWAY	hsa04660: T cell receptor signaling pathway	12	2.04E-09	2.57E-06
KEGG_PATHWAY	hsa04630: Jak-STAT signaling pathway	11	2.99E-06	0.003745

Table 5.2: Gene annotation enrichment among top predicted genes in the generic IFNG and the IFNG-vaccine networks.

mote intelligent information retrieval and modeling. The Vaccine Ontology (VO) is a community-based ontology in the domain of vaccine and vaccination[8] (He et al., 2009). VO has classified a large number of existing vaccines in licensed use, on trial, or in research. Each subclass in VO has an "is_a" relationship with its parent class. This ensures that all vaccine subclasses (e.g., BCG) can be included when a parent class (e.g., "Mycobacterium tuberculosis vaccine" or "vaccine") is searched in literature mining. Currently, VO contains more than 400 vaccine names.

As discussed in the previous section the IFNG-vaccine subgraph of the generic IFNG network contains only the interactions that have been extracted from sentences that contain the term "vaccine" (or its variants like "vaccines", "vaccination", and "vaccinated"). However, there are many vaccine-related sentences in the literature where the term "vaccine" or its variants do not occur. Consider the sentence "These

[8]http://www.violinet.org/vaccineontology

results suggest that the BCG-CWS induces TNF-alpha secretion from DC via TLR2 and TLR4 and that the secreted TNF-alpha induces the maturation of DC per se" from (Tsuji et al., 2000). The term "vaccine" or its variants do not occur in the sentence or in the abstract. However, this sentence is vaccine-related, since "BCG" (Bacillus Calmette-Guerin) is a licensed tuberculosis vaccine. The "BCG" vaccine is included in the VO.

In this section, we investigate whether incorporating the Vaccine Ontology to our LBD system will enhance the literature-based discovery of IFN-γ and vaccine-mediated gene interaction networks. We extended the IFNG-vaccine network by including the interactions that have been extracted from sentences that contain one of the vaccine names included in the VO. The vaccine names that contain the term "vaccine" were filtered out, since this term is explicitly included in the query for selecting the vaccine-related sentences. In total 197 vaccines, which are the leaf nodes under the "vaccine" ontology hierarchy, were obtained from VO for this analysis. The resulting network (IFNG-vaccine-VO) is a subgraph of the generic IFNG network. It contains the small network (IFNG-vaccine) and also genes and interactions associated with specific VO vaccine terms or their synonyms (e.g., tuberculosis vaccine BCG). The three layers of IFNG-associated gene interaction networks are summarized in Figure 5.7. The application of VO allows discovery of 38 more genes and 60 more interactions (IFNG-vaccine-VO). These new genes and interactions were not identified if only the term vaccine (or its variants) were used (IFNG-vaccine network). Our results indicate VO significantly increases the retrieval of the IFNG-vaccine network. Analyzing and comparing the vaccine-specific networks generated with or without VO support provides new insights and hypotheses for future investigations.

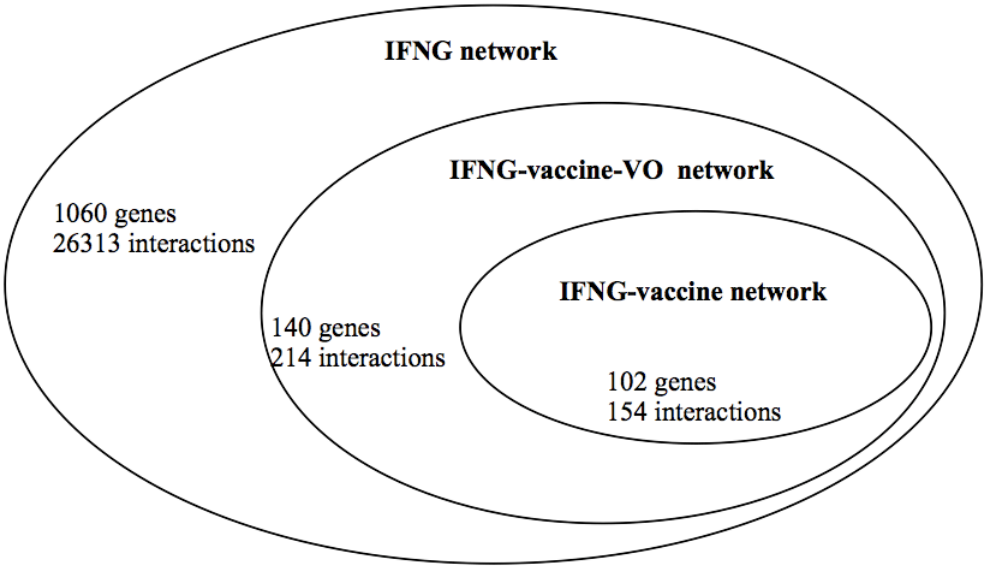


Figure 5.7: Three layers of IFNG-associated gene networks.

5.5.1 List of genes for vaccines or specific VO vaccine terms are predicted and sorted by centrality analyses

Figure 5.7 provides the general numbers of the different IFNG networks. To make more specific analysis, the most central genes (the genes ranked among the top 20 by at least one of the four centrality measures) are analyzed in more detail in Table 5.3. These genes (a total of 32 genes) are predicted to be associated with IFNG and relevant for the general vaccine or specific vaccine term(s). Literature evidence was manually curated for the vaccine development relatedness (Reference column in Table 5.3) of these genes. Based on Table 5.3, three different levels of prediction are available based on the comparison between the IFNG-vaccine network and the more specific IFNG-vaccine-VO network.

(i) Genes ranked high in both networks

23 genes were ranked high in both the IFNG-vaccine and IFNG-vaccine-VO networks. It suggests that the roles of certain genes (e.g., IL6) in vaccine research have

widely been recognized but studied in more depth in certain vaccines.

(ii) Genes ranked high in the IFNG-vaccine-VO network but not in the IFNG-vaccine network

Six genes (marked with "*") are included in this group, i.e., NFKB1, TLR2, NCAM1, CXCL10, CD86, and CCL2. These genes are found in the IFNG-vaccine network, but are not inferred as genes important for vaccine development, although there exists supporting literature evidence (Table 5.3). Using the VO enabled the identification of these vaccine-related genes.

(iii) Genes ranked high in the IFNG-vaccine-VO network but not found in the IFNG-vaccine network

This group includes three genes (marked with "**"), i.e., TLR4, TP53, and FCGR2B. These genes are not contained in the IFNG-vaccine network. Using the VO enabled the discovery of these genes as belonging to the IFNG-vaccine mediated gene interaction network and as genes important for vaccine research.

These gene lists provide new information to study vaccine-induced human gene networks associated with IFNG. For example, Toll-like receptor-4 (TLR4) is an important cell receptor that participates in many immune responses against pathogen infections. TLR4-active agents are often developed as vaccine adjuvants (Johnson, 2008). The finding of the presence of TLR4 in the IFNG-vaccine-VO network, but absence from the IFNG-vaccine network is a demonstration that our ontology-based method provides reasonable and useful information to better understand the vaccine-associated immune networks.

5.5.2 The predicted IFNG-BCG network

As an example of specific study on a single vaccine, Bacillus Calmette-Guérin (BCG) is a licensed tuberculosis vaccine to protect against infection of Mycobac-

Gene	Reference (PMID)	Gene	Reference (PMID)
IL2	8459207	CD40	11403919
TNF	16446013	CD28	12594842
IL10	10930151	C3	19477524
IL6	10225849	TLR4 **	12874299
IL4	8519092	TP53 **	10379742
CSF2	19459853	FCGR2B **	12874345
IL8	11378044	HSPD1	12218165
IL5	11138639	CD46	11757799
NFKB1 *	16971487	NCAM1 *	16316416
IL13	12232042	CXCL10 *	10799249
CD4	17298856	CD86 *	12594842
TLR2 *	12874299	IFNA1	19667099
IL7	17496983	CCL2 *	19833737
IL18	19467215	TPBG	16630022
EIF2AK2	19596385	GNLY	10644038
CD40LG	11403919	CD8A	18425263

Table 5.3: Predicted 32 genes related to IFN-γ and vaccine networks. These genes were ranked among the top 20 by at least one of the centrality measures in the literature-mined IFN-γ and vaccine network using VO (i.e. IFNG-vaccine-VO network). Genes marked with "*" were not ranked high in the IFNG-vaccine network built without using the VO (i.e. IFNG-vaccine network). Genes marked with "**" were not found in the IFNG-vaccine network. The PubMed PMIDs are listed to confirm the associations.

terium tuberculosis. In many cases, the term "BCG", instead of the term "vaccine" (or its variants), is used in sentences when talking about interaction with some other gene(s). Therefore, the sentence-based NLP text mining approaches won't retrieve those sentences with "BCG" when we only use the term "vaccine" for text retrieval. We used the "BCG" term and all its synonyms in VO to extract the network of interactions related to the BCG vaccine. The resulting network consists of 56 genes and 77 interactions (Figure 5.8). In total, 24 of these genes (colored with purple in Figure 5.8) were not found in the IFNG-vaccine network, which was constructed without using the "BCG" term in the VO.

The interactions between BCG treatment, TLR2 and TLR4 are interesting. BCG is able to activate TLR2 and TLR4 (PMID: 12874299). It induces the maturation of dendritic cells (DCs) via both TLR2 and TLR4 (PMID: 12630564), as well as the transcription and secretion of the chemokine CXCL8, by signalling through TLR2

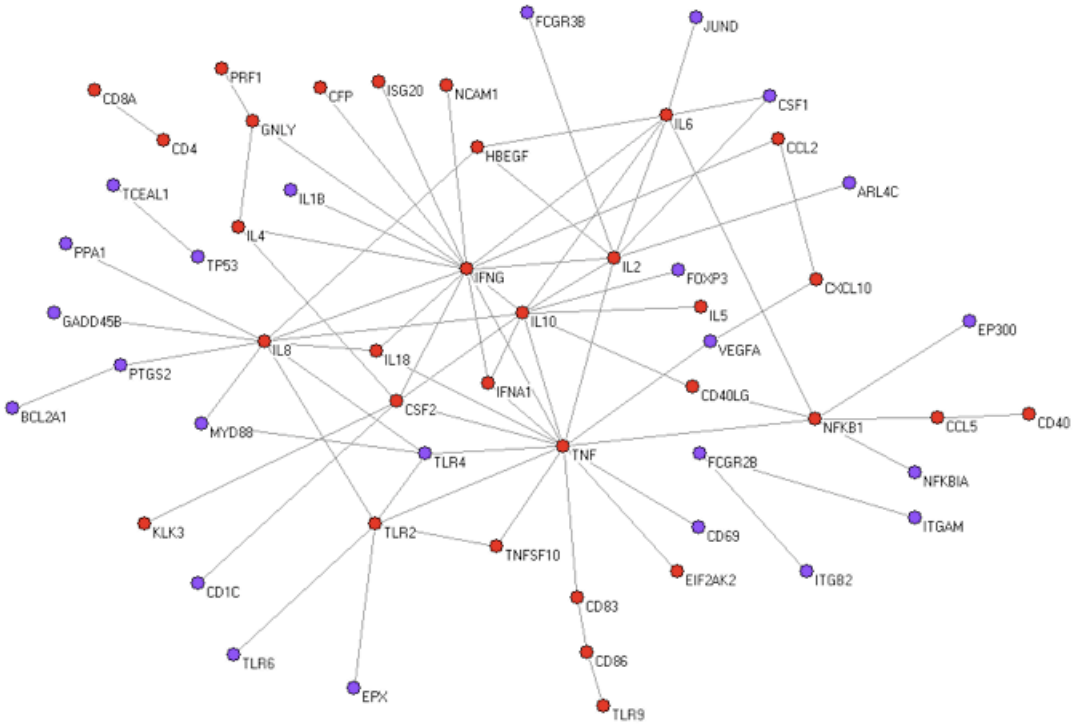


Figure 5.8: The IFNG-BCG network. All edges represent gene-gene interactions that are associated with the BCG vaccine. In total 24 new genes (colored with purple) are found by using the term BCG contained in the VO.

and TLR4 (PMID: 15760459). It can also induce TNF-alpha secretion from DC via TLR2 and TLR4 (PMID: 11083809).

As examples of more BCG-induced gene interactions, our system identified TNFSF10 (synonym: TRAIL) and TLR2 that are associated with BCG treatment (Fig. 2). It was reported that BCG can directly stimulate the release of tumor necrosis factor (TNF)-related apoptosis-inducing ligand (TRAIL, a synonym for TNFSF10) from polymorphonuclear leukocytes (PMN) through toll-like receptor-2 (TLR2) recognition that is augmented by IFNG (PMID: 18593617). BCG treatment on PMN triggers the induction of FCGR2B (synonym: CD32) (PMID: 12874345). BCG treatment also induces urinary IFNG, IP-10, TNF-alpha, and vascular endothelial growth factor (VEGF) (PMID: 10799249).

It is possible to generate new hypotheses by comparing the three layers of IFNG networks. For example, those new genes and interactions induced by BCG treatment may be possibly inferred to other vaccines (e.g., vaccines for intracellular pathogens such as Influenza vaccines or Brucella vaccines). Those genes and networks in the general IFNG or IFNG-vaccine network may provide new genes and interactions for inferring future BCG mechanism research.

5.6 Conclusion

In Chapter IV we proposed a centrality-based LBD approach to identify gene-disease associations and demonstrated that it is effective in discovering prostate cancer related genes, using 48,245 articles from PubMed Central (PMC) Open Access and 15 seed genes. In this chapter, we presented the general framework of the LBD method and showed that it can be generalized and used in different applications. We applied the LBD method to generate new hypotheses for IFNG and vaccine research

by using all abstracts in PubMed and only one gene (IFNG) as the seed gene.

Our analysis discovered a large number of genes that interact with IFNG and genes important for both IFNG and vaccine. Many of these genes have been studied but never been collected for systematic network analysis. Current databases contain limited information about IFNG gene interaction network. The Michigan Molecular Interactions (MiMI) database is a repository that includes interaction data from over 10 databases such as the Database of Interacting Proteins (DIP), the Human Protein Reference Database (HPRD), and the Biomolecular Interaction Network Database (BIND) (Tarcea et al., 2009). As of October 2009, MiMI contains only 12 genes that interact with IFNG and 27 interactions among these genes. Our IFNG gene interaction network contains more than 80 fold of genes that interact with IFNG. While the correctness of all these interactions require further confirmation, our manual confirmation of selected 56 interactions (Table 5.1) has already demonstrated the power of our literature-based discovery method. Since IFNG is an important immune regulator for vaccine-induced protective immunity, the systematical analysis of vaccine-induced IFNG-regulated gene network is critical to understand vaccine-induced immune mechanism and support rational vaccine design. Our selective analyses of the IFNG-vaccine subnetwork showed that genes potentially important for vaccine research can be predicted. Many predicted genes and gene networks deserve further experimental verifications.

We also investigated extending the centrality-based LBD approach by incorporating Vaccine Ontology (VO) support. Our study indicates that the application of VO significantly increases the discovery of IFNG and vaccine associated networks, leading to our finding of new genes and interactions that could not be found before.

CHAPTER VI

Conclusion

6.1 Summary of Contributions

Scientific publications are the main media through which researchers report their new findings. The huge amount and the continuing exponential growth of the number of published articles in biomedicine, has made it particularly difficult for researchers to access and utilize the knowledge contained in them. We had two main goals in this thesis: (i) develop methods to automatically extract biologically important information from published articles; (ii) use the information automatically extracted from the biomedical scientific literature to infer new knowledge (i.e., generate new scientific hypotheses). This chapter summarizes our main contributions and describes future directions for research. Chapters II and III target our first goal and contribute mainly to the areas of natural language processing (NLP) and information extraction (IE). Chapters IV and V address our second goal. While the main contributions of these chapters are in the area of literature-based discovery (LBD), the generated new hypotheses are contributions to the biomedical sciences.

In Chapter II we introduced a relation extraction method to identify protein-protein interactions in text. We proposed two kernel functions, i.e., path cosine kernel and path edit kernel, based on the paths between protein names in the de-

pendency parse trees of the sentences. Using these kernel functions we evaluated the performances of two classes of learning algorithms, Support Vector Machines (SVMs) and k-nearest-neighbor (kNN), and their semi-supervised counterparts, transductive SVMs (TSVM) and harmonic functions. We achieved significant improvement in protein-protein interaction extraction performance compared to results previously reported in the literature. To our knowledge, we presented the first effort of utilizing semi-supervised learning in this domain. We showed that semi-supervised algorithms perform better than their supervised versions by a wide margin when the amount of labeled data is limited. Harmonic functions achieve the best performance in such cases. When there is sufficient amount of labeled data, TSVM and SVM perform similarly to each other, and outperform kNN and harmonic functions. Unlike path cosine kernel, path edit kernel takes into account not only the common words on the dependency paths, but also the sequence of the words on the paths. Our results show that path edit kernel performs better than path cosine kernel in this domain. Chapters IV and V demonstrate the effectiveness of SVM with path edit kernel as a component of an LBD system for new hypothesis generation. We also used SVM with path edit kernel to contribute to the BioCreaive Meta-Server project by identifying abstracts that contain protein interaction information (Leitner et al., 2008)[1]. Another way that automatically extracted protein interactions can be used is to populate protein interaction databases. Our machine learning based approach achieved state-of-the-art F-measure performance for protein interaction extraction. However, in general, protein interaction databases favor high precision over high recall for higher user satisfaction. We developed a high-precision dependency tree rule-based interaction extraction system (GIN-IE). This system, which we describe in

[1] http://bcms.bioinfo.cnio.es/

the Appendix, has been integrated with the Michigan Molecular Interactions (MiMI) database[2] and made available to the end users (Tarcea et al., 2009).

Extracting protein-protein interactions from text is an active research area. Several new methods have been proposed (e.g. (Airola et al., 2008; Miwa et al., 2009; Wang, 2008)) after the work in Chapter II was published (Erkan, Özgür, & Radev, 2007a, 2007b). A recent study by Tikk et al. (2010) evaluates nine recent kernel methods for protein interaction extraction, including the path cosine kernel and the path edit kernel presented in Chapter II. The performances of our kernels are comparable to the current state-of-the-art dependency tree based kernel methods such as all-paths graph kernel (Airola et al., 2008) and k-band shortest path spectrum kernel (Tikk et al., 2010), and are better compared to syntactic parse tree based kernels such as subtree kernel (Vishwanathan & Smola, 2003), subset tree kernel (Collins & Duffy, 2001), and partial tree kernel (Moschitti, 2006).

Researchers often use speculative language in scientific articles when they are not certain about the statements that they make. It is important to distinguish factual information from speculative information. Previous studies on speculation detection approach the problem as a sentence classification task. In other words, sentences are classified as speculative or not. However, there are many sentences that contain both speculative and factual parts. In Chapter III, which was first published as (Özgür & Radev, 2009), we present one of the first efforts for identifying speculative fragments of sentences. A machine learning based method for detecting speculative sentence portions was independently proposed by (Morante & Daelemans, 2009). We approached the task in two steps, identifying speculation keywords and resolving their scopes. We used supervised classification to identify the speculation

[2]http://mimi.ncibi.org/

keywords. We introduced several linguistic features representing the contexts of the keywords and evaluated them using SVM with linear kernel. Our speculation keyword identification results (Abstracts: 91.69% F-measure, Full text: 82.58% F-measure) are close to the upper bound of human inter-annotator agreement scores for the BioScope corpus (Vincze et al., 2008). The best results were obtained by using all the features that we proposed: keyword-specific features (i.e., keyword, part of speech tag, stem), co-occurring keywords in the sentence, surrounding words with window size of one, positional features of the sentence in the article, and dependency tree relation features. To determine the scopes of the speculation keywords, we developed a rule-based system that exploits the syntactic structures of the sentences. This system achieved a significantly better performance (Abstracts: 79.89.% Accuracy, Full text: 61.13% Accuracy) compared to the baseline methods. Especially, the considerably lower performance of the baseline method that assigns the scope of a keyword to the whole sentence, emphasizes the importance of detecting speculative portions of the sentences. The fact that our scope resolution results are lower than the upper bound of human inter-annotator agreement, suggests that there is still room for improvement.

In Chapter IV we proposed a literature-based discovery (LBD) approach for identifying gene-disease associations. Most previous LBD systems are based on *Swanson's ABC* model and make use of the co-occurrence statistics among the entities. The novelty of our approach is that it integrates natural language processing (NLP) with network analysis to infer new relationships among entities. Given a concept of interest, we start with a set of one or more genes known to be related to the concept. We build a concept-specific gene interaction network by extracting the interactions of the concept-related genes from the biomedical literature using the path edit kernel

introduced in Chapter II with SVM. We analyze the constructed literature-mined network using network centrality methods. Our hypothesis is that genes central in the concept-specific network are also likely to be related to the concept. We showed that our approach is effective in identifying prostate cancer related genes. We were able to find genes which are not marked as being relayed to prostate cancer by the curated databases such as OMIM and PGDB, but there are recent articles in the literature that provide evidence for the relatedness of these genes to the disease. Our study also identified several genes which are currently not know to be related to prostate cancer, but are good candidates for further experimental studies, since they are found to be important in the prostate cancer specific network. Predicting good candidate genes is particularly important, given that wet lab experiments are costly and time consuming.

In Chapter V we showed that the LBD method proposed in Chapter IV to predict gene-disease associations can be generalized and applied to other problems. We adapted the approach to discover genes important for vaccine development starting with a single known vaccine-related gene, i.e., the interferon-gamma gene (IFNG). We analyzed all the article abstracts available in PubMed and reported the first high-throughput literature mining of human interferon-gamma and vaccine-mediated gene interaction networks. We created three different literature-mined networks. The first one is the generic IFNG network. The other two are its vaccine-specific sub-networks. The first vaccine-specific network is built by selecting the sentences that contain the concept term "vaccine". The second network extends the first one by including the terms in the Vaccine Ontology (VO) to the sentence selection process. Comparative analysis of these three layers of networks from graph centrality perspective led to the generation of several hypotheses. The evidences provided from the literature suggest

that many of the predicted genes are good candidates for further IFNG and vaccine development studies. Our results also showed that incorporating VO support to the LBD method enhanced the retrieval of IFNG and vaccine associated genes and provided new insights and hypotheses for future investigations.

In this thesis we focused on natural language processing and text mining in the biomedical domain. We designed and evaluated the proposed techniques for problems in biomedicine. However, it is important to note that, except for gene name identification, we did not use any tools that were specifically designed for the biomedical domain. Although verifying their success requires further investigations, the methods that we proposed here can be potentially applied to other domains. The relation extraction method that we introduced for protein interaction extraction in Chapter II, can be applied to extract pairwise relationships between other types of entities, e.g., employee-organization relationships. The characteristics of speculative language can differ across domains. However, given annotated training data the speculation detection method proposed in Chapter III can be trained for other domains. The scope resolution component is a rule-based system. However, the rules do not depend on the specific speculation keywords, but are based on the part of speech tags of the keywords. Therefore, in principle they can be generalized to other domains. In the LBD approach that we introduced in Chapters IV and V, we used genes as our entities and diseases or vaccines as our concepts of interest. The approach can be applied to problems in other domains with different types of entities and concepts. For example, we can start with a set of people who are known to be influential in a certain scientific field and build a co-authorship network around them. Identifying the most central people in this graph, can enable the detection of overlooked researchers who have implicitly influenced the development of the field.

6.2 Future Directions

Most information extraction systems including the methods that are proposed in this thesis operate on a sentence basis, neglecting the wider context information. Some types of information are not always found in the sentence, but need to be extracted non-locally from the entire document. For example, none of the sentences in the abstract in Figure 1.5 contain information regarding the species of the proteins and the experimental methods used to identify the interactions. However, the full text of the paper describes the experiments that were performed (e.g. yeast two hybrid, immunoprecipitation, and immunofluorescence microscopy) and the species studied (e.g. human) (Sato et al., 2005). Developing strategies to address this problem using global features from the entire article, considering not only the text but also the figures, tables, their legends, and the citation information is an interesting future direction for research.

In Chapters IV and V, even with a simple network design, where the nodes are the proteins/genes and the edges represent undirected and unweighted relationships among them, we were able to discover novel genes related to prostate cancer, and novel genes important for vaccine development. A possible avenue of research is enriching the network of interactions by including context information such as interaction type and causality (directionality), and developing new network analysis strategies to predict unknown relationships from such enriched networks. Another direction of research is to develop knowledge discovery methods based on integrating information extracted from the literature with data from various other available heterogeneous sources such as experimental results and manually curated databases. There are several different experimental techniques that can be used to detect an interac-

tion between a protein pair, as well as several different manually curated databases, and various journals and conference proceedings. Each experimental method is associated with different error rates and confidence levels. Similarly, different databases, journals, and conference proceedings are associated with different quality and reliability values. Weighting the edges in the network based on the confidence, quality, and reliability of the source from which the interaction was obtained, can lead to inferences of higher quality. In addition, analyzing such large-scale enriched networks can enable the identification of contradictory and anomalous knowledge.

APPENDICES

APPENDIX A

GIN-IE: A System for Extracting High Precision Gene Interactions using Dependency Tree Rules

A.1 Introduction

GIN-IE (Gene Interaction - Information Extraction) is a system that is developed with the goal of making the literature-mined bio-molecular interactions accessible and useful to the end users. While in general the state-of-the-art machine learning based approaches for interaction extraction achieve more balanced precision-recall performances, rule-based methods achieve higher precision in the expense of recall. High precision is an important requirement for most real-life applications. Therefore, a dependency tree rule-based approach for extracting protein/gene interactions with high precision is developed and integrated with the MiMI database[1] (Tarcea et al., 2009). The integration of GIN-IE with MiMI is a joint work with the National Center for Integrative Biomedical Informatics (NCIBI)[2].

Most previous approaches on protein interaction extraction focus on extracting that *"there is an interaction"* between a pair of proteins. Consider the sentence *"ZIPK phosphorylated STAT3 on serine 727 (Ser727) and enhanced STAT3 transcriptional activity."* from the abstract in Figure 1.5. Besides the fact that there is

[1] http://mimi.ncibi.org/
[2] http://ncibi.org/

an interaction relationship between *ZIPK* and *STAT3*, the facts that the relationship type is *phosphorylation*, and that the directionality is from *ZIPK* to *STAT3* (i.e. *ZIPK* phosphorylated *STAT3*, not the other way around), are also very important for biomedical scientists. GIN-IE extracts not only the interacting protein pairs, but also the types and directionalities of the interactions between them. GIN-IE has also rules to detect negations and speculations.

A.2 System Description

A.2.1 Data

GIN-IE is integrated with the daily processing and updates pipeline of the BioNLP database in NCIBI. This database stores parsed and tagged text from NLM's PubMed literature database. GIN-IE obtains the sentences, tagged gene names, dependency parse trees, and word stems from NCIBI's BioNLP database, processes these data for protein interactions, and stores the results back in the same database.

The GIN-IE pipeline has also been adjusted to process the full text articles in NCIBI's Pubmed Central (PMCOA) database.

A.2.2 Dependency Tree Rules for Protein Interaction Extraction

The dependency parse trees in NCIBI's BioNLP database were obtained using the Stanford Parser[3] (de Marneffe et al., 2006). We examined the dependency trees of various sentences to define high precision protein interaction extraction rules. The rules that we defined are based on, first identifying the interaction keywords in the sentences and then, inferring the protein pair that are related with that interaction keyword. The interaction keywords are identified by matching all the words in the sentence against a list of predefined interaction keywords. Matching is done using the stemmed words. The matched keywords are further mapped to interaction types

[3]http://nlp.stanford.edu/software/lex-parser.shtml

using an interaction ontology developed collaboratively in NCIBI. For example, if the matched keyword is "inhibit", the mapped interaction type is "negative regulation". Some of the interaction types are directional, while others are symmetric. For example, "binding" is symmetric. If "A binds to B", "B binds to A" as well. On the other hand, "phosphorylation" is directional. "Phosphorylation of A by B", does not imply that "A phosphorylates B". This section describes the different rules that we extracted. GIN-IE's implementation allows new rules to be added easily.

Rule 1:

Rule 1 is applied as follows. There is an interaction between two proteins, if they are the children of an interaction keyword, and the dependency type between the interaction keyword and one of the children is "nsubj", while the dependency type between the interaction keyword and the other child is "dobj". An example sentence and the portion of its dependency parse tree that triggers this rule is shown below.

Sentence:
"Recombinant Sin3A bound Ebp1 directly, but recombinant HDAC2 failed to bind Ebp1."

Portion of the dependency parse tree:
nsubj(bound-3, Sin3A-2)
dobj(bound-3, Ebp1-4)

Extracted Interaction:
Agent: Sin3A
Target: Ebp1
Interaction word: bound

Rule 2:

Rule 2 captures interactions expressed in passive voice. It is applied as follows. There is an interaction between two proteins, if they are the children of an interaction keyword, and the dependency type between the interaction keyword and one of the children is "nsubjpass", while the dependency type between the interaction keyword

and the other child is “agent”. An example sentence, the portion of its dependency parse tree that triggers this rule, and the extracted interaction are shown below.

Sentence:
“Such a notion is supported by the findings that phosphorylation of pp32 by p56lck correlated with expression of the CD45 molecules and that in vitro phosphorylated pp32 was completely dephosphorylated by purified CD45.”

Portion of the dependency parse tree:
nsubjpass(dephosphorylated-30, pp32-27)
agent(dephosphorylated-30, CD45-33)

Extracted Interaction:
Agent: CD45
Target: Ebp1
Interaction word: dephosphorylated

Rule 3:

Rule 3 is applied as follows. There is an interaction between two proteins, if they are the children of an interaction keyword, and the dependency type between the interaction keyword and one of the children is “nsubj”, while the dependency type between the interaction keyword and the other child is “prep_with”. An example sentence, the portion of its dependency parse tree that triggers this rule, and the extracted interaction are shown below.

Sentence:
“Taken together, these results indicate that the Ras-interacting region on AF-6 is structurally similar to that on Raf-1 and on RalGDS and that AF-6 interacts with activated Ras and ZO-1 in vivo. ”

Portion of the dependency parse tree:
nsubj(interacts-26, AF-6-25)
prep_with(interacts-26, Ras-29)

Extracted Interaction:
Agent: AF-6
Target: Ras
Interaction word: interacts

Rule 4:

Rule 4 is applied as follows. There is an interaction between two proteins, if they are the children of an interaction keyword, and the dependency type between the interaction keyword and one of the children is "prep_of", while the dependency type between the interaction keyword and the other child is "prep_by". An example sentence, the portion of its dependency parse tree that triggers this rule, and the extracted interaction are shown below.

Sentence:
"The activation of PKBbeta and PKBgamma by PDK1 was accompanied by the phosphorylation of the residues equivalent to Thr308 in PKBalpha, namely Thr309 (PKBbeta) and Thr305 (PKBgamma). "

Portion of the dependency parse tree:
prep_of(activation-2, PKBbeta-4)
prep_by(activation-2, PDK1-8)

Extracted Interaction:
Agent: PDK1
Target: PKBbeta
Interaction word: activation

Rule 5:

Rule 5 is applied as follows. There is an interaction between two proteins, if one of them is a child of an interaction keyword, and the dependency type between the interaction keyword and that child is "prep_between", while the other protein is a child of the first protein and the dependency type between them is "conj_and". An example sentence, the portion of its dependency parse tree that triggers this rule, and the extracted interaction are shown below.

Sentence:
"The interaction between Shank2 and NHE3 was further confirmed
by immunoprecipitation and surface plasmon resonance studies."

Portion of the dependency parse tree:
prep_between(interaction-2, Shank2-4)
conj_and(Shank2-4, NHE3-6)

Extracted Interaction:
Agent: Shank2
Target: NHE3-6
Interaction word: interaction

Rule 6:

Rule 6 is applied as follows. There is an interaction between two proteins, if they are the children of an interaction keyword, and the dependency type between the interaction keyword and one of the children is "prep_of", while the dependency type between the interaction keyword and the other child is "prep_with". An example sentence, the portion of its dependency parse tree that triggers this rule, and the extracted interaction are shown below.

Sentence:
"Interactions of RPB2, ERH, NDR1 and PRMT5 with FCP1
were confirmed by co-immunoprecipitation or in vitro pull-down assays."

Portion of the dependency parse tree:
prep_of(interactions-1, RPB2-3)
prep_with(interactions-1, FCP1-9)

Extracted Interaction:
Agent: FCP1
Target: RPB2
Interaction word: interactions

Rule 7:

Rule 7 is applied as follows. There is an interaction between two proteins, if they are the children of an interaction keyword, and the dependency type between the interaction keyword and one of the children is "nsubj", while the dependency type between the interaction keyword and the other child is "prep_to". An example

sentence, the portion of its dependency parse tree that triggers this rule, and the extracted interaction are shown below.

Sentence:
"These results suggest a model in which p53 binds to TBP and interferes with transcriptional initiation."

Portion of the dependency parse tree:
nsubj(binds-9, p53-8)
prep_to(binds-9, TBP-11)

Extracted Interaction:
Agent: p53
Target: TBP
Interaction word: binds

Rule 8:

Rule 8 is applied as follows. There is an interaction between two proteins, if one of them is a child of an interaction keyword, and the dependency type between the interaction keyword and that child is "prep_of", while the other protein is a child of the first protein and the dependency type between them is "prep_with". An example sentence, the portion of its dependency parse tree that triggers this rule, and the extracted interaction are shown below.

Sentence:
"The interaction of Ngb with flotillin-1 was confirmed by glutathione S-transferase pull-down experiments."

Portion of the dependency parse tree:
prep_of(interaction-2, Ngb-4)
prep_with(Ngb-4, flotillin-1-6)

Extracted Interaction:
Agent: flotillin-1
Target: Ngb
Interaction word: interaction

A.2.3 Dependency Tree Simplification

The rules described in the previous section, were implemented by extracting and analyzing the dependency tree paths from the interaction keywords to the protein names in the sentences. If the interaction keyword and/or the protein names consist of multiple tokens. The shortest path between the tokens is used.

Consider the sentence "These results demonstrate that Duplin inhibits not only Tcf-4 but also STAT3.", whose dependency tree is shown in Figure A.1.

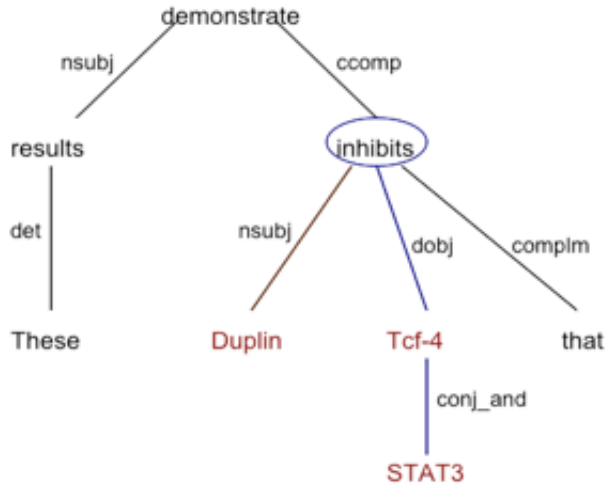


Figure A.1: The dependency tree of the sentence "*These results demonstrate that Duplin inhibits not only Tcf-4 but also STAT3.*" The proteins are shown in red and the interaction keyword is circled.

The sentence describes two interactions:

- Interaction 1:
  - Type: Negative regulation
  - Interaction keyword: inhibits
  - Agent: Duplin
  - Target: Tcf-4

- Interaction 2:
  - Type: Negative regulation
  - Interaction keyword: inhibits
  - Agent: Duplin
  - Target: STAT3

The first interaction is captured by *Rule 1*. The dependency tree path from the interaction keyword "inhibits" to "Duplin" is "nsubj" (noun subject), which encodes that "Duplin" is the agent of the interaction signaled by the keyword "inhibits". Similarly, the dependency tree path, "dobj" (direct object), from "inhibits" to "Tcf-4", encodes the information that "Tcf-4" is the target of the interaction. However, *Rule 1* fails to detect the second interaction. The dependency tree path from "inhibits" to the target of the second interaction "STAT3" is "dobj Tcf-4 conj_and". The main information that "STAT3" is a target of the interaction signaled by "inhinits" is encoded by the dependency type "dobj" on the path. "Tcf-4" and "conj_and" on the path don't modify the role of "STAT3". The words such as Tcf-4 on the path might result in over-fitting and poor generalization. Another observation is that dependency relations on the path such as conjunctions, abbreviations, determiners, numbers, and appositives generally don't modify the meaning for the relation, so can be eliminated for better generalization. For example, removing "Tcf-4" and "conj_and" from the path from "inhibits" to "STAT3", wouldn't change the semantics of the information encoded by the path for the second interaction. In addition, better generalization is achieved since, the targets of both interactions are now represented with the same dependency path "dobj". We implemented dependency tree simplification in GIN-IE, which resulted in higher recall with no loss in precision.

A.2.4 Negation and Speculation Detection

GIN-IE contains rules to detect negations and speculations. The dependency parse trees of the sentences are used to detect negations with "not" and "no". An extracted interaction is negated if one of the following dependency tree patterns are matched.

- The interaction keyword has a child connected to it with a dependency type "neg" (negation).
- The interaction keyword has a child "no" connected to it with a dependency type "det" (determiner).

Besides the dependency tree rules for negation, GIN-IE contains additional sentence pattern matching rules to detect negations with "fail to", "neither nor", and "lack of".

Speculative sentences are identified matching a set of manually derived speculation keywords including suggest, likely, may, putative, hypothesis, probable, speculate, investigate, examine, explore, and might. As future work, we will integrate the method proposed in Chapter III for detecting speculative sentence fragments to GIN-IE.

A.2.5 Evaluation

The MiMI database contains interactions integrated from several manually curated protein interaction databases. Some of these interactions are associated with the PubMed ID's of the articles from where they were curated. To compile a protein interaction data set, we used the abstracts of the articles for which there is an interaction reported in MiMI. 200 sentences that contain protein pairs that are reported as interacting in MiMI were randomly selected and manually annotated for inter-

actions. The annotation includes the interacting protein pair, interaction keyword, and directionality. 100 sentences were used for training and the remaining 100 were used to test. The training and test sets contain 114 and 90 interactions, respectively.

The performance results of GIN-IE over the test set are shown in Table A.1. Dependency tree simplification improves not only recall, but also precision.

Dependency Tree Simplification	Precision	Recall	F-Measure
No	0.92	0.14	0.25
Yes	0.95	0.22	0.36

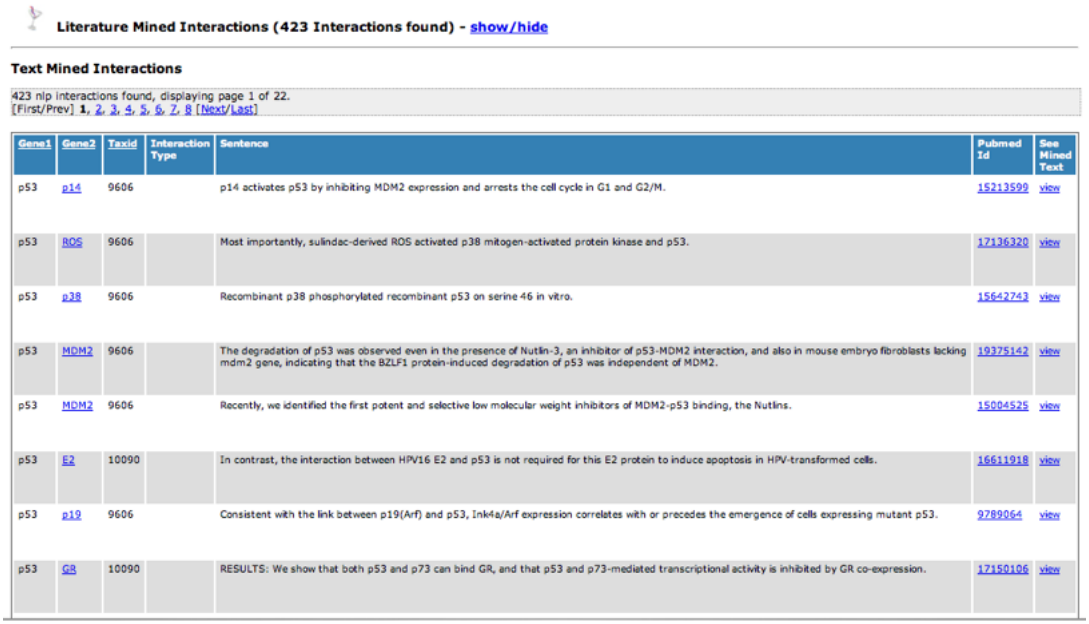
Table A.1: GIN-IE results over the test set.

A.3 Availability

The interactions extracted by GIN-IE are accessible through the MiMI Web interface (http://mimi.ncibi.org/). Currently the are over $30,000$ interactions in MiMI that were extracted by GIN-IE. Figure A.2 shows the screen shot from MiMI Web displaying the GIN-IE extracted interactions of the TP53 protein. The interacting protein pair, the sentence from where the interaction was extracted, together with a link to the corresponding abstract in PubMed are shown. Including the interaction type to MiMI is ongoing work.

GIN-IE is run on a daily basis, together with the NCIBI BioNLP database to extract the most recent interactions in PubMed. The new interactions are published as an RSS feed (http://gin.ncibi.org/rss/gin-ie/interactions.rss).

The GIN-IE source code is included to Clairlib (http://www.clairlib.org/), which is an open source library of Perl modules to simplify generic tasks in natural language processing, information retrieval, and network analysis.



Literature Mined Interactions (423 Interactions found) - show/hide

Text Mined Interactions

423 nlp interactions found, displaying page 1 of 22.
[First/Prev] **1**, 2, 3, 4, 5, 6, 7, 8 [Next/Last]

Gene1	Gene2	Taxid	Interaction Type	Sentence	Pubmed Id	See Mined Text
p53	p14	9606		p14 activates p53 by inhibiting MDM2 expression and arrests the cell cycle in G1 and G2/M.	15213599	view
p53	ROS	9606		Most importantly, sulindac-derived ROS activated p38 mitogen-activated protein kinase and p53.	17136320	view
p53	p38	9606		Recombinant p38 phosphorylated recombinant p53 on serine 46 in vitro.	15642743	view
p53	MDM2	9606		The degradation of p53 was observed even in the presence of Nutlin-3, an inhibitor of p53-MDM2 interaction, and also in mouse embryo fibroblasts lacking mdm2 gene, indicating that the BZLF1 protein-induced degradation of p53 was independent of MDM2.	19375142	view
p53	MDM2	9606		Recently, we identified the first potent and selective low molecular weight inhibitors of MDM2-p53 binding, the Nutlins.	15004525	view
p53	E2	10090		In contrast, the interaction between HPV16 E2 and p53 is not required for this E2 protein to induce apoptosis in HPV-transformed cells.	16611918	view
p53	p19	9606		Consistent with the link between p19(Arf) and p53, Ink4a/Arf expression correlates with or precedes the emergence of cells expressing mutant p53.	9789064	view
p53	GR	10090		RESULTS: We show that both p53 and p73 can bind GR, and that p53 and p73-mediated transcriptional activity is inhibited by GR co-expression.	17150106	view

Figure A.2: Screen shot from MiMI Web showing the interactions of the TP53 protein extracted by GIN-IE.

APPENDIX B

GIN-NA: A system for Gene Network Analysis

B.1 System Description

GIN-NA (Gene Interaction - Network Analysis) is a system for analysing molecule interaction networks. The interaction networks are retrieved from the MiMI database, which integrates protein interactions from diverse biological data sources. Analysis of two types of networks are performed, namely molecule-specific networks and disease-specific networks. Molecule-specific networks are the networks of interactions in the neighborhood of a molecule or a list of molecules. Besides the general network statistics such as average degree, power-law degree distribution, clustering coefficient, and shortest path statistics, GIN-NA ranks the molecules in the network based on graph centrality measures and second neighbor statistics. Network statistics and network centrality scores are computed using Clairlib (http://www.clairlib.org/). Disease-specific networks are built by compiling lists of known disease genes and retrieving the interactions among these genes and their neighborhood. We hypothesize that the genes central in the disease-specific gene interaction network are likely to be related to the disease and rank the genes based on their centrality scores. Currently, GIN-NA provides disease-specific networks for the Prostate Cancer, Type 1 Diabetes, Type 2 Diabetes, and Bipolar Disorder.

B.2 Availability

The GIN-NA web system is available at http://gin.ncibi.org/. The query page of GIN-NA is shown in Figure B.1. The user can query for a molecule or a list of molecules, and restrict his search by organism, molecule type, and data source.

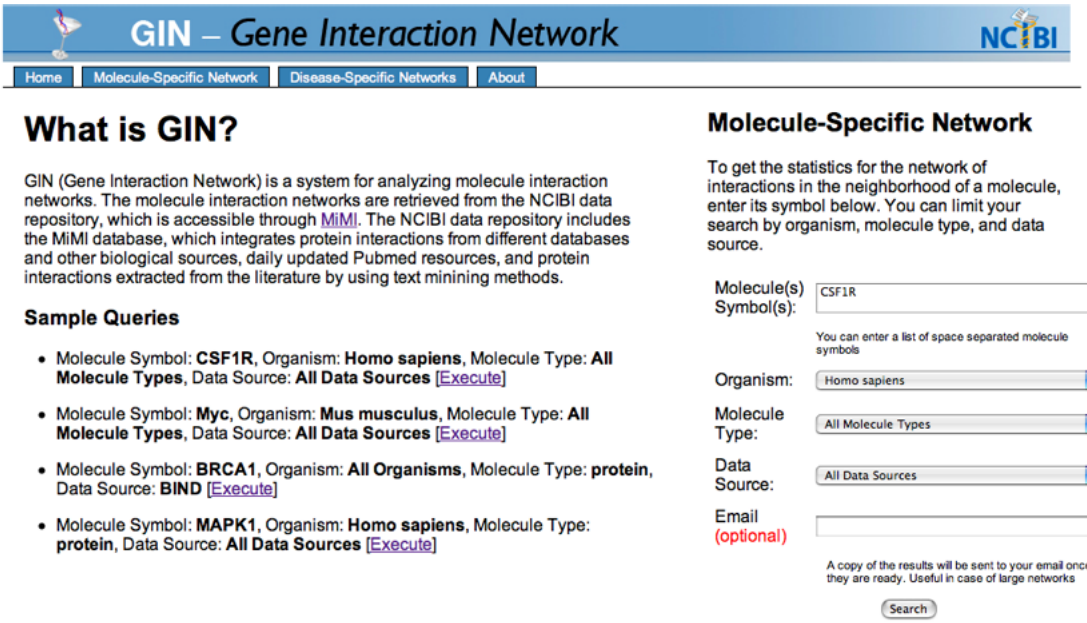


Figure B.1: Molecule query screen of GIN-NA.

The computed network analysis results are displayed to the user on the GIN-NA web-site and/or emailed to him if he provides an email address. The screen shot showing the molecule specific network analysis results for the "CSF1R" molecule is presented in Figure B.2. The screen shot showing the disease-specific network analysis results for prostate cancer is presented in Figure B.3.

GIN-NA is also accessible through web-services, where the user can provide a network of interactions in edge-list format and get the network analysis results: http://clair.si.umich.edu/clair/webservice/gin-na/netserver.cgi.

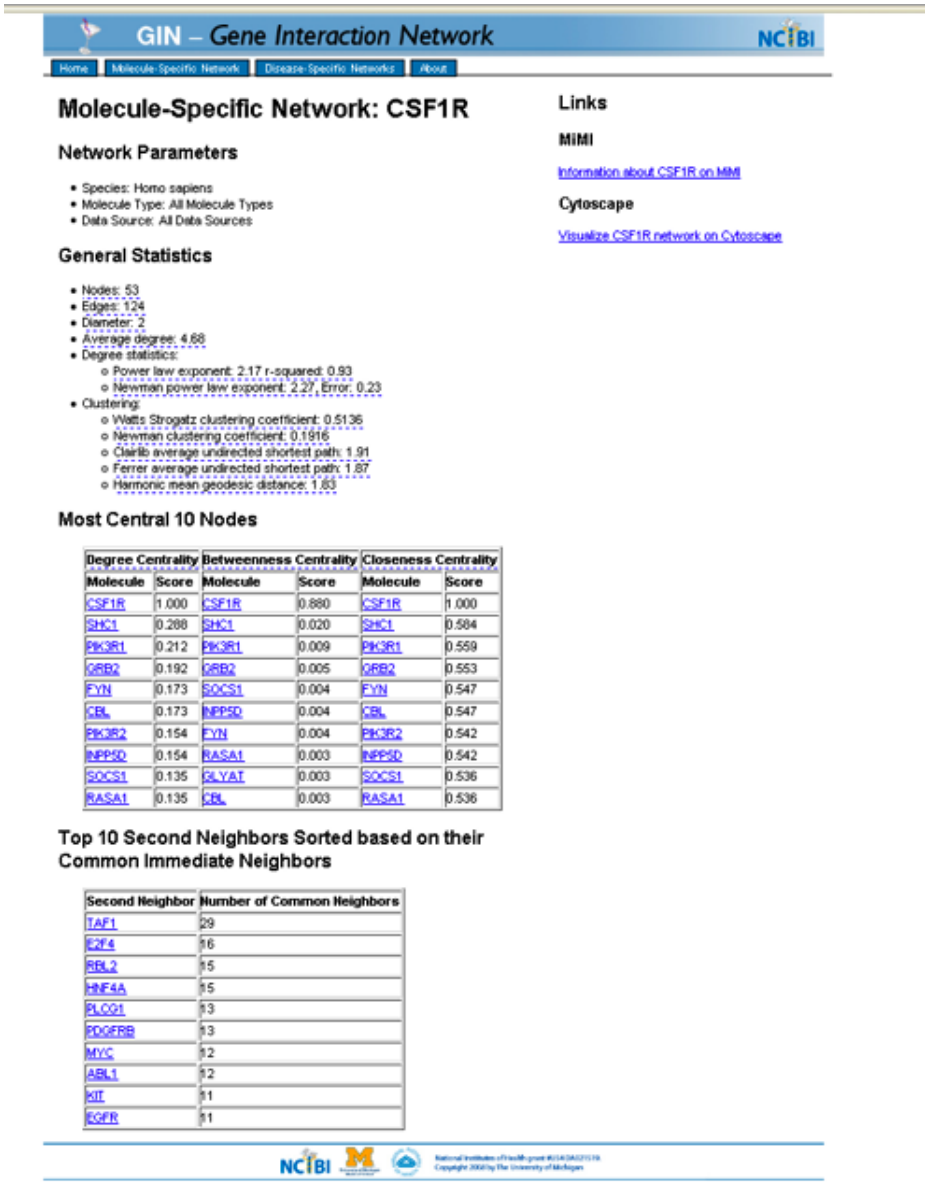


Figure B.2: Molecule-specific network analysis for CSF1R using GIN-NA.

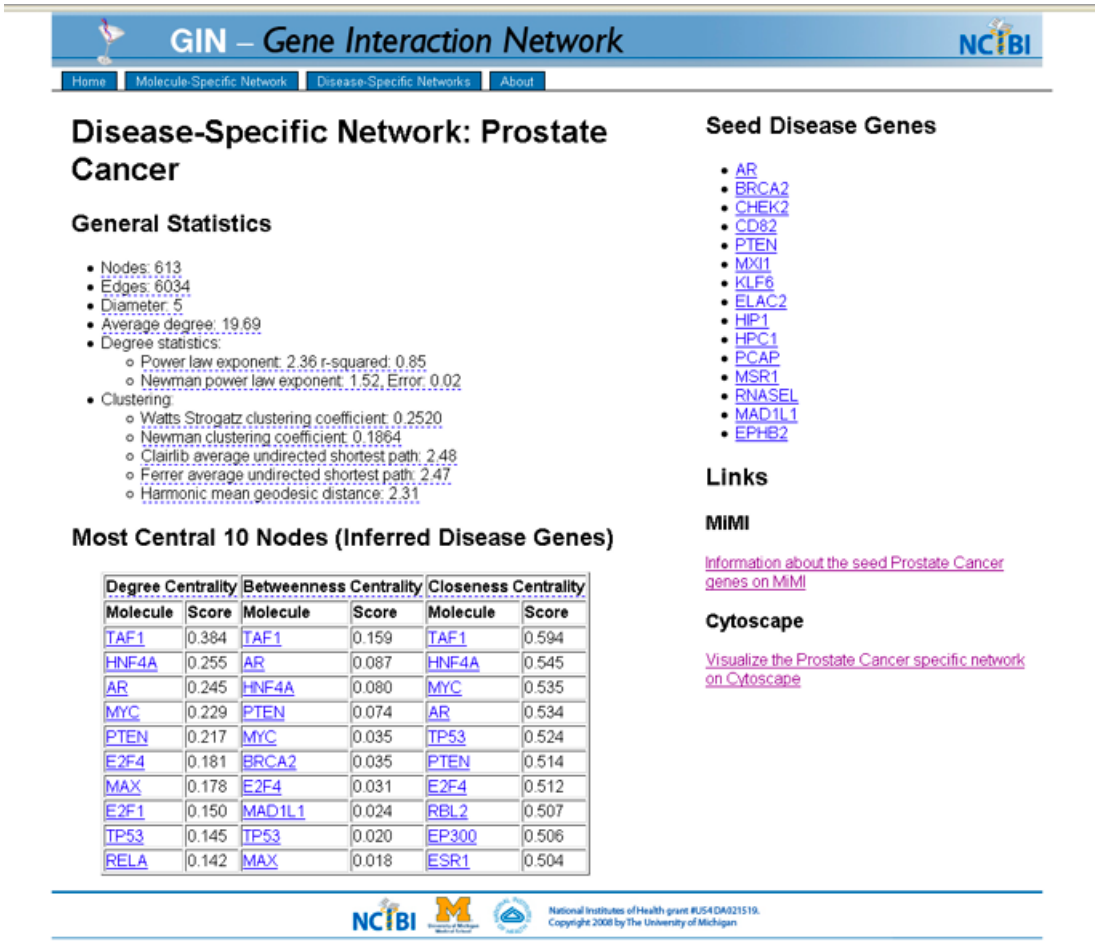


Figure B.3: Disease-specific network analysis for prostate cancer using GIN-NA.

BIBLIOGRAPHY

BIBLIOGRAPHY

Lada A. Adamic, Dennis Wilkinson, Bernardo A. Huberman, and Eytan Adar (2002). A literature based method for identifying gene-disease connections. In *IEEE Computer Society Bioinformatics Conference*.

Alex Ade, Zach Wright, and David J. States (2007). Gene2MeSH [Internet]. Ann Arbor (MI): National Center for Integrative Biomedical Informatics, Available from http://gene2mesh.ncibi.org.

Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski (2008). All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics*, *9 Suppl 11*.

Hisham Al-Mubaid and Rajit K. Singh (2005). A new text mining approach for finding protein-to-disease associations. *American Journal of Biochemistry and Biotechnology*, *1*(3), 145–152.

Réka Albert and Albert L. Barabási (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47–97.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29.

Yaron Avitzur, Esther Galindo-Mata, and Nicola L. Jones (2005). Oral vaccination against helicobacter pylori infection is not effective in mice with fas ligand deficiency. *Digestive Diseases and Sciences*, *50*(12), 2300–2306.

Gary D. Bader, Doron Betel, and Christopher W. V. Hogue (2003). BIND - The Biomolecular Interaction Network Database. *Nucleic Acids Research*, *31*(1), 248–250.

Amos Bairoch and Rolf Apweiler (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, *28*(1), 45–48.

Amos Bairoch, Rolf Apweiler, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su L. Yeh (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Research*, *33*(1), D154–D159.

Nancy C. Baker and Bradley M. Hemminger (2010). Mining connections between chemicals, proteins, and diseases extracted from Medline annotations. *J. of Biomedical Informatics*, *43*(4), 510–519.

Albert-László Barabási, Hawoong Jeong, Zoltan Neda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, *311*(3–4), 590–614.

Chitta Baral, Hasan Davulcu, Graciela Gonzalez, Geeta Joshi-Tope, Mutsumi Nakamura, Prabhdeep Singh, Luis Tari, and Lian Yu (2005). CBioC: Web-based collaborative curation of molecular interaction data from biomedical literature. In *The Genetics Society of America 1st International Biocurator Meeting*.

Kevin G. Becker, Kathleen C. Barnes, Tiffani J. Bright, and S. Alex Wang (2004). The genetic association database. *Nature Genetics*, *36*(5), 431–432.

Alfons Billiau and Patrick Matthys (2009). Interferon-gamma: a historical perspective. *Cytokine and Growth Factor Reviews*, *20*(2), 97–113.

Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski (2009). Extracting complex biological events with rich graph-based feature sets. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pp. 10–18 Morristown, NJ, USA. Association for Computational Linguistics.

Christian Blaschke, Miguel A. Andrade, Christos Ouzounis, and Alfonso Valencia (1999). Automatic extraction of biological information from scientific text: Protein-protein interactions. In *Proceedings of the AAAI Conference on Intelligent Systems for Molecular Biology (ISMB 1999)*, pp. 60–67.

Christian Blaschke and Alfonso Valencia (2002). The frame-based module of the Suiseki information extraction system. *IEEE Intelligent Systems*, pp. 14–20.

Ursula Bommhardt, Katherine C. Chang, Paul E. Swanson, Tracey H. Wagner, Kevin W. Tinsley, Irene E. Karl, and Richard S. Hotchkiss (2004). Akt decreases lymphocyte apoptosis and improves survival in sepsis. *Journal of Immunology*, *172*(12), 7583–7591.

Kevin R. Brown and Igor Jurisica (2005). Online Predicted Human Interaction Database OPHID. *Bioinformatics*, *21*(9), 2076–2082.

Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, and Yuk Wah Wong (2005). Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, *33*(2), 139–155.

Razvan Bunescu and Raymond J. Mooney (2005a). A shortest path dependency kernel for relation extraction. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 724–731 Vancouver, B.C.

Razvan Bunescu and Raymond J. Mooney (2005b). Subsequence kernels for relation extraction. In *Proceedings of the 19th Conference on Neural Information Processing Systems (NIPS)* Vancouver, B.C.

Razvan Bunescu and Raymond J. Mooney (2007). Extracting relations from text: From word sequences to dependency paths. In Anne Kao and Stephen R. Poteet, editors, *Natural Language Processing and Text Mining*, chap. 3, pp. 29–44. Springer London.

Christopher J.C. Burges (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167.

Hao Chen and Burt M. Sharp (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*, *5*(1).

Jake Yue Chen, Changyu Shen, and Andrey Y. Sivachenko (2006). Mining Alzheimer disease relevant proteins from integrated protein interactome data. In *Pac Symp Biocomput*, pp. 367–378.

Michele Clamp, Ben Fry, Mike Kamal, Xiaohui Xie, James Cuff, Michael F. Lin, Manolis Kellis, Kerstin Lindblad-Toh, and Eric S. Lander (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences*, *104*(49), 19428–19433.

Aaron M. Cohen and William R. Hersh (2005). A survey of current work in biomedical text mining. *Brief Bioinform*, *6*(1), 57–71.

K. Bretonnel Cohen and Lawrence Hunter (2004). *Artificial Intelligence Methods and Tools for Systems Biology*, chap. Natural language processing and systems biology, pp. 147–174. Springer.

Nigel Collier, Hyun S. Park, Norihiro Ogata, Yuka Tateishi, Chikashi Nobata, Tomoko Ohta, Tateshi Sekimizu, Hisao Imai, Katsutoshi Ibushi, and Jun I. Tsujii (1999). The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pp. 271–272. Association for Computational Linguistics.

Michael Collins and Nigel Duffy (2001). Convolution kernels for natural language. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *NIPS*, pp. 625–632. MIT Press.

Richard R. Copley (2008). The animal in the genome: comparative genomics and evolution. *Philos Trans R Soc Lond B Biol Sci.*, *363*(1496), 1453–1461.

Corinna Cortes, Patrick Haffner, and Mehryar Mohri (2004). Rational kernels: Theory and algorithms. *Journal of Machine Learning Research*, pp. 1035–1062.

Kenneth A. Cory (1997). Discovering hidden analogies in an online humanities database. *Computers and the Humanities*, *31*, 1–12.

Mark Craven and Johan Kumlien (1999). Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pp. 77–86. AAAI Press.

Aron Culotta and Jeffrey Sorensen (2004). Dependency tree kernels for relation extraction. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 423 Morristown, NJ, USA. Association for Computational Linguistics.

Sarah Danson, Emma J. Dean, Caroline Dive, and Malcolm R. Ranson (2007). IAPs as a Target for Anticancer Therapy. *Current Cancer Drug Targets*, *7*(8), 785–794.

Nikolai Daraselia, Anton Yuryev, Sergei Egorov, Svetalana Novichkova, Er Nikitin, and Ilya Mazo (2004). Extracting Human Protein Interactions from MEDLINE Using a Full-Sentence Parser. *Bioinformatics*, *20*(5), 604–611.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of LREC-06*.

Charles A. Dinarello (1999). Interleukin-18. *Methods*, *19*(1), 121–132.

Dmitriy Dligach and Martha Palmer (2008). Novel Semantic Features for Verb Sense Disambiguation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Rainer Doffinger, Matthew R. Helbert, Gabriela Barcenas-Morales, Kun Yang, Stephanie Dupuis, Lourdes Ceron-Gutierrez, Clara Espitia-Pinzon, Neil Barnes, Graham Bothamley, Jean-Laurent Casanova, Hilary J. Longhurst, and Dinakantha S. Kumararatne (2004). Autoantibodies to interferon-gamma in a patient with selective susceptibility to mycobacterial infection and organ-specific autoimmunity. *Clinical Infectious Diseases*, *38*(1), e10–e14.

Ian Donaldson, Joel Martin, Berry de Bruijn, Cheryl Wolting, Vicki Lay, Brigitte Tuekam, Shudong Zhang, Berivan Baskin, Gary D. Bader, Katerina Michalickova, Tony Pawson, and Christopher W. V. Hogue (2003). PreBIND and Textomy - Mining the Biomedical Literature for Protein-Protein Interactions Using a Support Vector Machine. *BMC Bioinformatics*, *4*, 11.

Peter G. Doyle and J. Laurie Snell (1984). *Random Walks and Electric Networks*. Mathematical Association of America.

Güneş Erkan (2007). *Using Graphs And Random Walks For Discovering Latent Semantic Relationships In Text*. Ph.D. thesis, The University of Michigan, Ann Arbor, Michigan.

Güneş Erkan, Arzucan Özgür, and Dragomir R. Radev (2007a). Extracting interacting protein pairs and evidence sentences by using dependency parsing and machine learning techniques. In *Proceedings of the Second BioCreAtIvE Challenge Workshop - Critical Assessment of Information Extraction in Molecular Biology*.

Güneş Erkan, Arzucan Özgür, and Dragomir R. Radev (2007b). Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 228–237.

Güneş Erkan and Dragomir R. Radev (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*, *22*, 457–479.

Anthony Fader, Dragomir R. Radev, Michael H. Crespin, Burt L. Monroe, Kevin M. Quinn, and Michael Colaresi (2007). MavenRank: Identifying influential members of the US senate using lexical centrality. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 658–666.

Christiane Fellbaum (Ed.). (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Michel D. Ferrari (1992). Biochemistry of migraine. *Pathol. Biol. (Paris)*, *40*(4), 287–292.

Ronald A. Fisher (1970). *Statistical Methods for Research Workers* (Fourteenth edition). Collier-Macmillian.

Helen A. Fletcher (2007). Correlates of immune protection from tuberculosis. *Current Molecular Medicine*, *7*(3), 319–325.

Linton C. Freeman (1977). A set of measures of centrality based on betweenness. *Sociometry*, *40*(1), 35–41.

Linton C. Freeman (1979). Centrality in social networks: Conceptual clarification. *Social Networks*, *1*(3), 215–239.

Jan Freudenberg and P. Propping (2002). A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, *18*(Suppl. 2), S110–S115.

Kenichiro Fukuda, A. Tamura, Tatsuhiko Tsunoda, and Toshihisa Takagi (1998). Toward information extraction: identifying protein names from biological papers. In *Pac Symp Biocomput*, pp. 707–718.

Simone Fulda (2008). Targeting inhibitor of apoptosis proteins (IAPs) for cancer therapy. *Anti-cancer agents in medicinal chemistry*, *8*(5), 533–539.

Sherrilynne S. Fuller, Debra Revere, Paul F. Bugni, and George M. Martin (2004). A knowledge-base system to enhance scientific discovery: Telemakus. *Biomed Digit Libr*, *1*(1), 2.

Katrin Fundel, Robert Kuffner, and Ralf Zimmer (2007). Relex–relation extraction using dependency parse trees. *Bioinformatics*, *23*(3), 365–371.

Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Laszlo Barabasi (2007). The human disease network. *PNAS*, *104*(21), 8685–8690.

Graciela Gonzalez, Juan C. Uribe, Luis Tari, Colleen Brophy, and Chitta Baral (2007). Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures. In *Pac Symp Biocomput*, pp. 28–39.

Michael Gordon, Robert K. Lindsay, and Weiguo Fan (2002). Literature-based discovery on the World Wide Web. *ACM Trans. Internet Technol.*, *2*(4), 261–275.

Michael D. Gordon and Susan Dumais (1998). Using latent semantic indexing for literature based discovery. *J. Am. Soc. Inf. Sci.*, *49*(8), 674–685.

Daniel J. Gough, David E. Levy, Ricky W. Johnstone, and Christopher J. Clarke (2008). Ifnγ signaling-does it mean jak-stat. *Cytokine and Growth Factor Reviews*, *19*(5-6), 383–394.

Zhou GuoDong and Su Jian (2004). Exploring deep knowledge resources in biomedical name recognition. In *JNLPBA'04: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 96–99 Morristown, NJ, USA. Association for Computational Linguistics.

Matthew W. Hahn and Andrew D. Kern (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, *22*(4).

Feng Hao, Mingqi Tan, Xuemin Xu, Jiahuai Han, Duane D. Miller, Gabor Tigyi, and Mei-Zhen Cui (2007). Lysophosphatidic acid induces prostate cancer PC3 cell migration via activation of LPA(1), p42 and p38alpha. *Biochim Biophys Acta.*, *1771*(7), 883–892.

Yongqun He, Lindsay Cowell, Alexander Diehl, Harry Mobley, Bjoern Peters, Alan Ruttenberg, Richard Scheuermann, Ryan Brinkman, Melanie Courtot, Chris Mungall, Zuoshuang Xiang, Fang Chen, Thomas Todd, Lesley Colby, Howard Rush, Trish Whetzel, Mark Musen, Brian Athey, Gilbert Omenn, and Barry Smith (2009). VO: Vaccine Ontology. In *International Conference on Biomedical Ontology, Nature Precedings*.

Yongqun He, Ramesh Vemulapalli, and Gerhardt G. Schurig (2002). Ochrobactrum anthropi expressing Brucella abortus Cu,Zn superoxide dismutase protects mice against B. abortus infection only after switching of immune responses to Th1 type. *Infection and Immunity*, *70*(5), 2535–2543.

Yongqun He, Ramesh Vemulapalli, Ahmet Zeytun, and Gerhardt G. Schurig (2001). Inducion of specific cytotoxic lymphocytes in mice vaccinated with Brucella abortus RB51. *Infection and Immunity*, *69*(9), 5502–5508.

LaDeana W. Hillier, Alan Coulson, John I. Murray, Zhirong Bao, John E. Sulston, and Robert H. Waterston (2005). Genomics in C. elegans: So many genes, such a little worm. *Genome Res*, *15*, 1651–1660.

Gloria Yuen Fun Ho, Arnold Melman, S.M. Liu, Maomi Li, H. Yu, Abdissa Negassa, Robert Burk, Ann W. Hsing, R. Ghavamian, and Streamson Coo Chua, Jr. (2003). Polymorphism of the insulin gene is associated with increased prostate cancer risk. *Br J Cancer*, *88*(2), 263–269.

Robert Hoffmann and Alfonso Valencia (2005). Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, *21 Suppl 2*.

Dimitar Hristovski, Borut Peterlin, Joyce A. Mitchell, and Susanne M. Humphrey (2005). Using literature-based discovery to identify disease candidate genes. *Int J Med Inform*, *74*(2-4), 289–298.

Chun-Nan Hsu, Yu-Ming Chang, Cheng-Ju Kuo, Yu-Shi Lin, Han-Shen Huang, and I-Fang Chung (2008). Integrating high dimensional bi-directional parsing models for gene mention tagging. *Bioinformatics*, *24*(13).

Zhenjun Hu, Joseph Mellor, Jie Wu, and Charles DeLisi (2004). VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics*, *5*.

Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, *4*(1), 44–57.

Minlie Huang, Xiaoyan Zhu, Yu Hao, Donald G. Payan, Kunbin Qu, and Ming Li (2004). Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*, *20*(18), 3604–3612.

Bernardo A. Huberman and Lada A. Adamic (1999). Growth dynamics of the World-Wide Web. *Nature*, *401*(6749), 131.

Lawrence Hunter and K. Bretonnel Cohen (2006). Biomedical language processing: What's beyond pubmed. *Molecular cell*, *21*, 589–594.

Ken Hyland (1998). *Hedging in Scientific Research Articles*. John Benjamins Publishing Co.

Rob Jelier, Guido Jenster, Lambert C. Dorssers, Christiaan van der Eijk, Erik M. van Mulligen, Barend Mons, and Jan A. Kors (2005). Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics*, *21*(9), 2049–2058.

Hawoong Jeong, Sean P. Mason, Albert-László Barabási, and Zoltan N. Oltvai (2001). Lethality and centrality in protein networks. *Nature*, *411*(6833), 41–42.

Thorsten Joachims (1999a). *Advances in Kernel Methods-Support Vector Learning*, chap. Making Large-Scale SVM Learning Practical. MIT-Press.

Thorsten Joachims (1999b). Transductive Inference for Text Classification using Support Vector Machines. In Ivan Bratko and Saso Dzeroski, editors, *Proceedings of ICML-99, 16th International Conference on Machine Learning*, pp. 200–209. Morgan Kaufmann Publishers, San Francisco, US.

David A. Johnson (2008). Synthetic TLR4-active glycolipids as vaccine adjuvants and standalone immunotherapeutics. *Curr Top Med Chem*, *8*(2), 64–79.

Emmanuelle Jouanguy, Frederic Altare, Salma Lamhamedi, Patrick Revy, Jean-Francois Emile, Melanie Newport, Michael Levin, Stephane Blanche, Eric Seboun, Alain Fischer, and Jean-Laurent Casanova (1996). Interferon-gamma-receptor deficiency in an infant with fatal bacille Calmette-Guerin infection. *New England Journal of Medicine*, *335*(26), 1956–1961.

Maliackal Poulo Joy, Amy Brock, Donald E. Ingber, and Sui Huang (2005). High-betweenness proteins in the yeast protein interaction network. *Journal of Biomedicine and Biotechnology*, *2*.

Minoru Kanehisa and Susumu Goto (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30.

Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, *38*(Database issue), D355–360.

Minoru Kanehisa, Susumu Goto, Masahiro Hattori, Kiyoko F. Aoki-Kinoshita, Masumi Itoh, Shuichi Kawashima, Toshiaki Katayama, Michihiro Araki, and Mika Hirakawa (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic acids research*, *34*(Database issue), D354–357.

T. S. Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan, Arivusudar Marimuthu, Sutopa Banerjee, Devi S. Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, and et al. (2009). Human Protein Reference Database–2009 update. *Nucleic acids research*, *37*(Database issue), D767–D772.

Halil Kilicoglu and Sabine Bergler (2008). Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, *9*(Suppl 11).

Jin D. Kim, Tomoko Ohta, and Jun'ichi Tsujii (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, *9*(1).

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii (2009). Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop*.

Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, and et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.

Bettina Langhans, Susann Schweitzer, Ingrid Braunschweiger, Monika Schulz, Tilman Sauerbruch, and Ulrich Spengler (2005). Cytotoxic capacity of hepatitis C virus (HCV)-specific lymphocytes after in vitro immunization with HCV-derived lipopeptides. *Cytometry A*, *65*(1), 59–68.

Robert Leaman and Graciela Gonzalez (2008). BANNER: an executable survey of advances in biomedical named entity recognition. In *Pac Symp Biocomput*, pp. 652–663.

Florian Leitner, Martin Krallinger, Carlos R. Penagos, Jorg Hakenberg, Conrad Plake, Cheng J. Kuo, Chun N. Hsu, Richard Tsai, Hsi C. Hung, William Lau, Calvin Johnson, Rune Saetre, Kazuhiro Yoshida, Yan Chen, Sun Kim, Soo Y. Shin, Byoung T. Zhang, William Baumgartner, Lawrence Hunter, Barry Haddow, Michael Matthews, Xinglong Wang, Patrick Ruch, Frederic Ehrler, Arzucan Özgür, Güneş Erkan, Dragomir R. Radev, and et al. (2008). Introducing meta-services for biomedical information extraction. *Genome Biology*, *9*(Suppl 2), S6+.

Haifeng Li and Tao Jiang (2005). A Class of Edit Kernels for SVMs to Predict Translation Initiation Sites in Eukaryotic mRNAs. *Journal of Computational Biology*, *12*(6), 702–718.

Long-Cheng Li, Hong Zhao, Hiroaki Shiina, Christopher J. Kane, and Rajvir Dahiya (2003). PGDB: a curated and integrated database of genes related to the prostate. *Nucleic Acids Res.*, *31*(1), 291–293.

Marc Light, Xin Ying Qiu, and Padmini Srinivasan (2004). The Language of Bioscience: Facts, Speculations, and Statements In Between. In Lynette Hirschman and James Pustejovsky, editors, *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*, pp. 17–24 Boston, Massachusetts, USA. Association for Computational Linguistics.

Robert K. Lindsay and Michael D. Gordon (1999). Literature-based discovery by lexical statistics. *J. Am. Soc. Inf. Sci.*, *50*(7), 574–587.

Pasquale Mansueto, Giustina Vitale, Gabriele Di Lorenzo, Giovam Battista Rini, Serafino Mansueto, and Enrico Cillari (2007). Immunopathology of leishmaniasis: an update. *International Journal of Immunopathology and Pharmacology*, *20*(3), 435–445.

Lisa Matthews, Gopal Gopinath, Marc Gillespie, Michael Caudy, David Croft, Bernard de Bono, Phani Garapati, Jill Hemish, Henning Hermjakob, Bijay Jassal, Alex Kanapin, Suzanna Lewis, Shahana Mahajan, Bruce May, Esther Schmidt, Imre Vastrik, Guanming Wu, Ewan Birney, Lincoln Stein, and Peter D'Eustachio (2009). Reactome knowledgebase of human biological pathways and processes. *Nucl. Acids Res.*, *37*(suppl_1), D619–622.

Alexa T. McCray, Suresh Srinivasan, and Allen C. Browne (1994). Lexical methods for managing variation in biomedical terminologies. In *Proc Annu Symp Comput Appl Med Care*, pp. 235–239.

Ryan Mcdonald and Fernando Pereira (2005). Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics*, *6*(Suppl 1), S6+.

Ben Medlock and Ted Briscoe (2007). Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 992–999 Prague, Czech Republic. Association for Computational Linguistics.

Joseph C. Mellor, Itai Yanai, Karl H. Clodfelter, Julian Mintseris, and Charles DeLisi (2002). Predictome: a database of putative functional links between proteins. *Nucleic Acids Res*, *30*(1), 306–309.

Stanley Milgram (1967). The small world problem. *Psychology Today*, *1*(1), 60–67.

Scott Miller, Heidi Fox, Lance Ramshaw, and Ralph Weischedel (2000). A novel use of statistical parsing to extract information from text. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pp. 226–233 San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Tomohiro Mitsumori, Masaki Murata, Yasushi Fukuda, Kouichi Doi, and Hirohumi Doi (2006). Extracting Protein-Protein Interaction Information from Biomedical Text with SVM. *IEICE Transactions on Information and Systems*, *E89-D*(8), 2464–2466.

Makoto Miwa, Rune Saetre, Yusuke Miyao, and Jun'ichi Tsujii (2009). A rich feature vector for protein-protein interaction extraction from multiple corpora. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 121–130 Morristown, NJ, USA. Association for Computational Linguistics.

Andres Montoyo, Armando Suarez, German Rigau, and Manuel Palomar (2005). Combining knowledge- and corpus-based word-sense-disambiguation methods. *Journal of Artificial Intelligence Research*, *23*, 299–330.

Roser Morante and Walter Daelemans (2009). Learning the scope of hedge cues in biomedical texts. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pp. 28–36 Morristown, NJ, USA. Association for Computational Linguistics.

Alessandro Moschitti (2006). Efficient convolution kernels for dependency and constituent syntactic trees. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *Machine Learning: ECML 2006*, Vol. 4212, chap. 32, pp. 318–329. Springer Berlin Heidelberg, Berlin, Heidelberg.

Mark E. J. Newman (2003). The structure and function of complex networks. *SIAM Review*, *45*, 167.

Hwee Tou Ng and Hian Beng Lee (1996). Integrating multiple knowledge sources to disambiguate word senses: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*.

OMIM (2007). Online Mendelian Inheritance in Man, OMIM (TM). http://www.ncbi.nlm.nih.gov/omim/. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD).

Toshihide Ono, Haretsugu Hishigaki, Akira Tanigami, and Toshihisa Takagi (2001). Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, *17*(2), 155–161.

Arzucan Özgür and Dragomir R. Radev (2009). Detecting speculations and their scopes in scientific text. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 1398–1407 Singapore. Association for Computational Linguistics.

Arzucan Özgür, Dragomir R. Radev, and Yongqun He (2010). Mining of vaccine-associated IFN-gamma gene interaction networks using the Vaccine Ontology. In *Bio-Ontologies 2010: Semantic Applications in Life Sciences Workshop at ISMB*.

Arzucan Özgür, Thuy Vu, Güneş Erkan, and Dragomir R. Radev (2008). Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, *24*, i277–i285.

Arzucan Özgür, Zhuohuang Xiang, Dragomir R. Radev, and Yongqun He (2010). Literature-based discovery of ifn-gamma and vaccine-mediated gene interaction networks. *Journal of Biomedicine and Biotechnology*, *2010*, 13 pages.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd (1999). The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford Digital Library Technologies Project.

Carolina Perez-Iratxeta, Peer Bork, and Miguel A. Andrade (2002). Association of genes to genetically inherited diseases using data mining. *Nat Genet*, *31*(3), 316–319.

Carolina Perez-Iratxeta, Matthias Wjst, Peer Bork, and Miguel A. Andrade (2005). G2D: a tool for mining genes associated with disease. *BMC Genet.*, *6*, 45.

Eric M. Phizicky and Stanley Fields (1995). Protein-protein interactions: methods for detection and analysis. *Microbiol. Rev.*, *59*(1), 94–123.

Martin F. Porter (1980). An algorithm for suffix stripping. *Program*, *3*(14), 130–137.

James Pustejovsky, Jose Castano, Jason Zhang, Maciej Kotecki, and Brent Cochran (2002). Robust relational parsing over biomedical literature: Extracting inhibit relations. In *Proceedings of the seventh Pacific Symposium on Biocomputing (PSB 2002)*, pp. 362–373.

Dragomir R. Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Celebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Adam Winkel, and Zhang Zhu (2004). MEAD - a platform for multidocument multilingual text summarization. In *Proceedings of LREC 2004*.

Nabih M. Ramadan, Herbert R. Halvorson, Ana M.Q. Vande Linde, Steven R. Levine, Joseph A Helpern, and K. Michael Welch (1989). Low brain magnesium in migraine. *Headache*, *29*(9), 590–593.

Arun Ramani, Razvan Bunescu, Raymond J. Mooney, and Edward Marcotte (2005). Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology*, *6*(5), R40.

Jeffrey C. Reynar and Adwait Ratnaparkhi (1997). A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing* Washington, D.C, USA.

Thomas C. Rindflesch, Lorraine Tanabe, John N. Weinstein, and Lawrence Hunter (2000). EDGAR: extraction of drugs, genes and relations from the biomedical literature. In *Pac Symp Biocomput*, pp. 517–528.

Barbara Rosario and Marti A. Hearst (2004). Classifying semantic relations in bioscience texts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 430. Association for Computational Linguistics.

Sunita Sarawagi (2008). Information extraction. *FnT Databases*, *1*(3).

Sami Sarfaraz, Farrukh Afaq, Vaqar M. Adhami, Arshi Malik, and Hasan Mukhtar (2006). Cannabinoid receptor agonist-induced apoptosis of human prostate cancer cells LNCaP proceeds through sustained activation of ERK1/2 leading to G1 cell cycle arrest. *J Biol Chem.*, *281*(51), 39480–39491.

Noriko Sato, Taro Kawai, Kenji Sugiyama, Ryuta Muromoto, Seiyu Imoto, Yuichi Sekine, Masato Ishida, Shizuo Akira, and Tadashi Matsuda (2005). Physical and functional interactions between STAT3 and ZIP kinase. *Int Immunol.*, *17*(12), 1543–1552.

Kate Schroder, Paul J. Hertzog, Timothy Ravasi, and David A. Hume (2004). Interferon-gamma: an overview of signals, mechanisms and functions. *J Leukoc Biol*, *75*(2), 163–189.

Benno Schwikowski, Peter Uetz, and Stanley Fields (2000). A network of protein-protein interactions in yeast. *Nat Biotechnol*, *18*(12), 1257–1261.

Burr Settles (2005). ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, *21*(14), 3191–3192.

Guixiu Shi, Jianning Mao, Guang Yu, Jun Zhang, and Jiangping Wu (2005). Tumor vaccine based on cell surface expression of DcR3/TR6. *Journal of Immunology*, *174*(8), 4727–4735.

Neil R. Smalheiser and Don R. Swanson (1994). Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease. *Neurosci. Res. Commun.*, *15*(1), 1–9.

Neil R. Smalheiser and Don R. Swanson (1996a). Indomethacin and Alzheimers disease. *Neurology*, *46*, 583.

Neil R. Smalheiser and Don R. Swanson (1996b). Linking estrogen to Alzheimer's disease: an informatics approach. *Neurology*, *47*(3), 809–810.

Neil R. Smalheiser and Don R. Swanson (1998). Calcium-independent phospholipase A2 and schizophrenia. *Arch. Gen. Psychiatry*, *55*(8), 752–753.

Larry Smith, Lorraine Tanabe, Rie Ando, Cheng J. Kuo, Fang I. Chung, Chun N. Hsu, Yu S. Lin, Roman Klinger, Christoph Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig Struble, Richard Povinelli, Andreas Vlachos, William Baumgartner, Lawrence Hunter, Bob Carpenter, Richard Tsai, Hong J. Dai, Feng Liu, and et al. (2008). Overview of BioCreative II gene mention recognition. *Genome Biology*, *9*(Suppl 2).

Victor Spirin and Leonid A. Mirny (2003). Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA*, *100*(21), 12123–12128.

Padmini Srinivasan (2004). Text mining: generating hypotheses from MEDLINE. *J. Am. Soc. Inf. Sci. Technol.*, *55*(5), 396–413.

Hendrik Streeck, Nicole Frahm, and Bruce D. Walker (2009). The role of IFN-gamma Elispot assay in HIV vaccine research. *Nature Protocols*, *4*(4), 461–469.

Michael P. Stumpf, Thomas Thorne, Eric de Silva, Ronald Stewart, Hyeong Jun J. An, Michael Lappe, and Carsten Wiuf (2008). Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(19), 6959–6964.

Kazunari Sugiyama, Kenji Hatano, Masatoshi Yoshikawa, and Shunsuke Uemura (2003). Extracting information on protein-protein interactions from biological literature based on machine learning approaches. *Genome Informatics*, *14*, 699–700.

Don R. Swanson (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, *30*(1), 7–18.

Don R. Swanson (1988). Migraine and magnesium: eleven neglected connections. *Perspectives in biology and medicine*, *31*(4), 526–557.

Don R. Swanson and Neil R. Smalheiser (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artif. Intell.*, *91*(2), 183–203.

Don R. Swanson, Neil R. Smalheiser, and Vetle I. Torvik (2006). Ranking indirect connections in literature-based discovery: The role of medical subject headings: Research articles. *J. Am. Soc. Inf. Sci. Technol.*, *57*(11), 1427–1439.

Gyorgy Szarvas (2008). Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *ACL 2008*.

Hiroshi Takayanagi, Kojiro Sato, Akinori Takaoka, and Tadatsugu Taniguchi (2005). Interplay between interferon and other cytokine systems in bone metabolism. *Immunological Reviews*, *208*, 181–193.

Lorraine Tanabe, Natalie Xie, Lynne H. Thom, Wayne Matten, and W. John Wilbur (2005). GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics*, *6 Suppl 1*.

V. Glenn Tarcea, Terry Weymouth, Alex Ade, Aaron Bookvich, Jing Gao, Vasudeva Mahavisno, Zach Wright, Adriane Chapman, Magesh Jayapandian, Arzucan Özgür, Yuanyuan Tian, Jim Cavalcoli, Barbara Mirel, Jignesh Patel, Dragomir Radev, Brian Athey, David States, and H. V. Jagadish (2009). Michigan molecular interactions r2: from interacting proteins to pathways. *Nucl. Acids Res.*, *37*(suppl_1), D642–646.

Joshua M. Temkin and Mark R. Gilder (2003). Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics*, *19*, 2046–2053.

Domonkos Tikk, Philippe Thomas, Peter Palaga, Jörg Hakenberg, and Ulf Leser (2010). A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. *PLoS Comput Biol*, *6*(7), e1000837+.

Richard Tsai, Cheng L. Sung, Hong J. Dai, Hsieh C. Hung, Ting Y. Sung, and Wen L. Hsu (2006). NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. *BMC Bioinformatics*, *7*(Suppl 5).

Shoutaro Tsuji, Misako Matsumoto, Osamu Takeuchi, Shizuo Akira, Ichiro Azuma, Akira Hayashi, Kumao Toyoshima, and Tsukasa Seya (2000). Maturation of human dendritic cells by cell wall skeleton of Mycobacterium bovis bacillus Calmette-Guerin: involvement of toll-like receptors. *Infect Immun*, *68*(12), 6883–6890.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii (2005). Developing a robust part-of-speech tagger for biomedical text. In *Proceedings of the 10th Panhellenic Conference on Informatics*, pp. 382–392.

Marc A. van Driel, Koen Cuelenaere, Patrick P. C. W. Kemmeren, Jack A. M. Leunissen, and Han G. Brunner (2002). A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur J Hum Genet.*, *11*(1), 57–63.

J. Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, Mark Yandell, Cheryl A. Evans, Robert A. Holt, Jeannine D. Gocayne, Peter Amanatides, Richard M. Ballew, Daniel H. Huson, Jennifer R. Wortman, Qing Zhang, Chinnappa D. Kodira, Xiangqun H. Zheng, Lin Chen, Marian Skupski, and et al. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351.

Veronika Vincze, Gyorgy Szarvas, Richard Farkas, Gyorgy Mora, and Janos Csirik (2008). The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, *9*(Suppl 11).

S. V. N. Vishwanathan and Alexander J. Smola (2003). Fast kernels for string and tree matching. In *Advances in Neural Information Processing Systems 15*, pp. 569–576. MIT Press.

Hester M. Wain, Michael J. Lush, Fabrice Ducluzeau, Varsha K. Khodiyar, and Sue Povey (2004). Genew: the Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Res.*, *32*(D255-7), 1257–1261.

Robert S. Wallis, T. Mark Doherty, Phillip Onyebujoh, Mahnaz Vahedi, Hannu Laang, Ole Olesen, Shreemanta Parida, and Alimuddin Zumla (2009). Biomarkers for tuberculosis disease activity, cure, and relapse. *The Lancet Infectious Diseases*, *9*(3), 162–172.

Hui Wang, Dong Yu, Sudhir Agrawal, and Ruiwen Zhang (2003). Experimental therapy of human prostate cancer by inhibiting MDM2 expression with novel mixed-backbone antisense oligonucleotides: in vitro and in vivo activities and mechanisms. *Prostate*, *54*(3), 194–205.

Mengqui Wang (2008). A re-examination of dependency path kernels for relation extraction. In *IJCNLP'08: Proceedings of the third International Conference on Natural Language Processing*.

Duncan J. Watts and Steven H. Strogatz (1998). Collective dynamics of small-world networks. *Nature*, *393*, 440–442.

Marc Weeber, Henny Klein, Alan R. Aronson, James G. Mork, Lolkje T. W. de Jong-van den Berg, and Rein Vos (2000). Text-Based Discovery in Biomedicine: The Architecture of the DAD-system. In *Proceedings of AMIA, the Annual Conference of the American Medical Informatics Association*, pp. 903–907.

Marc Weeber, Jan A. Kors, and Barend Mons (2005). Online tools to support literature-based discovery in the life sciences. *Brief Bioinform*, *6*(3), 277–286.

Q. Wei, M. Li, Xuping Fu, Rong Tang, Y. Na, M. Jiang, and Yao Li (2007). Global analysis of differentially expressed genes in androgen-independent prostate cancer. *Prostate Cancer Prostatic Dis.*, *10*(2), 167–174.

Thomas Wieder, Heidi Braumuller, Manfred Kneilling, Bernd Pichler, and Martin Rocken (2008). T cell-mediated help against tumors. *Cell Cycle*, *7*(19), 2974–2977.

Jonathan D. Wren (2004). Extending the mutual information measure to rank inferred literature relationships. *BMC Bioinformatics*, *5*, 145.

Stefan Wuchty, Zoltan N. Oltvai, and Albert-László Barabási (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat Genet*, *35*(2), 176–179.

Ioannis Xenarios, Esteban Fernandez, Lukasz Salwinski, Xiaoqun Joyce Duan, Michael J. Thompson, Edward M. Marcotte, and David Eisenberg (2001). DIP: The Database of Interacting Proteins: 2001 update. *Nucleic Acids Res.*, *29*, 239 – 241.

Zuoshuang Xiang, Yuying Tian, and Yongqun He (2007). PHIDIAS: a pathogen-host interaction data integration and analysis system. *Genome Biology*, *8*(7), R150.

Akane Yakushiji, Yusuke Miyao, Yuka Tateisi, and Jun'ichi Tsujii (2005). Biomedical information extraction with predicate-argument structure patterns. In *Proceedings of The Eleventh Annual Meeting of The Association for Natural Language Processing*, pp. 93–96.

Meliha Yetisgen-Yildiz and Wanda Pratt (2006). Using statistical and knowledge-based approaches for literature-based discovery. *J. of Biomedical Informatics*, *39*(6), 600–611.

Soon-Hyung Yook, Hawoong Jeong, and Albert-László Barabási (2002). Modeling the Internet's large-scale topology. *PNAS*, *99*(21), 13382–13386.

Andreas Zanzoni, Luisa Montecchi-Palazzi, Michele Quondam, Gabriele Ausiello, Manuela Helmer-Citterich, and Gianni Cesareni (2002). MINT: A Molecular INTeraction Database. *FEBS Letters*, *513*, 135–140.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella (2003). Kernel methods for relation extraction. *J. Mach. Learn. Res.*, *3*, 1083–1106.

Aidong Zhang (2009). *Protein Interaction Networks: Computational Analysis*. Cambridge University Press.

Zhuo Zhang, Mao Li, Hui Wang, Sudhir Agrawal, and Ruiwen Zhang (2003). Antisense therapy targeting MDM2 oncogene in prostate cancer: Effects on proliferation, apoptosis, multiple gene expression, and chemotherapy. *Proc Natl Acad Sci*, *100*(20), 11636–11641.

Shaojun Zhao (2004). Named entity recognition in biomedical texts using an HMM model. In *JNLPBA'04: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 84–87 Morristown, NJ, USA. Association for Computational Linguistics.

Xiaojin Zhu (2005). Semi-supervised learning literature survey. Tech. rep. 1530, Computer Sciences, University of Wisconsin-Madison. http://www.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf.

Xiaojin Zhu, Zoubin Ghahramani, and John D. Lafferty (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, pp. 912–919. AAAI Press.

Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B. Cohen (2007). Frontiers of biomedical text mining: current progress. *Brief Bioinform*, *8*(5), 358–375.