Biomedical Information Extraction: Mining Disease Associated Genes from Literature

A Thesis

Submitted to the Faculty

of

Drexel University

by

Zhong Huang

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

February 2014

This thesis is dedicated to my parents and my family, for their love, support, and encouragement over the years.

Acknowledgments

Without strong support and encouragement from my supervisor and my family, my long journey pursuing Ph.D. degree wouldn't make possible.

My most sincere gratitude goes first to my supervisor, Dr. Xiaohua Tony Hu, who has inspired me in the beginning of this difficult yet rewarding journey and guided me through the past years. As a mentor and friend, Tony reminds me the role-model I shall follow. His excellent scientific characteristics, profound knowledge and extensive experience in information science and technology have constantly motivated me to overcome obstacles I encountered in my study. As one of his many Ph.D. students, I enjoyed very much working with him in an open environment that encourages independent thinking while emphasizes rigorous scientific training. His supervision from technical details to systematic view of theoretical framework will be my life-long fortune for my future research career.

I am grateful to support and critic guidance I received from my thesis committee members, Dr.Yuan An, Dr.Zekarias Berhane, Dr.Weimao Ke, and Dr.Susan Gasson. Over the past two years they have devoted time and provided me valuable advices and insights to guide me moving from dissertation proposal to completion of the study. I would also like to thank Dr. Jiexun Jason Li for his advise on my thesis proposal.

I would like to express my gratitude to Dr.Susan Wiedenbeck and Dr.Il-Yeol Song, for their help and academic advices during my study in the Ph.D. program. My gratitude extends to the college of computing and informatics for providing me travel funds to attend international conferences related to my work.

It is also my great pleasure to work with former and current fellow Ph.D. students in Dr.Hu's group. Especially I would like to thank Drs. Xuheng Xu, Xiaodan Zhang, Xiaohua Zhou, and Daniel Duanqing Wu, for their help and collaboration in formulating my research ideas and constructive technical discussion. Their success in their own Ph.D. study is always a great inspiration for me along the way.

I would like to thank Dr.Franca Cambi, a professor and neurologist, for her encouragement and long-term friendship. My interests in bioinformatics starts from my previous works in her laboratory at Thomas Jefferson University. Her enthusiasm on tackling biomedical research questions by applying multiple disciplinary methods influenced me to take challenges on a different career road.

Finally, I would deeply express my gratitude to my parents, my sister, and my family. Without their fully support and love, it is hard to imagine that I can reach this particular stage of completing the final thesis. I thank my little boy, Eric, for bring me endless joy by asking various of kids' questions that I enjoyed so much.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Biomedical Information Extraction: Mining Disease Associated Genes from Literature

Zhong Huang

Xiaohua Tony Hu, Ph.D.

Disease associated gene discovery is a critical step to realize the future of personalized medicine. However empirical and clinical validation of disease associated genes are time consuming and expensive. In silico discovery of disease associated genes from literature is therefore becoming the first essential step for biomarker discovery to support hypothesis formulation and decision making. Completion of human genome project and advent of high-throughput technology have produced tremendous amount of data, which results in exponential growing of biomedical knowledge deposited in literature database. The sheer quantity of unexplored information causes information overflow for biomedical researchers, and poses big challenge for informatics researchers to address user's information extraction needs. This thesis focused on mining disease associated genes from PubMed literature database using machine learning and graph theory based information extraction (IE) methods. Mining disease associated genes is not trivial and requires pipelines of information extraction steps and methods. Beginning from named entity recognition (NER), the author introduced semantic concept type into feature space for conditional random fields machine learning and demonstrated the effectiveness of the concept feature for disease NER. The effects of domain specific POS tagging, domain specific dictionaries, and named entity encoding scheme on NER performance were also explored.

Experimental results show that by combining knowledge base with concept feature space, it can significantly improve the overall disease NER performance. It has also shown that shallow linguistic features of global and local word sequence context can be used with string kernel based supporting vector machine (SVM) for efficient disease-gene relation extraction. Lastly, the disease-associated gene network was constructed by utilizing concept co-occurrence matrix computed from disease focused document collection, and subjected to systematic topology analysis. The gene network was then merged with a seed-gene expanded network to form heterogeneous disease-gene network. The author identified and prioritized disease-associated genes by graph centrality measurements. This novel approach provides a new mean for disease associated gene extraction from large corpora.

CHAPTER 1. INTRODUCTION

During the past decades, high-throughput proteomics techniques have been widely employed for identifying disease associated genes, proteins, and metabolites. It led to rapid accumulation of experimental data and research reports. Identifying biomarkers and their interaction network underlying different diseases has become an important step to realize the future personal medicine. Based on NIH definition, the biomarker is a wide range of markers that can be objectively measured and evaluated to indicate normal biological or pathogenic processes (Biomarkers Definitions Working Group 2001). To aid biomarker discovery, text mining techniques have been utilized to analyze heterogeneous data sources. PubMed database comprises 23 million literature citations in biomedical fields and have been undergoing rapid update with growing experimental data analysis from high-throughput -omics study. To develop an efficient text mining approach to reveal underlying disease associated biomarkers from huge amount of literature are therefore extremely needed. Biomarkers show significant diversity ranging from genes, proteins, nucleic acid, and small metabolites, and have been applied throughout disease prediction, prognosis, and during various stages of drug discovery. Moreover, due to the nature of high variability of gene, protein, and disease names used in biomedicine literature reports, semantic search and information extraction played an important role in biomarker mining from literature. Named Entity Recognition (NER) combined with semantic annotation of biological entities including domain specific ontology, dictionary and thesaurus are often used to extract biological entities from text in order to achieve high accuracy and recall. In the context of this thesis, biomarker candidates discovery is considered as discovery of hidden semantic relations between diseases and genes. With the advancement of text mining technology

and accumulation of proteomics data in faster pace, more and more researches have been focused on finding potential biomarker candidates from literature database as the first step of biomarker discovery. However, Finding disease associated genes from literature is a difficult task involving variety aspects of information extraction, from named entity recognition to relation extraction and extracted gene ranking. As a result, current knowledge on biomarkers in biomedical literature has largely remained unexplored. In this thesis we will present systemic approaches using information extraction theory to address several research questions related to the issue.

1.1. Biomarker introduction

Biomarker is biological substance that is commonly used in clinic tests and basic life science research to indicate certain biological states including disease. The NCI thesaurus defines biomarker as "*a variation in cellular or biochemical components or processes, structures, or functions that is objectively measurable in a biological system and that characterizes normal biologic processes, pathogenic processes, an organism's state of health or disease, likelihood of developing a disease, prognosis, or response to a particular therapeutic intervention*". Accordingly, biomarker can be classified into different categories for their specific role. Early detection biomarker is used as indicator of early stage of diseases, ranging from diabetes to cancer, and is becoming increasingly important as medicine paradigm is shifting from traditional passively reacting to disease towards proactively predicting and preventing of diseases. Diagnostic biomarker is routinely used in clinical tests as laboratory evidence of some diseases. Currently oncology and neurology are two major driven forces for diagnostic biomarker research and development. Prognostic biomarker determines the chances of patient to recover from disease or disease recurring. Surrogate biomarker is regarded as valid substitute of clinical

outcomes that are impractical to measure directly, such as death. As biomarker concept has been adopted by pharmaceutical industry in their R&D, data with surrogate biomarker has also been submitted to FDA for new drug application in recent years. Efficacy and toxicity biomarker are important indicators of efficacy or toxicological effects for a drug treatment in an in vivo or in vitro system. With advancement of translational medicine, there is also a need to bridge the preclinical research with clinical application using translational biomarker, which serves as the cross-species indicator of treatment response in both animal/organism models (preclinical setting) and human (clinical setting). Although anatomical structures acquired by imaging techniques are included in biomarker category as imaging biomarker, it is out of the scope for this thesis. We will focus discussion on biomarker of biological molecule origin with predictive power in medicine, typically genes, proteins, and metabolic products.

Biomarker discovery is traditionally based on hypothesis guided research using low-throughput laboratory techniques. In this model, scientists focus on only a few genes of interests that are guided by hypothesis and generated from prior knowledge. The advantage of this approach is that biomarkers and its participating cell signaling pathways are well characterized and the results are often validated empirically by independent laboratories. The disadvantage is obvious, due to the extreme complexity of genome (estimated 30,000 genes) and proteome (estimated 1,000,000 proteins and their derivatives), the traditional biomarker discovery approach is time-consuming and inefficient. New biomarker discovery platform is built on genomics, proteomics, lipidomics, and metabolomics data. The '-omics' data are produced by modern high-throughput technologies represented by DNA microarray for genomics study, and 2D electrophoresis, mass spectrometry, protein microarray for proteomics study. Generally at

least two groups of samples, one from health control subjects and another from patients or treated subjects, are needed to identify biomarkers. The ultimate goal of biomarker discovery is to reliably differentiate protein patterns among different groups. At the last stage of biomarker discovery, the differentially revealed proteins or peptide fingerprints are further validated using variety of computing and empirical methods.

Biomarker discovery is the critical step to realize the future personalized medicine. In this paradigm shifting view of medicine, the genetic background of individual is being taken into full consideration for disease prediction, prevention, diagnosis, and treatment. On one hand, current medicine failed to address individual variations that lead to high percentage of non-responsiveness among population for some treatment regimens. For example 50-100% cancer patients (lung, breast, brain) are not responding well to chemotherapy (Jones 2002). On the other hand, rapid advancement of full-genome sequencing technology is making individual's full genome sequencing more readily available to general population. In late 2006, Biomarker Consortium was founded in an aim to bring pharmaceutical industry, academia, healthcare organizations, NIH, and FDA together to accelerate and standardize the biomarker-centered basic and translational research. It is expected in the future biomarker will be widely applied on basic research and development, therapy, public disease prevention etc under the new framework of personalized medicine.

1.2. In silico discovery of biomarkers and information extraction

Published scientific papers amount to significant part of knowledge expressed as natural language to describe genes, proteins, metabolic molecules, drugs, diseases, and their semantic relationships. However, it poses great challenge for text mining systems to parse and extract

valuable information from such unstructured and noisy textual data. In 1950s Zellig Harris (Harris 1958) had formulated the idea of linguistic transformation of scientific papers into set of kernel sentences as the semantic structure. Modern information extraction (IE) methods follow Harris's philosophy by transforming unstructured text data into annotated corpora and utilizing statistical modeling to learn the underlying structures, with the ultimate goal of applying learned models on automatic extraction of structured semantic data from large unstructured text sources.

To assist the annotation process and further provide domain specific background information, ontology and metathesaurus including Gene Ontology, UMLS metathesurus etc, have been developed and integrated into most state-of-the-art IE systems. For example, Semantic relationships between biomedical entities are defined in UMLS semantic network. Currently it contains 134 entity types and 54 relations between those entity types. There are five major semantic types including organism, anatomical structure, biologic function, chemical, physical object, idea or concept. The primary link between the semantic types is the "isa" link which connects semantic types to a hierarchical tree. Other major semantic relationships include physically related to, spatially related to, temporally related to, functionally related to, and conceptually related to. This semantic network provides an invaluable tool for variety of IE tasks.

Like most IE tasks, mining biomarkers, e.g. genes associated with disease in the context of this thesis, is not trivial and requires pipelines of information extraction steps and methodologies. As will be discussed in detail in chapter 2, the pipeline generally include text preprocessing, feature representation, named entity recognition, relationship extraction, and prioritizing or ranking of extracted information. So far, web based tools including PolySearch (Cheng et al. 2008), iHop (Hoffmann and Valencia 2005), EBIMed (Rebholz-Schuhmann et al.

2006), and Semedico (Wermter et al. 2009) are four representative systems that can be used for biological entity associations mining from biomedical literatures. However above methods are largely based on dictionary, bag-of-words machine learning, and rule-based approaches. Therefore it is still an open research question to represent and utilize the semantic contextual features in entity association mining.

1.3. Motivations and research questions

Several challenges must overcome to improve the IE performance for disease associated gene mining. Firstly, biomedical named entities are highly variable and ambiguous compared with other domains, largely due to lack of naming conventions in different area of study, frequent use of abbreviations, synonyms etc. Recognition and disambiguation are two important steps to map variations of biomedical names in the text to unique biomedical entities in the curated databases. This problem is especially prominent in disease named entity recognition and need to be tackled for disease-associated gene mining. Secondly, despite wide application of IE on biomedical domain, the specific disease associated gene extraction is still new and much more works are needed based on current IE framework. At each step of entity recognition, normalization, and relation extraction, it is critical for machine learning approaches to capture the most representative textual features and semantic contextual information. Finally, although text graph representation to information retrieval has been studied in past years and has been shown to be a powerful representation model (Blanco and Lioma 2011), so far not much work has been done to apply graph theory on disease associated gene mining.

Motivated by above challenges, in this thesis different approaches were proposed to address following research questions:

1. How to better represent text with concept features to improve disease NER using machine learning based approach?

2. How to utilize linguistic features to develop efficient relation extraction model for disease-gene relation extraction?

3. How to represent gene-gene and gene-disease network in concept space and achieve dimension reduction for the concept text graph? How to incorporate concepts mined from literature with empirical data from protein interaction database to reveal and prioritize disease-associated genes by network topology analysis?

The rest of the thesis is organized as follows:

In chapter 2 the literature review on information extraction including document feature representation and concept space modeling is introduced. State-of-the-art information extraction algorithms related to machine learning, statistical modeling, and graph theory are highlighted.

In chapter 3 we attempt to address research question 1 on document feature representation and utilization of concept feature for conditional random fields modeling in disease and gene named entity recognition. Two annotated biomedical corpora will be used to experiment text preprocessing and different feature set for conditional random fields based learning of disease and gene NER.

In chapter 4 we will address research question 2 on relation extraction modeling by exploring the effect of contextual features on disease-gene relation extraction, using string kernel based support vector machine classification approach.

In chapter 5 a graph theory based IE framework will be proposed to answer research question 3 in the context of specific disease associated gene mining, e.g. how to represent gene-gene and gene-disease network in concept space and achieve dimension reduction for the concept text graph? And how to incorporate concepts mined from literature with empirical data from protein interaction database to reveal and prioritize disease-associated genes by network topology analysis? In the proposed integrated approach, concepts extracted from the disease focused literature will be semantically filtered, normalized, and used to construct text graph by concept co-occurrence to model the disease-associated gene network. The network will be further expanded by utilizing protein interaction data. And finally the network topology will be analyzed to identify and rank genes associated with the disease by centrality measurements.

In chapter 6 we will summarize what have been learned and discuss future works.

CHAPTER 2. LITERATURE REVIEW

The thesis concerns itself with information extraction of disease associated genes from biomedical text. More specifically, it focuses on recognition and extraction of genes, diseases, and their relationships from PubMed literature database. The chapter will give background and literature review on related text mining fields and general text mining workflow. Advances on information extraction (IE) including machine learning based and graphical model based IE methods will be introduced in more details.

2.1. Text mining and its core processing steps

Text mining (TM) can be broadly defined as a knowledge discovery process from large corpora of unstructured text collections. It is derived from data mining framework that utilizing machine learning and statistical methods to extract explicit rules and patterns from large and noisy data. Additionally, due to complexity of human languages, extra steps including natural language processing (NLP), information retrieval (IR), and knowledge management are also required as part of integrated text mining process. The mined information, often represented by a statistical model, can then be applied to real-world data for text classification, clustering, question and answering, or summarization tasks.

Statistical modeling and machine learning methods play a central role in modern text mining. Two critical steps are involved. The first is the feature selection which converts the unstructured text into structured data, and represent document with set of features and associated statistics. The second is the model selection which tries to model the random process by using

the statistics collected in the first step. The generated model can then be applied on real-world textual data to predict the outcome of the process.

The source of document collection for text mining come from dynamic online and static offline text repositories. In biomedical domain, PubMed database from National Library of Medicine (NLM) (NCBI) is a major source of dynamic information which contains more than 23 millions of abstracts in life science and is growing at an estimated rate of 40,000 new records each month (Pustejovsky et al. 2002). It has attracted much attention in computer and linguistic research fields in order to solve the information overloading problem when querying such huge database. In this thesis we utilized PubMed as a major source of text collection for disease-associated gene mining.

Analogue to data preprocessing in data mining, large collections of documents also need to be preprocessed for heterogeneous text input formats standardization and document representation. During this step the original textual data are normalized and non-informative data are removed by techniques of format converting, stop words removing, tokenization, part-of-speech (POS) tagging etc. Furthermore, the text needs to be represented by a set of document features, normally modeled by the representational model, to transform the unstructured document to its structured counterpart. Compared with data mining system, the textual feature sets are generally much larger in dimensionality and requires deliberate consideration for different text mining tasks. Indeed, most text mining algorithms and methods rely on this representative feature set to retrieve, extract, classify, and clustering information. It is noted that feature sparsity is the characteristic of text mining which is caused by high dimensionality of feature set for a large document collection while only small portion of the feature set is present in

each document (Feldman and Sanger 2007). To help reaching the balance between including rich set of features representing raw text more accurately and selecting the most essential features in terms of computation efficiency, external ontology and knowledge base pertaining to the underlying domain are often needed.

Text features can be generated from characters, words, terms, concepts, phrases, character n-grams, syntactic parse trees in the document. Characters are the most basic text unit consisting of letters, numbers, special symbols etc. Despite its high dimensionality, character feature space is regarded as the most comprehensive representation of the document. Document can also be represented by word level features after stop word filtering and tokenization. Tokenization algorithms parse the document by removing punctuations, numbers etc from the text. The term feature consists of either single word or multi-words phrases extracted from document after tokenization, lemmatization, and POS tagging. Lemmatization is used to normalize variants of word that share the same root (e.g. 'is', 'was', 'were' can be lemmatized to their root word 'be'). Similar to term feature, the concept feature is a single word or multi-words phrases that describes a concept extracted from document using annotated corpora, domain ontology, or lexicon. The difference between term and concept feature is that the later doesn't necessarily contain words/phrases from the document. For example the concept *apoptosis* can be used to represent *programmed cell death* in the document even though the concept word itself doesn't occur in the text. Concept feature has been implemented in several text mining systems including KDT (Feldman and Dagan 1995), which utilized concept hierarchy to represent the document.

As an important branch of data mining, text mining has become increasingly important in biomedical domain due to exponential explosion of clinical and research data. How to represent knowledge in a computational efficient way, to help researchers with knowledge visualization in multi-dimension space, and to facilitate the knowledge discovery process, is remarkably challenging involving multiple disciplinary efforts. From the cognitive science point of view, there are generally two goals for information and data representation. They are explanatory and constructive modeling respectively. Explanatory modeling formulates theories that are subject to experimental or simulation test. The constructive modeling, on the other hand, designs and builds artifacts that can accomplish certain cognitive tasks. There are generally two approaches to build information and data model. One is symbolic approach focusing on symbol manipulation. Another is associationism approach that attempts to associate and connect different information elements to form a semantic information network. In (Gärdenfors 2004) Gaerdenfors articulates that above methods are not adequate to model some cognitive phenomena and thus advocates a third modeling method that is based on geometrical structure of the information space. This new way of representing is termed conceptual modeling. Under the theory of conceptual spaces, Gaerdenfors proposed to represent information on the conceptual level using it as the framework.

Quality dimensions are used to represent qualities of objects and form framework that connects different objects by relationships. In conceptual space a collection of quality dimensions defines the space. Conceptual spaces are considered to be facilitator of knowledge sharing. Moreover, the paradigm shifts of disciplines can be regarded as conceptual spaces shift.

To present the conceptual model in mathematical way, conceptual space *S* can be described as multiple quality dimensions D_1,..., D_n space with each point in the space represented as vector $v = <d1, ..., dn>$. *n* is the number of dimensions.

The concept can be a region in the conceptual space. To precisely define the region of the concept in the space, it is necessary to follow criterion to define the topology structure of concepts in the multiple quality dimensions. Criterion *P* is described as follows:

Criterion *P*: A natural concept is a convex region of a conceptual space. By this criterion, every point between two points v_1 and v_2 in the region should also be localized in the region. This is also the 'betweenness' notion often mentioned in cognitive psychology. Natural concept notion is the key for the conceptual space modeling.

Convexity of space region works very well when it is applied on Prototype theories. In prototype theory, members of the objects are not equally representative. Some members are regarded as more representative than others, thus belongs to prototypical members. In convex region, a point can be judged as per its centrality. Those points with high centrality can form the prototype members.

Voronoi tessellation method is another example that convex space fits well with prototypic theories (Aurenhammer 1991). For a set of prototypical points $(P_1, ..., P_i)$ of the categories, every point *P* in the space can be measured by its distance to each of the points in the set P_i's. Based on the distance similarity, point *p* may belong to the same category as set of P_i. Therefore it will partition the space into convex areas. This technique has been used by others (PETITOT 1988) for characterization of the categorical perception of phonemes.

Gardenfors discussed semantics in the conceptual space framework in (Gärdenfors 2004) and proposed the criterion *L*. lexical expressions are represented semantically as natural concepts.

It still remains great challenge to apply conceptual space model on various domains, as the author has admitted. Thus it will also lead to great potentials for researchers in their domain of expertise to discover and build the underlying geographical structures of quality dimensions. It should be noted that in Gaerdenfors' conceptual space the quality dimensions are identified and measured through human's perception, which is different to the objective measurement in Physical world. Gaerdenfors' conceptual space theory may improve the organizing, sharing, visualization, and potentially re-discovery of knowledge in biomedical domain. In chapter 3 and 5, we will present our works that integrate semantic concept feature for biological entity recognition and their relation network modeling.

So far we have discussed how to convert unstructured textual data into structured data represented by document features as a whole. Depending on the information needs, further process are needed for information retrieval, extraction, and ultimately knowledge discovery. In 2.2. we will focus on information extraction (IE) and in 2.3 we will give a background review on the application of IE on biomedical domain, e.g. finding disease-associated genes from literature.

2.2. Information extraction

Information extraction is the process of recognition and extraction of entities and their relationships from text. IE has been widely applied on news wire, customer care and other commercial domains. In biomedical domain, information extraction is particularly attracting to researchers seeking novel relations between entities like genes, proteins, and drugs.

Same natural language processing steps can be applied on information extraction as other text mining tasks. Those include pipeline consisting of section and sentence splitting, tokenization, lemmatization, POS tagging, linguistic parsing and dependency analysis steps. The main goal of IE is to extract structured data including named entities and their predicted relations from unstructured and often noisy text.

2.2.1. Named entity recognition

Named Entity Recognition (NER) first appeared in Message Understanding Conferences (MUC) for recognition and classification of persons, organizations, locations (Grishman and Sundheim 1996). When applied on biomedical domain, NER has shown to be more challenging than general domains due to its versatile naming conventions, spelling variations, abbreviation, and synonyms. In general, approaches to NER can be categorized as being dictionary and rule-based, machine learning based, and the hybrid method.

2.2.1.1. Dictionary and rule-based NER

Dictionary approach is the most straight-forward method to identify named entity through dictionary matching. Rule based extraction relies on hand crafted or learned rules from annotated examples. The rule can be defined as list of contextual patterns that capture prominent properties of entities and the context in which they appear. The pattern is generally based on bag of features for tokens, which include but not limited to token itself, orthographical and morphological properties, dictionary entry matches, and POS. Taken the gene name "Epithelial Growth Factor" appearing in the text as example, the rule can be defined as two conditions shown below:

({Dictionary Lookup=Gene}{Orthography Type=Capitalized word}{3}) → Gene Names

Similar to the regular expression pattern matching, above example specifies a condition that the token matches with a entry in gene dictionary, and a condition that the token is capitalized consecutively for three times. In general, the hand-crafted rule is highly dependent on the domain knowledge.

If the source corpus is manually annotated, machine learning algorithms can be applied to automatically induce rules from a set of annotated training text by following a greedy hill climbing strategy. Such heuristic rule learning algorithms were proposed and implemented in $(LP)^2$ (Ciravegna 2001), FOIL (Quinlan 1990), and WHISK (Stephen Soderland, Claire Cardie 1999).

2.2.1.2. Machine learning (ML) based NER

Machine learning based NER approach, on the other hand, is language independent and more robust in terms of system performance. ML based approaches can be further divided into supervised learning and semi-supervised learning methods. Supervised ML utilizes large annotated corpus while semi-supervised ML only needs small size of annotated corpus (seeds) along with large un-annotated corpus.

ML approach based on probabilistic models have been shown to give better accuracy and robustness against noisy in NER as well other IE tasks. Among them, hidden Markov models (HMMs), maximal entropy (ME), and conditional random fields (CRFs) are prominent methods for ML based NER. HMM is the extension of Naive Bayes model and both belong to generative

approach modeling the joint probability distribution. CRF is regarded as the extension of ME model and both belong to discriminative approach modeling the conditional distribution.

Hidden Markov model (HMM)

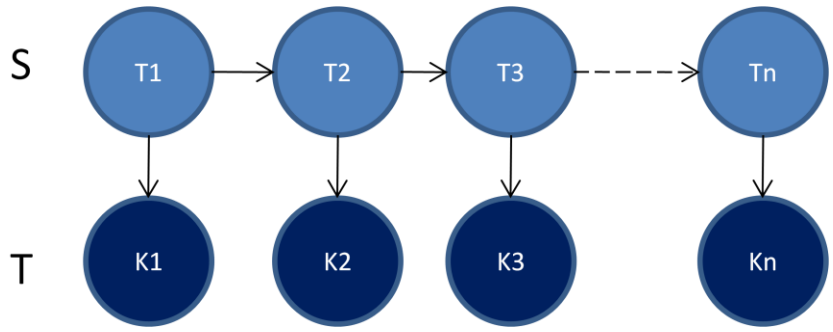


Figure 2-1. Hidden Markov model (HMM) for NER. The observation probability for token $k_i \in T$ $\{k_1, k_2, ..., k_n\}$ depends only on its current state t_i. And the current state t_i depends only on its previous state t_{i-1}.

HMMs are based on finite-state machine (FSM) which models the probability of finite state transitions and symbol emissions. The theory was first published by Baum etc (Baum et al. 1970) and was later applied on speech recognition. When it is applied on natural language NER, the problem can be formulated as a sequence labeling problem to find the likelihood of stochastic tag or label sequence $S = \{t_1, t_2, ..., t_n\}$ for a observed sequence of tokens $T = \{k_1, k_2, ..., k_n\}$ that maximizes the joint probability $P(S,T)$. Figure 2-1 illustrates the Bayes network graph of the HMM. S can be regarded as set of states of a finite state machine with each state corresponding to a named entity tag or label. Each observed token k is defined as $<f, w>$ where w is the token and f is the feature set for w. Each hidden tag t can be defined as $<p, c, f>$ where p is the position of current token in the named entity, c is its entity class, and f is its feature set. Formally, the joint probability $P(S,T)$ is defined as:

$$P(S,T) = P(S|T)P(S) \qquad 2\text{-}1$$

There are two assumptions concerning to this model. First is the so called Markov property which assumes the state t_i depends only on its previous state t_{i-1}. Second is the assumption that each observed token k_i depends only on state t_i. Therefore 2-1 can be represented as:

$$P(S,T) = \prod_{i=1}^{n} P(k_i|t_i)P(t_i|t_{i-1}) \qquad 2\text{-}2$$

If we relax the first assumption to assume state t_i depends on its previous state t_{i-1} and t_{i-2}. the first-order equation of 2-2 can be extended to following second-order form:

$$P(S,T) = \prod_{i=1}^{n} P(k_i|t_i)P(t_i|t_{i-1}, t_{i-2,}) \qquad 2\text{-}3$$

The solution is thus to find the sequence of states that maximizes the probability in 2-2 and 2-3 among all possible state sequences. However, Given a HMM and a training corpus, it is computationally prohibitive to calculate all probabilities exhaustibly. Instead, this problem can be efficiently solved by Vertibi algorithm (Viterbi 1967), a dynamic programming algorithm, using three probability distributions shown below.

$$\hat{x} = argmax_x \prod_{n=1}^{N} P(q_n|x_n)P(x_n|x_{n-1})P(x_0) \qquad 2\text{-}4$$

where $P(x_0)$ is the initial probabilities of state x_0, $P(x_n|x_{n-1})$ is the state transition probabilities, and $P(q_n|x_n)$ is the observation probabilities of the observed token q_n.

Maximum entropy model (ME)

ME was first proposed by Jaynes in (Jaynes 1957). In information theory, entropy is defined as measurement of uncertainty in a random variable *x*, i.e. the higher the uncertainty, the bigger the entropy. It can be formally written as :

$$H(p) = -\sum p(x) \log_2 p(x) \qquad 2\text{-}5$$

The philosophy of ME comes from the statistical inference on the basis of partial knowledge which makes as few assumptions or constraints as possible for the model output. In other words, ME model contains the maximum entropy with only those information constraints that are justified by the empirical data but not any arbitrary constraints. As a consequence, ME model preserves as much uncertainty or information content as possible (Ratnaparkhi 1997).

For natural language processing (NLP) tasks including NER, the problem can be stated as to estimate the probability of class *a* for a given context *b* in which *a* occurs, e.g. *P(a,b)*. The ME solution to this problem can be represented below to maximize the entropy:

$$H(p) = -\sum_{x \in \varepsilon} p(x) \log p(x) \qquad 2\text{-}6$$

where *x=(a,b)*, *a* belongs to set of possible classes *A*, *b* belongs to set of possible contexts *B*, and $\varepsilon = A \times B$.

By ME principle, equation 2-6 should accord with known facts about the partial knowledge. The known facts, also termed features, are expressed as a binary function shown in example below:

$$f_j = \begin{cases} 1, & condition\ on\ certain\ event \\ 0, & otherwise \end{cases}$$

Let *k* be the number of features and $1 \leq j \leq k$. The constraints can be expressed as:

$$E_p f_j = E_{\bar{p}} f_j \qquad 2\text{-}7$$

where $E_p f_j$ is the ME model *p*'s expectation of f_j, and $E_{\bar{p}} f_j$ is the observed expectation of f_j from sample data. According to 2-6, they can be represented as:

$$E_p f_j = \sum_{x \in \varepsilon} p(x) f_j\,(x)$$

$$E_{\bar{p}} f_j = \sum_{x \in \varepsilon} \bar{p}(x) f_j\,(x)$$

We then can define *P* set of all conditional probability distributions conforming to the constraints.

$$P = \{p \mid E_p f_j = E_{\bar{p}} f_j,\ \mathrm{j} = \{1, 2, \ldots, \mathrm{k}\}\} \qquad 2\text{-}8$$

It is worth note that ME models the conditional probability distribution while HMM models joint distribution. By applying the ME principle, we can choose the most informative model with the maximum entropy:

$$p^* = \mathrm{argmax}_{p \in P}\, H(p) \qquad 2\text{-}9$$

Conditional Random Fields (CRF)

CRF described in (Lafferty et al. 2001) is the state-of-the-art ML method for sequence classification problems including NER. Given a sequence of observations $x=\{x_1, \ldots, x_n\}$ the CRF

tries to model the probability $p(y|x)$ of output $y=\{y_1, ..., y_n\}$. CRF combines the idea of Hidden Markov Model (HMM) which deals with sequences problem, and Max-Entropy (ME) that utilizes many correlated features. In the meantime, it avoided label bias problem compared to Maximum Entropy Markov Models (MEMM) (McCallum et al. 2000), and is capable of handling arbitrary features with relaxed independence assumption as compared to HMM.

In text mining fields, the sequence of words is regarded as special case of linear chain of output nodes as illustrated below.

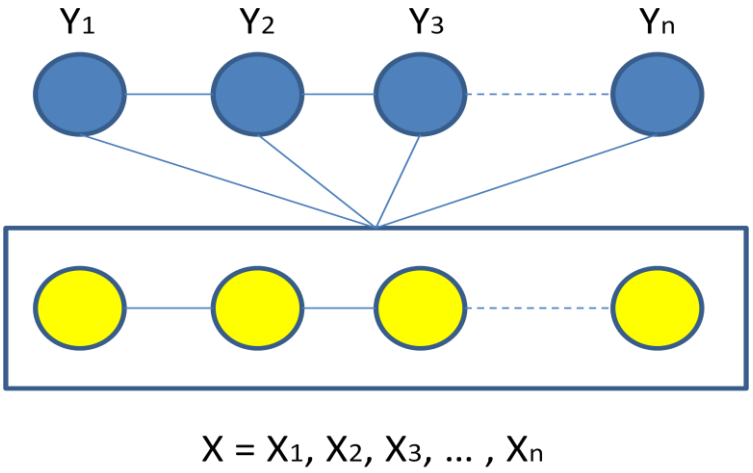


Figure 2-2. Simple illustration of first-order linear chain CRF graph. Y is sequence of output and X is the sequence of observations.

Lets define the undirected graph $G = (V, E)$ such that a node $v \in V$ and the random variable represents an element Y_v of Y which is indexed by the vertices of G. The (Y, X) is a conditional random field when conditioned on X, and the random field Y_v obeys the Markov property with respect to G. e.g. $p(Y_v \mid X, Y_w, w \neq v) = p(Y_v \mid X, Y_w, w \sim v)$ where $w \sim v$ denotes the neighbors in G. Therefore the CRF is a random field globally conditioned on the observation X.

For text labeling problem, let $o = \{o_1, o_2, \ldots, o_T\}$ be the observed sequence of words from a sentence with length r. Let S be a set of states in a finite state machine with each associated a label. The conditional probability of a state sequence $s = \{s_1, s_2, \ldots, s_T\}$ is calculated as:

$$P_{\wedge}(s \mid o) = \frac{1}{Z_o} \exp\left(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k (s_{t-1}, s_t, o, t) \right) \qquad (2\text{-}10)$$

where $f_k (s_{t-1}, s_t, o, t)$ is a feature function with λ_k as weight that can be learned during model training. The Z_o is a normalization factor of all state sequences which is used to sum up all conditional probabilities to 1 and is calculated as:

$$Z_o = \sum_{s} \exp\left(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k (s_{t-1}, s_t, o, t) \right) \qquad (2\text{-}11)$$

The objective function to be maximized in CRF model training is the log-likelihood of the state sequences given observation sequences:

$$L_{\wedge} = \sum_{i=1}^{N} \log\left(P_{\wedge}\left(s^{(i)} \mid o^{(i)} \right) \right) - \sum_{k} \frac{\lambda_k^2}{2\sigma^2} \qquad (2\text{-}12)$$

where $(s^{(i)} \mid o^{(i)})$ is the empirical distribution of training data. The L-BFGS algorithm is used for CRF parameter estimation and can be treated as a black-box optimization procedure (McCallum 2003).

In a nutshell, given a sentence of n words for named entity labeling problem (figure 2-3), we want to predict the tag T for a given word W using linear-chain CRF such that $P(T \mid W) = \frac{1}{Z} \exp\left(\theta \cdot F(T)\right)$ and maximize the weight $\theta \cdot F(T)$.

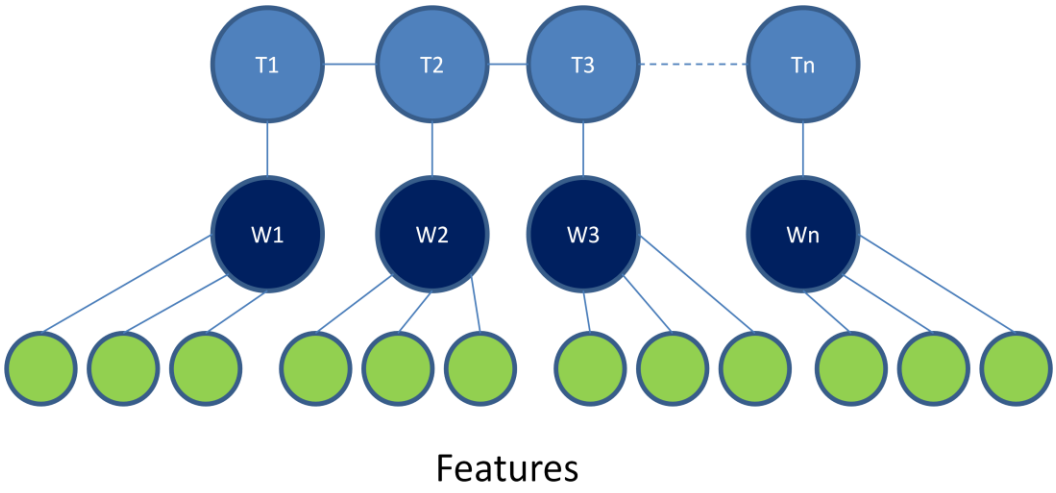


Figure 2-3. Illustration of linear-chain CRF as a labeling problem. W_1-W_n is sequence of observation (words) and T_1-T_n is sequence of tags.

In Chapter 3 we utilized second order linear-chain CRF for disease and gene named entity recognition.

2.2.2. Relation extraction

Relation extraction is one of the most important subject in IE. It refers to the method of identification and extraction of semantic relationships between named entities in the text. Broadly speaking, relations include semantic and grammatical relations, negation, and coreference etc. In biomedical domain, protein-protein interaction and disease-associated gene mining are two examples of relation extraction applications.

The relation extraction task can be defined as to identify the relations specified above between two entities in the text, normally at the sentence level, and assign the relation type to one of predefined relation types. Methods for relation extraction include supervised learning if

large corpora of annotated data is available, or semi-supervised and bootstrapping method if annotated corpora is limited.

For supervised learning utilizing annotated positive and negative examples, context information surrounding the related entities are extracted as features for learning the relation using statistical learning classifiers. Feature spaces that are useful for relation classification is reviewed in (Jiang and Zhai 2007). Among them, entity attributes e.g. entity types, bag of words, n-grams, grammar productions, dependency paths etc can be used as discriminative features for feature based classification. For corpora with large training set, the feature space is huge and makes it infeasible to search the space exhaustibly. In (Jiang and Zhai 2007) those feature spaces are systematically exploited by a bottom up approach, starting with a set of minimum features and adding more complex features to experiment the classification performances. Their results show that the basic unit features, which consists of bigrams and syntactic parse tree, is sufficient to achieve state-of-the-art performance while over fitting the classifier by adding complex features may decrease the overall performance. It suggests for each feature space, different feature representations may be redundant, even though it can increase robustness to noise but in the meantime may introduce more errors. Deliberate selection of most representative features is thus necessary to achieve better performance for feature based classification.

In (Zelenko et al. 2003) a kernel based relation classification method was introduced which is adapted from kernel method described in (Shawe-Taylor and Cristianini 2004). In contrast to feature based methods that directly rely on extracted features, kernel based methods utilize kernel function to compute the similarity score between pair of objects. Let $\{x^i, E_1^i, E_2^i, r^i\}$ represent an input training instance where x^i denotes the sentence, E_1^i and E_2^i denote entities, r^i

denotes the relationship and $r^i \in Y$ (relation types), $1 \leq i \leq N$ (N is the size of the training set). Let X_i denotes the $\{x^i, E_1^i, E_2^i\}$ of a training instance, and $X = \{x, E_1, E_2\}$ denotes a new instance for which the relation is to be predicated. The relation $\hat{r}$ for the new instance can be computed by:

$$\hat{r} = \text{argmax}_{r \in Y} \sum_{i=1}^{N} \alpha_{ir} K(X_i, X) \qquad 2\text{-}13$$

where $K(X_i, X)$ is the kernel function for similarity computing, and α_{ir} can be estimated during training process (Sarawagi 2007). Kernel function $K(X_i, X)$ is defined over structures like parse tree or dependency graph, without the need to convert those structures to flat sequence of features required by feature based methods. In chapter 4 we will present our work of extracting disease-gene relationship from text corpora based on kernel method and SVM classifier.

2.3. Graph theory and information extraction

Graph theory plays an important role in many disciplines including biomedical domain, where biological network is found to be an invaluable tool to model the complex biological processes. In chapter 5 we will apply graph theory on disease associated gene networks construction. In this section we will review the fundamental basics of graph theory, focusing on undirected graph.

A graph G is a finite set of vertices $V(G)$ connected by set of edges $\varepsilon(G)$, defined as $G = \{V(G), \varepsilon(G)\}$. If the edge connecting two vertices is directed, the graph is a directed graph, or a undirected graph if otherwise. Most biological networks, including protein-protein interaction network and gene-disease network described in chapter 5, are treated as undirected graph.

For undirected graph, there exist at most one edge between any two vertices. The size or order of a graph is defined as its total number of vertices. Let *u* and *v* be the vertices in above graph G. The degree for a node *u* is the total number of edges at *u*, or its neighbors denoted as $|N(u)|$, e.g. $\deg(u) = |N(u)|$. For edge *uv* in the edge set $\varepsilon(G)$ of a graph, vertex *u*'s neighbors $|N(u)|$ is given by:

$$|N(u)| = \{v \in V(G) : uv \in \varepsilon(G)\} \qquad 2\text{-}14$$

where edge *uv* is equal to *vu* for undirected graph.

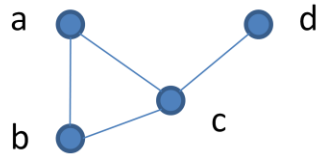
The degree distribution *P(k)* defines the probability distribution of all nodes with degree of *k*, e.g. $\frac{n_k}{n}$ where n is the total number of nodes in the graph and n_k is the number of nodes with exact degree of *k*. If *P(k)* distribution follows the power law, e.g. $P(k) \sim k^{-r}$, it is called a scale-free network (Barabási, A. 1999).

Given set of ordered vertices $v_1 \sim v_n$ and set of graph edges $\varepsilon(G)$, the undirected graph *G* can be mathematically represented as a binary symmetric adjacency matrix *A*:

$$a_{ij} = \begin{cases} 1, & if\ v_i v_j \in \varepsilon(G) \\ 0, & if\ v_i v_j \notin \varepsilon(G) \end{cases} \qquad 2\text{-}15$$

where $v_i\ and\ v_j$ are adjacent if $v_i v_j \in \varepsilon(G)$.

Example of the symmetric adjacency matrix *A* for a simple undirected graph is illustrated below:



	a	b	c	d
a	0	1	1	0
b	1	0	1	0
c	1	1	0	1
d	0	0	1	0

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Figure 2-4. Illustration of symmetric adjacency matrix for a simple undirected graph.

A walk in the graph *G* is a finite sequence of vertices and edges between the initial and terminal vertices *u*, *v*. If $u \neq v$ it is a open walk, otherwise it is a closed walk. In above example, a walk between *a* and *d* could be *a,b,c,d* or *a,c,d* and their connecting edges. The length of the path between *u* and *v* is *k*-1, where *k* denotes the number of vertices along the walk. The distance *δ(u,v)* is the shortest path between *u* and *v*. The diameter of the network is defined as the longest shortest paths of all calculated shortest path in the graph.

A local measurement for the degree of a node *u*'s clustering tendency is the clustering coefficient C_u, which equals to:

$$c_u = \frac{2e_u}{k_u(k_u - 1)} \qquad 2\text{-}16$$

where k_u is the number of neighbors of node *u*, e_u is the number of connected pairs between all neighbors of *u*. It can be understood as number of triangles pass through the node *u* divided by the maximum possible triangles that can be formed by its neighbors. In above example, node *c*'s clustering coefficient is 1/3, e.g. actual triangles pass through *c* of 1 (*abc*) divided by maximum possible triangles of 3 (*abc, acd, bcd*). Intuitively, it is an important

measurement of small-world network in which most nodes in the network are not connected directly and can only be reached from others by a small number of hubs. The characteristic of small-world network is that the distance (shortest path) between any two random nodes grows slowly to the number of nodes in the network *N*, e.g. proportionally to the logarithm of *N*.

Biology network consists of biological objects as nodes, and interactions between objects as edges. The biological objects account for genes, proteins, metabolites, and phenotypes or diseases. The structure properties or topology of complex real-world networks, including biology networks, are often exploited by comparing them with their random network counterpart (Erdős–Rényi random graph model) (P. Erdos 1960) which is stochastically generated by adding edges to same set of vertices with equal probability. Unlike the random network which follows a Poisson degree distribution and tend to have a lower average clustering coefficient, biology networks have been shown to have a power law degree distribution and much higher average clustering coefficient (Jeong et al. 2000) (Lee et al. 2009), and are organized by statistically significant motifs (Shen-Orr et al. 2002). Another characteristic of biology network is its small world property (Watts and Strogatz 1998), e.g. the diameter and average path lengths are small and proportional to the logarithmic of total node numbers. This phenomenon has been observed in variety of biology networks including metabolic networks (Wagner and Fell 2001), genetic networks (Tong et al. 2004), and protein interaction networks (Wagner 2001) (Yu et al. 2004). It is worth note however, current biology networks are based on sampled sub-networks consisting of only fraction of known biological objects instead of the complete network with all biological objects. Caution is needed when making conclusion on overall biology network structure based on aforementioned partial and sometimes inaccurate data (Mason and Verwoerd 2007a).

One important task for complex network modeling is to identify the most important vertices that are crucial to the network stability. In biology network it is to identify the most important genes and proteins that are critical to the network robustness and resistant to errors and attacks, as failure on those hubs will likely affect survival of the organism. In this regards, analysis of network centrality is an essential step. Commonly used centrality measures include degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality.

A. Degree centrality

The degree centrality is measured by nodes degrees. It has been reported in several protein interaction network studies that high degree nodes correlate with the essentiality of proteins (Zotenko et al. 2008).

B. Closeness centrality

Closeness centrality measures the distance $\delta(u, v)$ between nodes u and v. A node is deemed to be important when it can communicate more quickly with other nodes in the network. In protein interaction network, nodes with high closeness centrality plays role of bottleneck or cross-road that are often correlated with the degree centrality (Wuchty and Stadler 2003).

C. Betweenness centrality

Betweenness centrality measures the number of shortest paths passing through a node. Nodes lies between higher proportion of shortest paths are thought to be more important than nodes with fewer shortest paths passing through. In (Joy et al. 2005) it is found that yeast proteins with high betweenness but low degree are abundant in the network. The finding leads to

the hypothesis that proteins with high betweenness centrality but low degree connectivity is likely to be more essential.

D. Eigenvector centrality

Eigenvector centrality is calculated using the principal eigenvector of the adjacency matrix described before. In contrast to the degree centrality which assumes each neighbor contributes equally to its centrality, eigenvector centrality assigns high centrality scores to nodes that are connected to many central nodes. In other words, nodes with high eigenvector centrality scores receive more communications from other highly connected nodes and is thus more informative.

Applying different centrality measurements on biology networks is still an active research field. There is no simple unified solution to rank importance or essentiality for different types of biological objects and their interactions. Much more work is needed to disambiguate and further characterize the biology network topology. In chapter 5 we will further explore the topology of gene and disease-gene networks using different centrality measurements to identify and rank important disease-associated genes.

Term co-occurrence has been used to statistically represent text as graph model (Blanco and Lioma 2011). In this undirected text graph, vertices are terms and edges are term co-occurrence. It is assumed that co-occurring entities, including gene and protein, are functionally related. Co-occurrence based probabilistic models have been described for chemical compound-gene associations (Zhu et al. 2005), mutation-gene associations (Rebholz-Schuhmann et al. 2004), and cancer-gene associations (Zhu et al. 2006). By utilizing controlled vocabulary (MeSH

and GO), an algorithm was also proposed for scoring the possible associations between human genes and genetically inherited diseases based on co-occurrence (Perez-Iratxeta et al. 2002). In this thesis the gene-gene and gene-disease associations are extracted from biomedical text using concept co-occurrence. The network is further expanded using small number of seed genes and protein-protein interaction dataset. Our approach provide a novel way of identifying, prioritizing, and visualizing the important genes associated with specific disease.

CHAPTER 3. BIOMEDICAL NAMED ENTITY RECOGNITION BY MACHINE LEARNING

3.1. Introduction

Named Entity Recognition (NER) refers to the computational method to automatically recognize named entities (NE) in natural language documents, e.g. to relate it to a named entity (NE) in the domain of interest. For biomedical domain, an NE is defined as a term or phrase that denotes a biomedical object, for instance a protein, gene, disease, or drug with which a semantic hierarchy is associated. In this dissertation we will focus on gene, protein, and disease named entity type, which are directly associated with biomarker candidates discovery work presented here.

NER in biomedical text mining is particularly challenging. It is evidenced by the fact that many alias, different naming conventions, abbreviations, variety of organisms may refer a same protein/gene/disease with different terms, or a term may refer to biologically different entities. For example, named entity p53 may refer to a protein name in one context, but may also be used to denote the molecular weight of a protein with 53 Kd in another context.

Major classes of biomedical named entities includes genes, proteins, cells, drugs, chemicals, and diseases. Several high impact databases, including HUGO, Swiss-Prot, GenBànk, IPI, MedMaster, USP, UMLS, have been developed with intensive manual curation to support biomedical research community. Those databases provide rich resource for developing domain specific dictionaries, lexicons, and knowledge base for many text mining systems.

Some interesting patterns have been identified for biomedical NE. Linguistic pattern like upper case, comma, hyphen, slash, digit, and bracket have been noticed in examples such as 'proteolipid protein - 1', 'Thioredoxin h-type 1' etc. Many entities also contains semantic description (e.g. Epithelial Growth Hormone EGF, with description of protein expression location and function). However it is difficult to infer the functions for commonly seen abbreviations in literature without analysis of its semantic context in the place of occurrence (e.g. TCF may refer to gene 'T cell factor' or biomatrix 'Tissue culture fluid'). Due to the fact that biomedical names are expressed in various linguistic forms (plurals, compounds, abbreviations, anaphoric expressions) and relaxed forms of descriptions (prepositional phrases, relative phrases, phrases across sentences etc), the text mining system therefore should address above variations with respect to its problem-solving goals. A survey of name ambiguities, synonyms, and variations is given in table 3-1.

Table 3-1. Examples of biomedical entities and their linguistic or semantic form.

Example Biomedical Name	Linguistic or semantic form
Rpg1p/Tif32p	Compound name
TCF	Abbreviation
91 and 84 proteins	Coordination
p38 MAPKs	Plural
It, this protein, this enzyme	Anaphoric expression

Human epithelial growth factor	Semantic description
c-Jun N-terminal kinase (JNK)	Acronym
N-acetylcysteine, N-acetyl-cysteine, NAcytylCysteine	Synonym

3.1.1. Challenges in biomedical NER

As described above, biomedical NER is challenged by several aspects, e.g. ambiguous names, large amount of synonyms, acronyms, and linguistic variations. Furthermore, with rapid deposition of new literatures regarding novel gene and protein identifications, names of new biomedical entities needs to be taken into account for different text mining systems. It is especially true for biomarker candidates discovery which ideally should include finding associations between disease and new gene/protein names.

In light of the challenges in biomedical NER, the Critical Assessment of Information Extraction system in Biology (BioCreAtivE) was founded in 2004 which consists of a community-wide effort for evaluating information extraction in biomedical domain (Hirschman et al. 2005b). BioCreative II task 1A is concerned with the gene mention (GM) tagging, e.g. NE extraction of gene and gene product mentions in document. BioCreative II task 1B is human gene normalization (GN) task, which requires the text mining system to unambiguously map the human genes extracted from the text to the unique EntrezGene identifiers (Hirschman et al. 2005a). GN task is one step further after GM task in an aim to create distinct linkage between extracted NE and its biological database counterpart.

Another annotated corpus of MEDLINE abstracts, GENIA corpus, is also widely used as golden-standard for evaluation of NER algorithms (Kim et al. 2003).

3.1.2. Approaches for biomedical NER

Several text mining systems have been implemented for biomedical NER tasks using different approaches. Those approaches, in summary, can be categorized into following four categories.

3.1.2.1 Dictionary-based approaches

Dictionary-based approach is the most straightforward approach that tries to find all NE from text by looking up the dictionary. Some nomenclatures have been extensively applied on biomedical text mining. The HUGO Nomenclature for instance, provides more than 21,000 human gene entries (Cotton et al. 1998). The Swiss-Prot, the UniProt database containing more than 180,000 protein records has also been frequently used. The BioThesaurus collects comprehensive compilation of several million human protein and gene names mapped to UniProt knowledgebase entries using cross-reference in iProClass database (Liu et al. 2006). Unlike machine learning based approach, one advantage of dictionary based approach is that it has external database identifier (ID) built-in for each entry, thus provides external metadata annotation to the extracted names. However, it suffers from several limitations including false positive caused by name ambiguity, false negative cause by spelling variations and synonyms, and inability to cover newly created names. In addition, it heavily depends on creation and curation of lexicon for the specific domain which may consist of millions of entries and is very labor intensive. To address aforementioned spelling variation issue, (Tsuruoka and Tsujii 2004)

used approximate string searching and variant generator methods to achieve a significant improvement of F-measure (10.8%) on GENIA corpora evaluation as compared with exact matching algorithms.

3.1.2.2 Rule-based approaches

Rule-based approach can better deal with word orthographic and morphological structures, as compared with dictionary based approach. In (Fukuda et al. 1998) a method using surface clue on character strings was presented to identify core terms followed by handcrafted patterns and rules to concatenate adjacent words as named entity. The rule based approach largely depends on the domain specific named entities with common orthographic or morphologic characteristics. Thus makes it difficult to extend to other domains since the handcrafted rules are often domain specific and cannot be applied to a new domain due to different naming conventions.

3.1.2.3 Machine learning based approaches

Machine learning approaches are most frequently used and have achieved the best performance in BioCreative II gene/protein NER tasks. Different supervised machine learning methods including HMMs (Collier et al. 2000) (Zhou 2006), SVM (Jonnalagadda et al. 2013), MEMMs (McCallum et al. 2000), CRF (Lafferty et al. 2001), and Case-based reasoning (Neves et al. 2010) have been used in NER systems. In addition to supervised methods that utilize only the annotated text corpora, in order to solve data sparseness issue which often encountered when using large feature set on an relatively small training dataset, some semi-supervised methods are also presented recently to take advantage of large size of un-annotated text corpora. Such semi-supervised machine learning algorithms include semi-CRFs (Mann and McCallum 2007), semi-

SVMs (Kristin P. Bennett 1999), SVD-ASO (Ando and Zhang 2005), and FCG (Li et al.). Hybrid approach combining machine learning methods with dictionary or rule-based methods can also be used to improve the overall performance. For example in (Sasaki et al. 2008) a hybrid system combining dictionary and machine learning based statistical NER was used for protein name recognition. The critical step of machine learning approaches is to select the most discriminative feature. Commonly used features include orthographical word formation patterns, morphological patterns, part-of-speech POS tagging, lemmatization, token window, and conjunction of contextual features.

Machine learning (ML) based approaches use vector space to represent the text data and construct the model using labeled training data so that the model can be applied to predict unlabeled data. The key to success of ML based approaches lies on selecting vector features that have the most discriminative power. For NER task, the machine learning model is trained using training corpora which contains the specially formatted text and its associated annotation text. The annotation follows some guidelines tailored to certain collaborative activities such as BioCreative and BioNLP. An example of BioCreative training data is shown below:

```
P00001606T0076 Comparison with alkaline phosphatases and 5-nucleotidase
P00030937A0119 SGPT, SGOT, and alkaline phosphatase concentrations were
essentially normal in all subjects.
```

Text file: text sentence preceded by sentence identifier.

```
P00001606T0076|14 33|alkaline phosphatases
P00001606T0076|37 50|5-nucleotidase
P00030937A0119|0 3|SGPT
P00030937A0119|5 8|SGOT
P00030937A0119|13 31|alkaline phosphatase
```

Annotation file: annotation for each sentence proceeded by sentence identifier. The start and end position of each name are indicated (space not counted).

Two categories of corpora are commonly adopted by biomedical NER research community. One is golden-standard corpora (GSC) manually annotated by domain experts. The golden-standard corpora that have been widely cited include BioCreative (Hirschman et al. 2005b), JNLPBA (Kim et al. 2004), GENETAG (Tanabe et al. 2005), and PennBioIE (Kulick et al. 2004). Another corpora, also called silver-standard corpora (SSC), are those automatically annotated by NER systems. One such representative SSC is CALBC (Collaborative Annotation of a Large Biomedical Corpus) (Rebholz-Schuhmann et al. 2010). CALBC initiative aims to solve problem of small number of GSC (15,000-22,000 annotated sentences) due to labor intensive manual annotation, by automatically generating large scale named entity annotation (714,283 Medline abstracts) using a harmonized approach with annotations predicted by different NER systems.

Before using text as input to train machine learning model, the text preprocessing step is required to first divide document into sentences and tokens. Normalization techniques including stemming, lemmatization, part-of-speech POS tagging, and chunking are used at this step to provide local analysis of the token. Each token is subsequently tagged with the annotation scheme for the training corpora. Several annotation schemes have been applied on NER: The IO scheme tags token as either within (I) or outside (O) of the named entity. The BIO scheme is the most commonly used scheme which added beginning (B) of the named entity on top of IO scheme. The BMEWO scheme is used to further distinguish the NE containing multiple tokens and those containing only one token (W) by tagging the middle (M) and the end (E) of the token. The BIOLU scheme is used to indicate begin, inside, outside, last (L), and unit (U) (e.g. one word NE) of the token. Following example shows an original sentence taken from an MEDLINE abstract (PMID 10022891) and its annotation using BIO scheme.

```
We have identified a transcriptional repressor , Nrg1 , in a genetic screen
designed to reveal negative factors involved in the expression of STA1 ,
which encodes a glucoamylase .

We|O have|O identified|O a|O transcriptional|B-PROTEIN repressor|I-
PROTEIN ,|O Nrg1|B-PROTEIN ,|O in|O a|O genetic|O screen|O designed|O
to|O reveal|B-PROTEIN negative|I-PROTEIN factors|I-PROTEIN involved|O
in|O the|O expression|O of|O STA1|B-PROTEIN ,|O which|O encodes|O a|O
glucoamylase|B-PROTEIN .|O
```

Since feature representation and selection is a critical step required for NER machine learning, following paragraphs will review the current progress on text feature processing before discussing machine learning algorithms on NER.

Several surveys of state-of-the-art machine learning NER systems have been given in (Nadeau and Sekine 2007), (Leaman et al. 2008) and (David Campos , Sérgio Matos 2012). Among them, feature sets including orthographic features, morphological features, contextual features, and lexicons have been utilized to train variety of machine learning models. The authors concluded that those feature sets are essential to build a NER system with high F-measures as evaluated with golden-standard corpora (David Campos, Sérgio Matos 2012).

Orthographic features concerns with word formation. A linguistic orthography is a standard system to capture the token's word formation which includes capitalization, hyphenation, emphasis, punctuation, symbol, digit, and word breaks. Taken the example of a biological entity name "Interleukin-1 β", the first token starts with an upper case "L" followed by a hyphen and a Greek character, to denote a cytokine name. It is obviate that such orthographical feature can help to distinguish the named entity from other tokens within the context.

Morphological feature is used to analyze common structures of tokens being studied, which include suffixes/prefixes, char n-grams, and word shape. For example, the suffix '-ase' often denotes an enzyme, and '-in' often indicates a protein name. The char n-grams, on the other hand, extend the suffix/prefix to include characters in the middle of the token. The word shape pattern can be generalized to find the word/digit/symbol composition for a given token. For instance, the biomedical name "Interleukin-1" can be represented by the word pattern *Aaaaaaaaaaa#1* or *a#1*.

The local context of a token is also an important feature need to be captured. The relatedness measure between tokens and extracted features can be established through window or conjunctions to add contextual information to the token and utilize it as discriminative feature.

Compared with gene named entity recognition, so far disease named entity recognition has received much less attention and the performance needs to be improved (Leaman et al. 2008). In this chapter we attempt to address the research questions on how to improve the disease NER by incorporating domain knowledge base and semantic concept into preprocessing and feature representation, as the first step towards mining disease associated genes from literature.

3.2. Experiments design and methods

3.2.1. Data set

Two datasets were used for our NER experiments. For protein and gene name recognition we used the golden-standard GENETAG corpora from BioCreative II challenge of gene mention task (Hirschman et al. 2005b). The corpus contains 20,000 sentences chosen randomly from MEDLINE abstracts with low score of term similarity among documents to ensure its

heterogeneity. The corpus is divided into a training set of 15,000 sentences for model training, and a test set of 5,000 sentences for human judgment of participant's NER system performance. Training set was annotated by experts with biomedical background.

For human disease NER task the golden-standard BioText corpus was used. BioText corpus was originally annotated for disease and treatment mentions (Rosario and Hearst 2004) and is part of BioText Project at UC Berkley. The corpus was obtained from MEDLINE 2001 and contains 3655 annotated sentences. In our experiments, sentences labeled with <*TO SEE*> while lacking the close tag were removed and result in a final corpus of 3580 annotated sentences. Due to relatively small dataset, the 5 x 2 fold cross-validation (Dietterich 1998) was used for evaluation. The test is executed for 5 iterations of 2-fold cross-validation. Compared with 10 fold cross-validation, it is more powerful in terms of detecting real system performance differences rather than the biased splitting of testing data.

To extract the concepts from sentences we used semantic types of UMLS metathesaurus. It defines a comprehensive hierarchical tree of semantic network to represent all concepts in the UMLS metathesaurus as well as their relationships. This semantic network currently contains 133 semantic types and 54 relationships. Figure 3-1 shows the UMLS semantic network hierarchy related to disease.

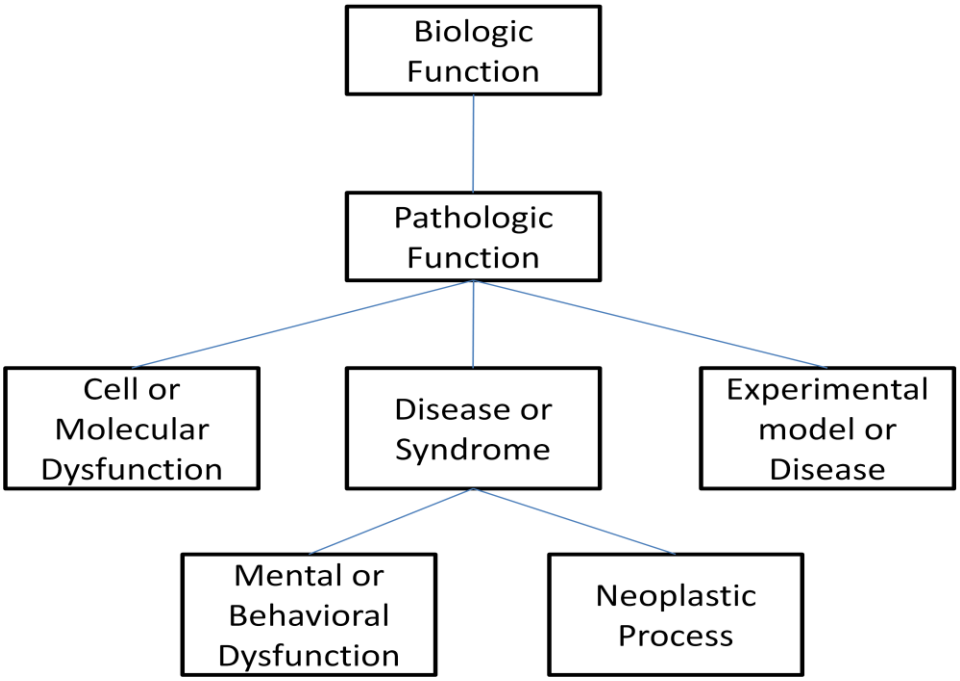


Figure 3-1. UMLS semantic network disease related semantic type hierarchy.

3.2.2. System architecture

Figure 3-2 shows the system architecture for disease NER. The corpus was first pre-processed by tokenization and lemmatization before feature extraction. Following (Leaman et al. 2008), we used feature set consisting of POS, lemma, orthographical and morphological features (patterns for word capitalization, letter and digit combinations, prefixes and suffixes). Numbers were normalized by converting digits to single digit "0". We used a simple tokenization method to tokenize the sentence. For POS tagging, we experimented with two different POS taggers implemented in Dragon Toolkit (Zhou et al. 2007), namely Hepple tagger and MedPost tagger. MedPost tagger is a POS tagger (Smith et al. 2004) specifically designed for biomedical text as compared with the more generic Hepple tagger (Hepple 2000).

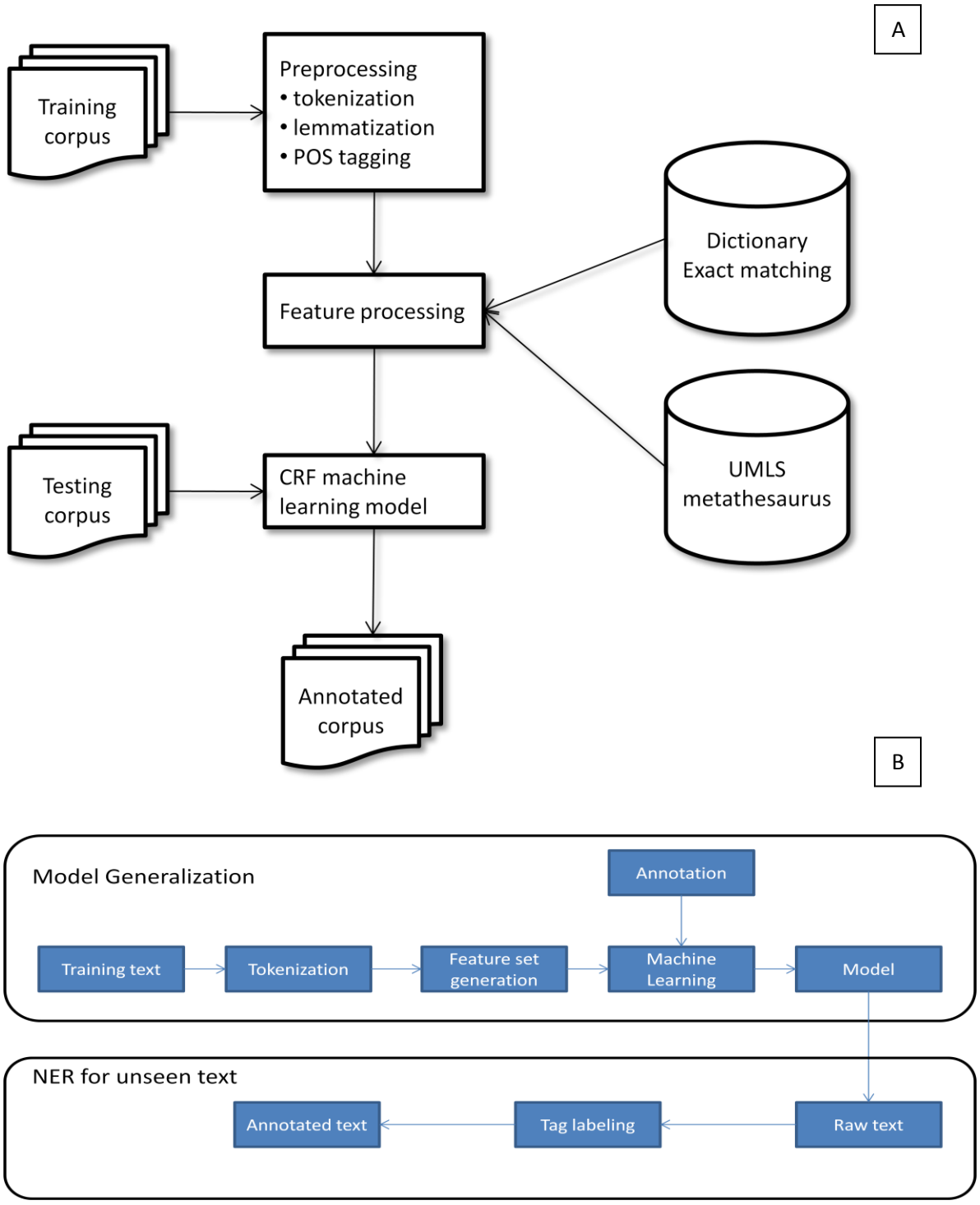


Figure 3-2. System architecture (A) and pipelines (B) for CRF machine learning based disease NER.

3.2.3. Feature engineering for NER

The architecture is a two-stage pipeline involving first stage of training the machine learning model for the NER, and the second stage of labeling the NE from raw text which is unseen in the training stage. The raw text of training data set is first tokenized using a tokenizer and the stop word is removed using a stop word list. At the feature processing step, collection of features is generated for each token, based on the experiment design detailed below in table 3-2.

Table 3-2. Feature set used for machine learning.

Category	Id	Features	Example and note
Orthographic features	Fcap	Capitalized word (start, end, all cap, mixed)	Interleukin, kappaB, MBP, RalGDS
	Fdig	Digits and counting	1, 12, 107
	Fsym	Symbols	-, /, [ ], \, :, ;, ., ”, *, =, %, ‘, ( ), +
Morphological features	Fwordshape	Word shapes	Represent “P50” as “A*
	Flem	Lemma	
	Fpos	POS tag	
	Fngrams	Char n-grams	
	Ffixes	Suffixes and prefixes	
Contextual features	Fwindow	Windows	For sentence “*Our data show that the transcriptional activity of IL-6 increases during*

			CVVH", the window feature of {-1, 1} for token "IL-6" can be captured as "of" and "increases".
	Fconjunction	Conjunctions	For sentence "*Our data show that the transcriptional activity of IL-6 increases during CVVH*", the conjunction feature of {-1, 1} of token "IL-6" can be captured as "of@-1_&_increases@1".
	Flexicon	Adding biomedical knowledge to the set of features using lexicon	BioThesaurus dictionary lookup

Exact dictionary matching using a disease dictionary was utilized to add biomedical knowledge semantic information to the feature.

One limitation of exact dictionary matching for NER is that it often gives false negative for spelling variations and newly created terms in the text. Moreover, it is highly dependent on the availability of domain specific dictionary which is not easily portable to other domains. For this reason, we used semantic types of UMLS metathesaurus to extract disease related concept from text as one of discriminative features, along with features described above, for NER machine learning. We used the approximate dictionary lookup algorithm in (Zhou et al. 2006) to capture the significant word in the text instead of capturing all words of the concept, and map it to the ontology term, e.g. UMLS semantic concept.

Let concept $c = \{s_1, s_2, s_3, \ldots, s_n\}$, where s_1-s_n are variant concept names that belong to c. $N(w)$ denotes number of concepts whose variant names contain word w.

The relative significance score of word w to the concept c is defined as:

$$I(w, c) = max\{I(w, s_j) \mid j \leq n\} \qquad (3\text{-}1)$$

where: $$I(w, s_j) = \begin{cases} 0 & w \notin s_j \\ \frac{1/N(w)}{\sum_i 1/N(w_{ji})} & w \in s_j \end{cases} \qquad (3\text{-}2)$$

The significant scores matrix containing normalized words as rows and concepts as columns were built using UMLS Metathesaurus (Zhou et al. 2006) and stored as sparse matrix for efficient retrieval. In equation 3-1 shown above, the w_{ji} denotes the word at *i-th* row which is found in concept s_j at *j-th* column.

The concept lookup algorithm uses rule-based pattern matching to search the word boundary and extract the concept term from text. In this study we used the default threshold score of 0.95 and the maximum number of skipped words of 1 which have been shown to give the best results for UMLS based biological concept extraction.

The word that is mapped to an UMLS concept is then filtered by its semantic type shown in figure 3-1. Only those concepts with semantic type of "DISEASE OR SYNDROME" are kept. The word with filtered semantic type is assigned a label and encoded as a new binary feature for model training at next step. The algorithm for the conceptual semantic feature generation is shown in Figure 3-3.

```
Algorithm concept feature engineering

Input:
1. sentence consisting of n words
2. semantic types of concept for extracted concept filtering

Output:
UMLS concept semantic type tagged sentence

SET sentence S = {w1, w2, ... , wn} where w1-wn is the pre-processed word
SET D = {d1, d2, ..., dm} where d1-dm is the semantic type of the concept to be kept
Initiate array of semantic type for w1-wn (ArrayW) and SET each value to 0
Initiate the list FL for final filtered concept names
Find next starting word ts
k = 0
C = {c | t ∈ T(c) } // T(c) is the set of words in concept c
For each c ∈ C, Sc = I(ts, c) // Sc is the significant score for word ts to concept c
WHILE next word t is not boundary word AND k < skip
    N = {c | t ∈ T(c) ∧ c ∈ C}
    IF N = Φ Then k = k + 1
    Else
      C = N
      For each c ∈ C
                Sc = Sc + I(t, c)
        End If
WhileEnd
C = {c | Sc > threshold ∧ c ∈ C }
If |C| > 0 then
    Return concept name and candidate concepts c ∈ C
End If

For each c ∈ C
    Get its UMLS term id tui
    Get the semantic type for tui
    If semantic type of the tui ∈ D
        Add the concept c to final list FL
For each word w in c ∈ FL
    Get its position index p (0≤p ≤n)
    Set ArrayW[p] = 1
Return array of semantic type for w1-wn (ArrayW)
```

Figure 3-3. Algorithm for the binary semantic concept feature engineering of disease named entity recognition.

The token is converted to name-value pair to feed the machine learning algorithm. As shown in table 3-3, each token is converted to list of binary features with value of either 1 (feature present) or 0 (feature not present) and associated with its name (token). Our feature

engineering approach was integrated into Banner toolkit (Leaman et al. 2008) to take advantage of its NER processing pipeline.

Table 3-3. The m x n matrix illustration of feature vectors for each token in the sentence.

	Feature 1	**Feature 2**	**Feature 3**	**…**	**Feature n**
Token 1	0	1	1	….	0
….	….	….	….	….	….
Token m	1	0	0	….	1

3.2.4. Conditional Random Fields (CRF)

We used conditional random fields (CRF) machine learning algorithm which has been proved to be a high performance method for label sequence problem. In (Lafferty et al. 2001) CRF was proposed as an undirected graphical model and the conditional probability of output nodes can be calculated based on other designated input nodes. The model defines a single log-linear distribution over label sequences of Y, given the observation sequence of X (Wallach 2004). In Chapter 2 (2-2-1-2) we have described the model in details. For our experiments, We used the 2-order CRF implemented in Mallet toolkit (McCallum 2002).

3.2.5. Evaluation method

Precision (P), recall (R), and F-measure (F-score) were used to evaluate NER experiments shown in formula below:

$$P = TP / (TP + FP) \qquad (3\text{-}3)$$

$$R = TP/ (TP + FN) \qquad (3\text{-}4)$$

$$F\text{-}score = (2 \times P \times R) / (P + R) \qquad (3\text{-}5)$$

where TP, FP, and FN are numbers of true positive, false positive, and false negative respectively. F-measure is a weighted average score combining both precision and recall, with score value ranging between 1 (the best) and 0 (the worst).

3.3. Result and discussion

3.3.1. Disease named entity recognition

We first compared the biomedical domain specific POS tagger MedPost tagger with generic Hepple tagger for disease NER task using BioText corpora. As shown in table 3-4, experimental results show an improvement in F-score by 1.23 using MedPost tagger over Hepple tagger when the disease specific dictionary is used. Compared with baseline Hepple tagger with non disease specific dictionary, the MedPost tagger with disease dictionary enhanced the F-measure by 1.67. The disease dictionary contains 25,944 entries of manually curated human disease names while non disease specific dictionary contains only gene and protein names. When a larger dictionary combining both non disease specific dictionary and disease specific dictionary was used, it slightly decreased precision, recall, and F-score of MedPost tagger.

Table 3-4. Evaluation with Hepple tagger and MedPost tagger. Non disease specific dictionary contains biological entities not specific to disease. The combined dictionary contains both non disease dictionary entries and the disease dictionary entries.

POS tagger + dictionary	Precision (%)	Recall (%)	F-score (%)
Hepple Tagger + non disease specific dictionary	62.82	47.79	54.28
Hepple Tagger + disease dictionary	63.29	48.21	54.72
MedPost Tagger + disease dictionary	64.93	49.15	55.95
MedPost Tagger + combined dictionary	64.45	48.80	55.54

We also compared different encoding scheme for disease named entity recognition. As discussed above in CRF section, NER can be modeled as a sequence labeling problem. Let $x = \{x_1, x_2, \ldots, x_n\}$ be the sequence of tokens for the input sentence, the problem is to determine the output sequence of labels $t = \{t_1, t_2, \ldots, t_n\}$ such that $t_i \in L$ (set of labels) for $1 \leq i \leq n$. The output label consists of two parts, e.g. the named entity type and its positional information. In this experiment we compared 3 named entity position encoding scheme, namely IO, BIO, and BIOEW. The IO coding is the simplest coding scheme that labels tokens as either Inside (I) or outside (O) of the named entity type. The BIO scheme adds Beginning (B) of the entity to IO scheme. The most complex coding is BIOEW which indicates the End (E) of entity and whether the token is a single word entity (W) on top of BIO scheme. Results shown in table 3-5 suggests the more complex coding schemes do not necessarily increase the F-score for BioText corpus NER task. The IO encoding scheme gives slightly better F-score than BIO and BIOEW schemes.

This is in agreement with the finding in (Leaman et al. 2008) that uses the BioCreative II corpus for gene/protein NER task. The IO setting is retained for our experiments.

Table 3-5. Results of evaluating different entity encoding scheme on BioText NER task. Hepple tagger and non disease specific dictionary were used as baseline for encoding scheme comparison.

Encoding scheme	Precision (%)	Recall (%)	F-score (%)
IO	62.82	47.79	54.28
BIO	63.40	47.13	54.07
BIOEW	63.11	46.61	53.61

As shown in table 3-4, the preliminary experiment using exact disease dictionary matching indicates the biomedical knowledge can improve the performance of disease NER. However, one limitation of exact dictionary matching is that it cannot handle spelling variants. We further experimented the effect of using concept semantic type as a new feature for disease NER. Table 3-6 shows results using the disease concept semantic type, e.g. "DISEASE OR SYNDROME" (type-1). The result without concept semantic type feature (type-0) is used as baseline for comparison.

Table 3-6. Results of evaluating effect of concept semantic types as feature for disease NER. Type-1 is "DISEASE OR SYNDROME" semantic type. Type-0 denotes no concept semantic feature added.

	Precision (%)	Recall (%)	F-score (%)
Type-0	64.93	49.15	55.95
Type-1	65.98	49.67	56.67

Table 3-6 shows that by adding "DISEASE OR SYNDROME" semantic type as feature to train the CRF model it achieves overall 0.72 increase of F-score, with 1.05 and 0.52 increase in precision and recall respectively.

Three NER systems for disease recognition using the BioText corpus and 5 x 2 cross-validation was reported in (Leaman et al. 2008). Comparing with reported results, our semantic concept type feature based method gives the highest F-score of 56.67 (BANNER: 54.84, ABNER: 53.44, and LingPipe: 51.15). This is largely due to the increase of recall (BANNER: 45.55, ABNER: 44.86, LingPipe: 47.50). The performance of disease NER using BioText by different systems are relatively poor, as compared with performance on gene and protein NER using BioCreative II gene mention task. This could be due to several reasons. First, the BioText golden-standard corpus is considerably small (3655 sentences versus 20,000 sentences for BioCreative II corpus), which is more likely to cause the data sparseness and out-of-vocabulary (OOV) issue. Secondly, unlike BioText that has only one annotation, the BioCreative II gene mention task provides an alternative annotation. Recently the silver-standard corpora (SSC), e.g. the automatically annotated corpora produced by machine learning models, have been used to supplement the golden-standard corpora (GSC) in an aim to boost the machine learning performance (Chowdhury and Lavelli 2011). It provides an alternative way to overcome above limitations caused by corpora size for disease NER task.

3.3.2. Gene named entity recognition

In addition to disease named entity recognition, we also exploited gene named entity recognition using BioCreative II corpora. As shown in table 3-7, BIO encoding scheme significantly enhanced the prediction performance by 7.64 as compared with IO scheme. BIOEW encoding only increased F-measure of IO scheme by 1.2. The BIO scheme is thus used for all gene NER tasks in rest of the thesis.

Table 3-7. Effect of encoding scheme on gene NER by CRF method (BioCreativeII corpora). MedTagger and non disease specific dictionary were used.

	Precision (%)	Recall (%)	F-measure (%)
IO	87.42	69.40	77.38
BIO	87.93	82.29	85.02
BIOEW	83.05	74.57	78.58

3.4. Conclusion and future work

The first challenge for our information extraction task is posed by the high variable nature of biomedical named entity. Named Entity Recognition (NER) has been an active research fields in biomedical text mining. In the past years, much attention has been focused on semantic types related to protein, gene, and other named entities in biology domain. Human disease named entity recognition in literatures, however, has not received much attention. Comparing the NER solutions for gene and protein named entities, existing machine learning solutions lacks same

level of precision and recall for disease named entity recognition. The development of machine learning based NER for disease named entity is largely focused on local features of tokens in the sentence by integrating its linguistic, orthographic, morphological, local contextual characteristics.

In this chapter we presented a new method of utilizing biomedical knowledge by both exact matching of disease dictionary and adding semantic concept feature through UMLS semantic type filtering, in order to improve the human disease named entity recognition by machine learning. By engineering the concept semantic type into feature set, we demonstrated the importance of domain knowledge on machine learning based disease NER. The background knowledge enriches the representation of named entity and helps to disambiguate terms in the context thereby improves the overall NER performance.

For the future work, it is interesting to further explore the effect of adding other relevant concept semantic types to feature set as high dimensional arbitrary features can be well handled in CRF model. It is also interesting to exploit the possibility of utilizing large silver-standard corpora, such as CALBC (Rebholz-Schuhmann et al. 2010), to train our concept based machine learning model and test it on the small size golden-standard corpus. It has been observed that by selecting those sentences of SSC containing annotations rather than the full SSC results gives the performance boost (Chowdhury and Lavelli 2011).

Another direction for the future work is to improve the computing efficiency by feature induction. Extraction of contextual features for each token by adding features of preceding and succeeding tokens through window, or by grouping features of preceding/succeeding tokens through conjunction has been studied in works (Zhang and Johnson 2003). Because the CRF is

log-lineal model, conjunction of features are necessary for projecting the feature space to a high dimensional space. On the other hand, considering for each token we have *n* features to select from to form the feature conjunction, it is important that the most informative features are selected. Although one significant advantage of CRF based sequence labeling over other machine learning algorithms such as HMMs is that it can handle arbitrary features without considering independence assumption, it is computationally infeasible to use complete set of contextual features surrounding the token, as it can result in extremely large feature set containing millions of features (Sha and Pereira 2003). In (McCallum 2003) a feature induction method was introduced to deal with the problem by automatically construct the most discriminative feature conjunctions. Starting from an empty feature set, feature induction algorithm takes input of list of user defined features and iteratively adds them to a dynamic feature set during training. Only those features with information gain will be preserved in the updated feature set. The feature induction algorithm given in (McCallum 2003) is summarized below:

algorithm feature induction for linear-chain CRF

input:

(1) Training set: paired sequences of feature vectors and labels.

(2) A finite state machine with labeled states and transition structure

output: A finite state CRF model that generate the most likely label sequences given an input sequence

Feature set K = 0

Do: Create list of candidate features using observational tests, conjunctions of observational tests with existing features

Limit number of conjunctions to those with highest information gain.

Add to K.

Apply an iterative quasi-Newton method to adjust CRF parameters to increase conditional likelihood of the label sequences given the input sequences.

while: convergence criteria is not met.

CHAPTER 4. INFORMATION EXTRACTION OF SEMANTIC RELATIONS BETWEEN DISEASE AND ITS ASSOCIATED GENES

4.1. Introduction

In this chapter we will discuss current research on information extraction (IE) of semantic relations from biomedical literature, and present our work on automatically extraction of disease gene relations utilizing the textual linguistic features with a string kernel based SVM classifier.

As discussed in chapter 1 and 2, disease associated biomarker mining from literature is a critical preliminary step prior to the laboratory research and clinical study phase. So far very few biomarkers have been identified and applied as clinical diagnostic and prognostic markers. On the other hand, the knowledge deposited in biomedical literature database doubles every 2-5 years which leads to accumulation of total 23 million citations in PubMed (NCBI). With current implementation of PubMed search engine, manual extraction of such information is the least efficient and most labor intensive way. It is therefore desirable to develop new method to automatically extract disease associated genes from literature. The problem can be formulated as semantic relations extraction from literature, which is a subject of information extraction study. Information extraction concerns itself with extraction of entities and their semantic relations from the unstructured text. Those relationships can be attributes of the entity, static facts, or dynamic events that exist between entities. In this chapter we are particularly interested in extracting fact relationships between biomedical entities, e.g. those facts that may imply a biological entity (gene or gene product) being a biomarker candidate of certain disease entity.

Most information extraction systems follow the bottom-up strategy to extract structured information frames from unstructured text. Like the relational database, the goal is to populate the predefined data frame with information extracted from the text. Taking a simple example of extracting name, title, contact number from highly heterogeneous web pages, the IE task is to parse the web content and extract person name as named entity, and its attributes including title, phone number etc. Generally it involves steps to tokenize the input text, analyze the morphological and lexical structure, analyze the syntactic structure, and integrate above annotated components in a domain knowledge framework representing the entities and their relationships. Because the natural language has characteristic of long-distance dependency (Jianfeng Gao 2005), one has to first resolve the co-reference or anaphora issue in order to extract relations between entities. It is particularly important for domains like social study, where names are frequently co-referenced between sentences. However this problem is much less significant in biomedical domain, therefore not tackled in this chapter. During each step the text words are disambiguated, syntactically parsed, and co-reference or anaphora resolution resolved. In previous chapter we have focused our work on biomedical named entity recognition, e.g. NER for gene and disease. In this chapter we will focus our efforts on extracting their semantic relationships in the context of biomarker definition, which can be viewed as an structured information framework containing disease and its associated genes or gene products.

4.2. Related works

Relation extraction has been extensively studied in newspapers, web content, emails etc. In biomedical domain, by querying PubMed with all known protein names it was found 269,000 out of 1.88 million PubMed abstracts were classified as being containing protein-protein

interaction relations (Donaldson et al. 2003). In another study ~150,000 gene and protein relations were extracted from one million PubMed abstracts (Fundel et al. 2007). And the number is soaring in recent years due to application of high-throughput technology. To facilitate automatic extraction of biomedical relations from the fast growing literature reports, BioCreative II (Hirschman et al. 2005b) and BioNLP (Pyysalo et al. 2012) (Björne et al. 2010) have included relation extraction tasks for protein-protein interaction, co-reference, and entity relations extraction. Both events rely on annotated GENIA corpora and focused on PPI, protein-component and subunit complex relation extraction. For relation extraction task the annotated corpora is indispensible for statistical machine learning based modeling, rule induction using rule-based methods, and for performance evaluation. Table 4-1 summarized current public annotated corpora for relation extraction in biomedical realm. For disease-gene relation extraction, to our knowledge, so far there is no publically available annotated corpora dedicated to this specific niche.

Table 4-1. Public biomedical corpora for relation extraction tasks. PPI denotes protein-protein interaction. AImed and HPRD50 are the only two corpora focusing on human PPI only.

Corpora	Corpora size	Type	References
AImed	225 abstracts	PPI (human)	(Bunescu et al. 2005)
BioInfer	1100 sentences	PPI	(Pyysalo et al. 2007)
HPRD50	145 sentences	PPI (human)	(Fundel et al. 2007)
IEPA	303 abstracts	PPI	(Ding et al. 2002)

LLL	77 sentences	PPI	(Nédellec 2005)
GENIA	9372 sentences	PPI	(Kim et al. 2003)
GREC	240 abstracts	Gene regulation	(Thompson et al. 2009)
IntAct	693 sentences	PPI	(Raja et al. 2013)

Generally speaking, relation extraction can be binary or multi-way of directed or undirected entity pairs. For directed pair in subject-object relation, the object of the relation is the target and the subject entity is the agent. The binary relation involves only two entities related to each other. While the multi-way relations involves three or more entities linked by the relationship. The protein-component and subunit complex relation extraction is a multi-way relation extraction where typically more than three proteins or protein subunits form a functional complex. Above two relations are illustrated in figure 4-1.

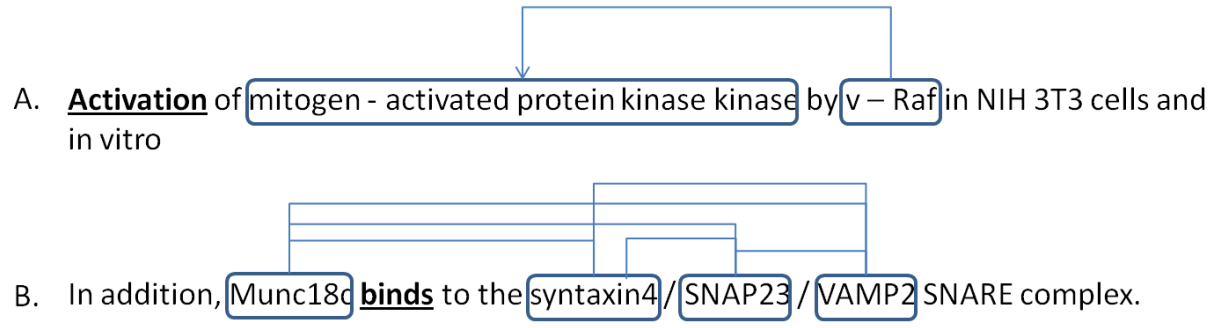


Figure 4-1. Illustration of common biomedical relations. A. Directed binary relation (activation) between two gene and protein pair. B. Undirected multi-way relation (binding) between subunits of a protein complex. (PMID 1326789, 16899085).

In following section we will discuss major relation extraction methods including statistical machine learning based approach, and rule-based approach.

4.2.1. Machine learning and statistics based relation extraction

First we define the relation as $r(e_1, e_2, \ldots, e_n)$ where e_i are entities with relation r in the text. The sentence s from which e_i are identified can be represented as $s = (w_1, w_2, e_1, \ldots, w_m, e_n)$, where $w_j (1 \leq j \leq m)$ is the word in the text. Given a corpora of positive and negative relation examples, in which e_i are annotated for an relation, the discriminative classifier can be trained using set of text features representing its local or global context shown below. Thus we can represent relation extraction as a classification problem that can be solved by supervised machine learning.

Commonly used features in relation extraction are summarized below:

- Bag-of-words, bigrams surrounding the entity (before, between, and after), lemma
- Entity types
- The distance between entities and the word sequence
- Syntactic parse tree paths, tree distance between entities

Parse tree is a tree graph representing syntactic structure of natural language based on formal grammar (Feldman and Sanger 2007). Constituent parse tree and dependency parse tree are two types of parse tree commonly used in text mining, with former one analyzed by constituency grammars (e.g. phrase structure grammars) and later one analyzed by dependency grammars without considering noun phrase (NP) or verb phrase (VP) categories. The

constituency parse tree of a given sentence is more complex than its corresponding dependency parse tree and therefore is more computationally expensive. The constituency parse tree and dependency parse tree from an example PubMed sentence are given below. We generated both parse trees using the annotation pipeline in Stanford Core NLP toolkit.

"Activation of mitogen-activated protein kinase kinase by v–Raf in NIH 3T3 cells and in vitro" (PMID 1326789)

constituency parse tree:

```
(ROOT
  (NP
    (NP
      (NP (NN Activation))
      (PP (IN of)
        (NP
          (NP
            (NP (JJ mitogen-activated) (NN protein) (NN kinase) (NN kinase))
            (PP (IN by)
              (NP (LS v))))
          (: --)
          (NP
            (NP (NN Raf))
            (PP (IN in)
              (NP (NN NIH) (NN 3T3) (NNS cells)))))))
    (CC and)
    (ADVP (FW in) (FW vitro))))
```

Dependency parse tree:

```
[Activation/NN
  prep_of:[kinase/NN
           amod:mitogen-activated/JJ
           nn:protein/NN
           nn:kinase/NN
           prep_by:v/LS
           dep:[Raf/NN prep_in:[cells/NNS nn:NIH/NN nn:3T3/NN]]]
  cc:and/CC
  advmod:[vitro/FW nn:in/FW]]
```

In (Jiang and Zhai 2007) feature spaces for relation extraction was systematically exploited using parse tree graph representation of the relation instance. It shows constituency parse tree feature gave better performance than dependency parse tree and sequence feature. But

the difference is small, suggesting each of three feature spaces is capable of capturing most structural information between entities. Further comparison between unigram, bigram, and trigram features shows the bigram performs significantly better than unigram, but trigram feature didn't improve it further.

For sentence $s = (w_1, w_2, e_1, \ldots, w_m, e_n)$ where w_j is the word and e_i is the entity with defined relation, the feature set ideally should include as much discriminative power as possible for e_i while minimizing the computational cost. Feature set containing full syntactic parsing is also called heavy-weighted feature set.

Based on the specific relation extraction problem e.g. binary or multiclass relation extraction, different classifiers including SVM, Max Entropy, Naive Bayes etc, can be used for the classification task. Support Vector Machine (SVM) is the most commonly used machine learning classifier for relation extraction.

Figure 4-2 shows the linear SVM model trained with samples from two classes by the hyperplane H. The machine learning task is to find the hyperplane that can separate two classes of vectors with maximum margin between two of them.

For a training set with sample size of *L*

(x_i, y_i), $x_i \in R^d$, $y_i \in \{+1,-1\}$, $i = 1, 2, \ldots, L$

The SVM is to find the hyperplane H: $W^T x + r = 0$

Maximize $W(\alpha) = \sum_{i=1}^{L} \alpha i - \frac{1}{2} \sum_{i,j=1}^{L} \alpha i \, \alpha j \, yi \, yi \, \langle xi, xj \rangle$

Subject to $\sum_{i=1}^{L} \alpha i \ yi = 0$

Where $\alpha i \geq 0, i = 1, 2, \ldots, L$

αi is the non-negative Lagrangian multipliers. Vector $\alpha i > 0$ when xi is a support vector, and $\alpha i = 0$ when it is not.

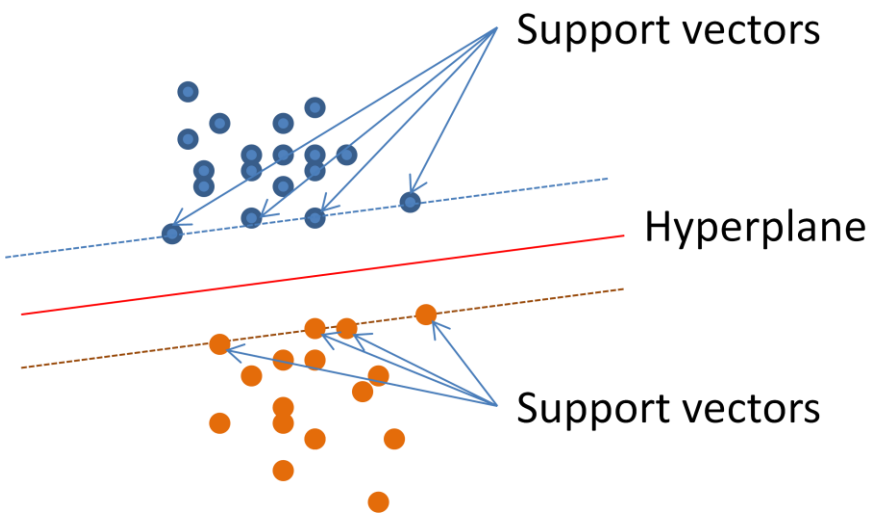


Figure 4-2. Illustration of Linear Support Vector and Hyperplane separation.

In case the data points between two classes are not linearly separable, a kernel function is needed to map dataset into higher dimensional space so that classes become separable. Kernel function $K(X_i, X)$ can also be thought of a similarity function for pair of structures X and X_i in the feature space (Kim et al. 2008). Lets represent each training data instance i as (x^i, E_1^i, E_2^i, r^i) and $X_i = (x^i, E_1^i, E_2^i)$, where r denotes the relationship, E denotes the entity, and x denotes the sentence. For a new instance $X = (x, E_1, E_2)$ we can classify it by predicting its relation $\hat{r}$ with formula 4-1 (Sarawagi 2007).

$$\hat{r} = \text{argmax}_{r \in y} \sum_{i=0}^{N} \propto_{ir} K(X_i, X) \qquad (4\text{-}1)$$

The $\propto_{ir}$ is the estimated weight for each training instance *i,* and *y* is the class. *N* denotes number of training sets.

Kernels for computing $K(X_i, X)$ used in relation extraction is based on string kernels proposed in (Lodhi et al. 2002), mathematically represented in 4-2 as:

$$K(X_i, X) = \sum_{u \in U} \emptyset_u(X_i)^T \emptyset_u(X) \qquad (4\text{-}2)$$

where *U* is the set of all possible sub-structure in structure *Xi* and *X*. The $\emptyset_u(X_i)$ and $\emptyset_u(X)$ are decay factor $\in (0,1)$. The term "structure" can be generalized to any object including string, sequence of words, parse tree etc. For relation extraction the structures are represented as word sequences before/between/after related entities using Bag of features kernel approach, or parse trees containing the entity using Tree kernel approach (Bach and Badaskar 2007). Kernels developed using above approaches include tree kernel (TK) (Zelenko et al. 2003), dependency tree kernel (DTK) (Culotta and Sorensen 2004), shortest path dependency kernel (SPDK) (Bunescu and Mooney 2005), subsequence kernel (SK) (Bunescu and Mooney 2006), composite kernel (CK) (Zhang 2006)(Zhang et al. 2011).

If the learning method utilizes only the labeled data for training, it is supervised machine learning. If it utilizes small set of labeled and large set of unlabeled data for training, it is semi-supervised. Semi-supervised methods rely on iterative learning by taking output of learner from last iteration and are becoming an important alternative to supervised approach, due to limited availability of high quality labeled data.

4.2.2. Pattern-based relation extraction

This approach uses handcrafted patterns or automatically generated patterns to extract relations. Patterns can be simple regular expression matching rules, or more complicated surface patterns consisting of POS tags and phrasal structures. The most sophisticated pattern representation involves syntactic and semantic structure analysis by full parsing, for instance to produce subject-verb-object (SVO) structure or predicate-argument structure (PAS) (Surdeanu et al. 2003). The major advantage of manual pattern is its high precision. The major disadvantage for handcrafted patterns is poor generalization from one domain to another, which also leads to relatively low recall because the manual pattern will not be able to cover all possible relation structures. This issue can be alleviated by automatically generated patterns. Bootstrapping methods, for example, extract patterns from small set of relation examples (seeds) and iteratively expand the seeds by applying them on new data (Agichtein and Gravano 2000).

4.2.3. Disease and gene relationship extraction

Disease-associated genes are important biomarker candidates which have been used as indicators of diagnosis, disease progression, and treatment efficacy for the past years. For example, in neurodegenerative diseases including Alzheimer's disease, Huntington's disease, Parkinson's disease, the genetic factor plays a critical role and consequently the disease-causing genes were studied extensively. On the other hand, the gene-disease relation extraction from literature haven't received similar level of attention as protein-protein interaction, protein and its sub-cellular localization. Therefore, there are still large rooms left to improve performance of disease-gene relation extraction. In this chapter, we applied machine learning kernel methods based on works in (Bunescu and Mooney 2005) and (Giuliano et al. 2006) to extract Huntington disease - gene relation from PubMed literatures.

In terms of information extraction needs, two approaches have been applied on disease-gene relation mining, namely global mining of general disease-gene association and selective mining of specific disease-gene associations. EDGAR is a system for global extraction of genes, drugs, and cell types interactions from PubMed literature and can be used to query the disease-gene associations (Rindflesch et al. 2000). BITOLA is a literature-based information extraction system designed to extract relations between different concepts, such as disease and gene association, by association rule algorithm (Hristovski et al. 2005). The association rule has form of $x \rightarrow y(confidence, support)$ where support is their co-occurrence frequency and confidence is the percentage of records containing *y* concept (e.g. pathological functions or symptoms) with all records containing *x* concept (e.g. disease). For disease (*x*) gene (*z*) relation association, the algorithm first finds all concepts *y* such that $x \rightarrow y$ and then finds all concepts *z* (e.g. genes) such that $y \rightarrow z$. The algorithm then filter out all concepts *z* whose chromosomal location do not co-localize with chromosomal location of disease concept *x* by using HUGO gene nomenclature and LocusLink genetic loci information. Finally, the remaining set of *z* concepts (genes or gene products) are ranked as candidate disease associated genes. Recently, an rule-based with keyword matching algorithm for disease-gene extraction was also presented (Jung et al. 2013). In another work (Chun et al. 2006a) the binary pair of gene-disease was extracted from PubMed sentences using dictionary based matching approach followed by machine learning NER filtering. The filtering step which removed large set of false positive introduced by dictionary matching, improved precision of relation extraction by 26.7%, suggesting the critical role of entity recognition step for overall performance of disease-gene relation extraction. For specific disease gene relation extraction, in (Chun et al. 2006b) annotated corpora for prostate and gastric cancers

from PubMed were constructed to train the maximum entropy based NER and relation extractor. The authors reported a 92.1% precision of topic-classified relation recognition.

4.3. Experiments and Results

4.3.1. Experiment design and datasets

In our experiment we focused on Huntington disease related gene extraction and casted it to a binary classification problem. Taken following NER tagged sentence describing association between HD and NR2B, NR2A as an example:

We conclude that these two genes, coding for <GENE>NR2B</GENE> and <GENE>NR2A</GENE> subtypes mainly expressed in the striatum, may influence the variability in AO of <DISEASE>HD</DISEASE>. (PMID 15742215)

In this example two gene entities and one disease entity were identified to have disease-gene association relations *r(NR2B, HD), r(NR2A, HD).* Here we consider the relation being all molecular interactions including expression, genetic variation, regulatory modification, or general description of associations in the text.

Since annotated corpora for machine learning based disease-gene relation extraction isn't available, we started by constructing it using the PubMed citations in Genetic Association Database (GAD) (Kevin Becker, Kathleaen Barnes, Tiffani Bright 2004). GAD is a database containing manually curated genetic association information for human disease with links to corresponding PubMed citations. We compiled list of all PubMed ids related to Huntington disease from GAD and retrieved all abstracts from PubMed using Entrez e-Util API. Abstracts were automatically split into sentences and tagged with NER tagger described in our work in chapter 3. Sentences with at least one gene mention and one disease mention were selected.

Because the disease-gene relation extraction is considered as a binary relation extraction, in case the sentence contains more than one gene or disease mention, our system automatically makes copies of the sentence (instance of the sentence) so that each gene-disease pair is tagged as a training example. After manual verification and curation of all tagged sentences, a training datasets consisting of 117 positive examples and 64 negative examples was constructed. The annotated corpora was then processed by contextual kernel functions, and used subsequently to train and test on SVM classifier by 10 fold cross-validation. The performance of the SVM classifier was compared against a protein-protein interaction golden standard corpora AImed, which collects only human protein interactions. Table 4-2 shows the statistics of the two corpora used in our experiments and figure 4-3 summarized the system architecture of our kernel based relation extraction system.

Table 4-2. Statistics of two corpora used in the experiments. The constructed Huntington disease corpora from PubMed contains 181 annotated sentences and the AImed corpora contains 5625 annotated sentences.

	AIMED dataset	**Huntington disease dataset**
Positive examples	1008	117
Negative examples	4617	64
Total	5625	181

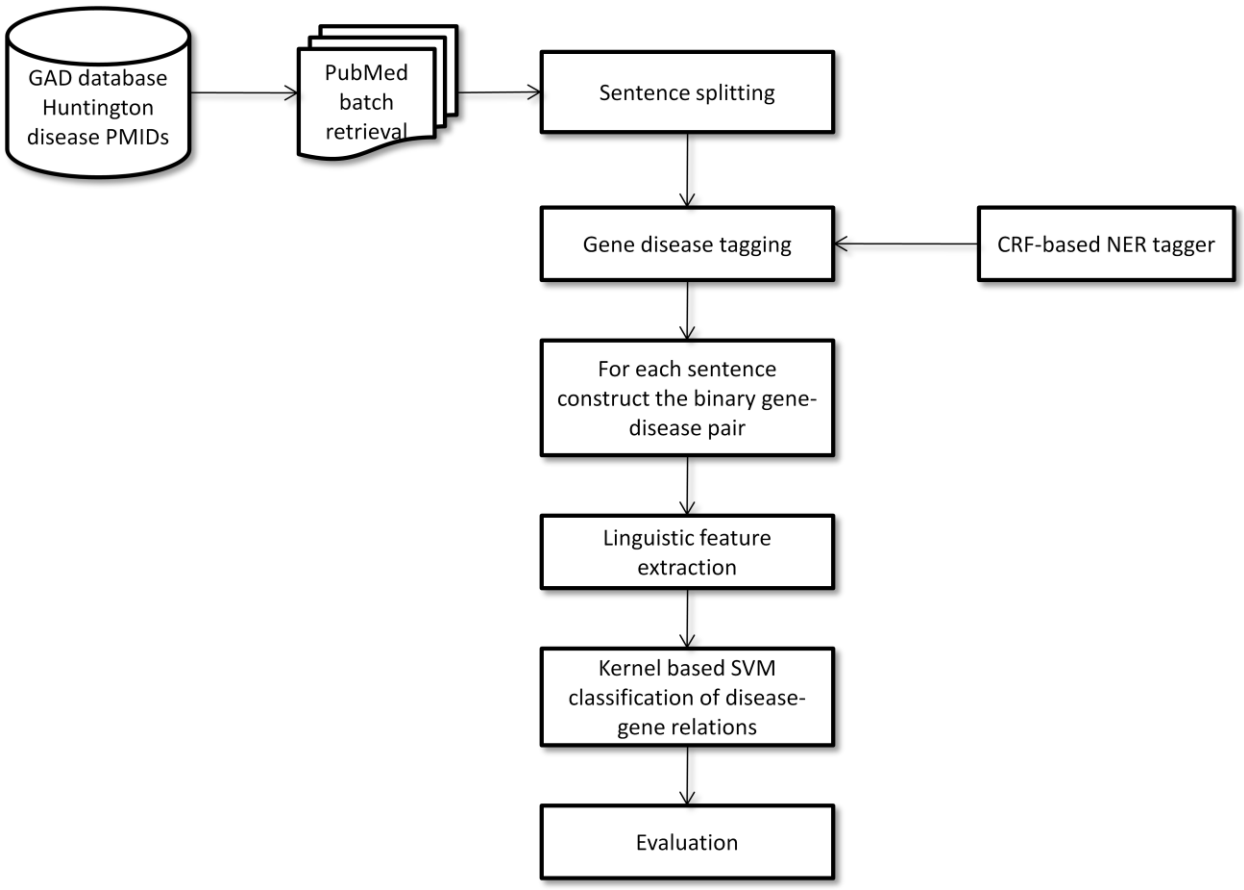


Figure 4-3. system architecture of kernel based Huntington disease-gene relation extraction system.

4.3.2. Kernel based SVM classifier for relation extraction

Kernel methods are used to map the input data into a high dimensional feature space so that linearly non-separable classes become separable by a linear algorithm. For our disease-gene classification problem, we used kernel functions implemented in JRSE package (Giuliano et al. 2006) shown below (4-3 to 4-8):

$$K(x_1, x_2) = \frac{(\emptyset(x_1), \emptyset(x_2))}{\|(\emptyset(x_1)\| \|\emptyset(x_2)\|} \quad (4\text{-}3)$$

The kernel is normalized by 2-norm of embedding vectors $\emptyset(x_1)$ and $\emptyset(x_2)$. It is based on string kernel using bag-of-features approach. Given two entities and an interaction relation shown in following examples, (Bunescu and Mooney 2006) found three patterns for words around related entities:

- Fore-Between (FB): relation is asserted using words before and between the entities. For example "interaction of **Entity_1** with **Entity_2**".
- Between (B): relation is asserted using words between entities. For example "**Entitye_1** is associated with **Entity_2**".
- Between-After (BA): relation is asserted using words between and after entities. For example "**Entity_1** and **Entity_2** interaction".

Formally, for the relation *R*, all three patterns (*P*) can be represented as a row vector:

$$\emptyset_P(R) = (tf(t_1, P), tf(t_2, P), , , tf(t_n, P)) \qquad (4\text{-}4)$$

where t_i $(1<i<n)$ is the token in the pattern and $tf(t_i,P)$ is its frequency of occurrence in pattern *P*. For all three patterns a kernel termed Global Context kernel K_{GC} is defined as:

$$K_{GC}(R1, R2) = K_{FB}(R1, R2) + K_B(R1, R2) + K_{BA}(R1, R2) \qquad (4\text{-}5)$$

where K_{FB}, K_B, and K_{BA} denotes kernels for Fore-Between, Between, and Between-After bag-of-words patterns based on 4-4 respectively.

It is observed in (Bunescu and Mooney 2006) above patterns use no more than 4 words to assert the relation. Therefore in our experiment for disease-gene relation classification, we used tri-grams contiguous tokens kernel.

In addition to above global context kernel K_{GC}, a Local Context kernel (LC) is define to take following four features of related entities into account.

- Token
- Lemma
- POS tag
- Orthographic (capitalization, punctuation, numerals)

The local context $LC = (t_{-w}, ..., t_{-1}, t_0, t_{+1}, ..., t_{+w})$ can be formally represented as a row vector:

$$\varphi_L(R) = (f_1(L), f_2(L), , , f_n(L)) \qquad (4\text{-}6)$$

For each feature at position L, the feature function f_i returns 1 if it is active, or 0 if otherwise. Here we used default window size of 1. The local context kernel K_{LC} for entity *E1* and *E2* is therefore defined as:

$$K_{LC}(R1, R2) = K_{E1}(R1, R2) + K_{E2}(R1, R2) \qquad (4\text{-}7)$$

Finally, the combo kernel K_{SL} combining both global and local context kernel is defined as:

$$K_{SL}(R1, R2) = K_{GC}(R1, R2) + K_{LC}(R1, R2) \qquad (4\text{-}8)$$

Table 4-3 summarized the kernels and its configuration used in our experiments.

Table 4-3. Kernels and configuration used in the experiments.

Kernel	Features	Configuration
Global context (GC)	Fore-between	Tri-gram
	Between	Tri-gram
	Between-after	Tri-gram
Local context (LC)	Token, lemma, POS, orthographic	Windows size = 1
Shallow Linguistic (SL)	GC + LC	Tri-gram, window = 1

All kernels in the toolkit are embedded into SVM package LIBSVM (Chang and Lin 2011) for model training and testing.

4.3.3. Evaluation of linguistic context based kernel method on AImed corpora

Before applying above kernel based classification methods on Huntington disease corpora, we evaluated their performance on the human protein-protein interaction corpora AImed. Table 4-4 shows the performance matrices (precision, recall, and F-measure) using 10-fold cross-validation. The results indicate global context kernel performs significantly better than local context kernel, and the combined kernel slightly increased the F-measure by 0.76%.

Table 4-4. Performance evaluation of three kernel based methods on human protein-protein interaction corpora AImed.

LC			GC			LC+GC		
Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
0.4424	0.7332	0.5493	0.6245	0.777	0.6912	0.6212	0.8014	0.6988

4.3.4. Disease-gene relation extraction from Huntington disease corpora

We applied the linguistic kernel based SVM classification on our Huntington disease corpora. Similar to the results in 4-3-3, table 4-5 shows global context kernel outperformed local context kernel in our binary disease-gene relation classification task, with significant increase of recall by 15.31% and F-measure by 9.34%. Compared with global context kernel, the combined kernel decreased the F-measure by 4.7% and recall by 7.93%. It suggests the most discriminative linguistic characteristics are largely contained in tri-grams global context before, between, and after two related entities in our annotated corpora.

Table 4-5. Kernel based disease-gene classification using annotated Huntington disease corpora.

LC			GC			LC+GC		
Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
0.9621	0.7205	0.8211	0.9623	0.8736	0.9145	0.9614	0.7943	0.8675

4.4. Conclusion and future work

In contrast to full syntactic tree parsing, shallow linguistic parsing computes the basic text structure using bag-of-features approach. Our preliminary results obtained using shallow linguistic kernel methods on an annotated Huntington disease corpora suggest the global tri-grams context surrounding related entities are critical for disease-gene relation extraction, which is in agreement with PPI relation extraction evaluation using AImed corpora. It is noted however, due to limited Huntington disease PubMed citations from GAD our annotated dataset is relatively small, which will likely miss some complicated sentences in real-world. Therefore for future work it is necessary to increase the corpora size by adding new PubMed abstracts referenced by other gene-disease relation databases, for example OMIM (NCBI). A similar corpora for hypertension gene relation extraction was constructed from GAD in (Tsai et al. 2009) with total 939 annotated sentences. It may not be surprising as comparing with 203864 abstracts returned by the PubMed query using "hypertention" as MeSH term, only 8843 abstracts were returned by PubMed query using "Huntington diesase" as MeSH term. An alternative way of expanding the less commonly seen diseases corpora is to use phrases in GeneRIF database as shown in work (Bundschus Markus et al. 2008), in which the authors extracted 5720 phrases with gene and disease associations.

In conclusion, in this chapter we exploited the linguistic kernel based machine learning approach in extracting relations between disease and gene. Our results suggest bag-of-features kernel-based SVM classification is a promising resolution for specific disease-gene association mining. With future expansion of the training corpora, it can be applied on real-world problem for known disease associated gene extraction and novel gene prediction.

CHAPTER 5. MINING DISEASE ASSOCIATED GENES USING INFORMATION EXTRACTION AND GRAPH THEORETIC APPROACHES

5.1 Introduction

In chapter 4 we focused on relation extraction of disease and disease associated genes from literature. The machine learning based approach has several advantages including robustness in processing noise dataset, but its application is largely limited by the availability of training collection and its quality and quantity. To overcome the problem of requiring annotated corpora for relation extraction in specific domain, in this chapter we present our novel strategy on concept co-occurrence based approach for disease and its associated gene extraction.

With completion of human genome project, thousands of genes have been identified to be linked with variety of human diseases. Online Mendelian Inheritance in Man (OMIM) (NCBI), and Genetic Association Database (GAD) (Kevin Becker, Kathleaen Barnes, Tiffani Bright 2004), among many other online databases, have been utilized extensively to aid discovery of new genetic factors leading to diseases. New findings ranging from basic research to clinical reports have been constantly published and indexed by PubMed. Biomedical researchers are increasingly depending on gathering published data of their interest to formulate research hypothesis before conducting basic and clinical study. Given millions of published papers deposited in PubMed, it is not surprising one can easily be overwhelmed. Information extraction

as a subject of broad information retrieval plays an important role in finding hidden relational information from unstructured biomedical text, for instance in identifying genetic roles in human disease. Large database such as OMIM, Entrez Gene (NCBI), and GAD contains validated links between genes and diseases with reference to literatures, but due to laborious manual curation process, novel findings especially those related to newly identified genes in more recent published papers are often not included.

Information extraction by text mining in biomedical domain is one of the hot spots in natural language processing community, and is also regarded as a more difficult task than general information retrieval as it not only requires retrieval of relevant documents containing the information, but also requires structured retrieval of entities, relationships, and their associated attributes. For the past decades several high impact methods were proposed to address the difficulty. Rule-based methods involve manual creation of rules and patterns but has major limitation on covering many variations in large unstructured and often noisy corpus. Statistical model based approaches overcome the issue by introducing Hidden Markov (HM), Maximum Entropy (ME), and Conditional Random Fields (CRF) models. Rule-based methods are generally easier to interpret while statistical model based methods are generally more robust to noisy unstructured text. Depending on the context of information extraction needs, above methods are commonly used in parallel or integrated way.

Extraction of relations from collections of documents follows two different ways. Given a relation between two entities *r{e1, e2}* where r denotes relation type and *e1/e2* denote two different entities, the rule based methods and statistical model based methods attempt to match or predict the relation pair by using rich set of local and global linguistic features. This is normally

done at sentence level and we have discussed it in more detail in chapter 4. However it doesn't take into consideration that the two entities may be correlated in different sentences at paragraph level or abstract level. Taken an example of sentences from PubMed (23341638), where gene mentions "AP-1" and "SOX2" are located at different sentence than the disease "Huntington disease" mentioning sentence.

*"On the basis of the sequence of regions that change in methylation, we identify **AP-1** and **SOX2** as transcriptional regulators associated with DNA methylation changes, and we confirm these hypotheses using genome-wide chromatin immunoprecipitation sequencing (ChIP-Seq). Our findings suggest new mechanisms for the effects of polyglutamine-expanded HTT. These results also raise important questions about the potential effects of changes in DNA methylation on neurogenesis and cognitive decline in patients with **Huntington disease**." (PubMed 23341638)*

To accommodate such scenario, a deep syntactic parse tree across sentences is often needed to address co-reference concerns. However it is computationally prohibitive for large scale text mining task. Another approach to this problem is to utilize term co-occurrence counts in 'bag of words' way at abstract level, or at specified word window, to determine whether they are significantly co-related. Extracted terms are then ranked for researchers to make informed decision and novel hypothesis formulation.

In this chapter, we are concerned with extraction of disease associated genes from literature and prioritize them using graph theory methods. We will propose an integrated text mining and graph analysis approach to identify disease associated genes, using Huntington's disease as a case study. We first prepared a corpora by querying the PubMed with MeSH terms

of the specific disease to collect all documents that use the disease MeSH term as their major topic. The text collection is indexed by UMLS metathesaurus concepts and co-occurrence matrix are constructed. The extracted concepts are further disambiguated to construct gene-gene and disease-gene network. We then compiled a list of known disease associated genes from GAD database as initial seed genes associated with the disease and expand them to a large gene interaction network. The literature mined interaction network was then merged with the seed gene expanded network to form a heterogeneous disease-gene network for network analysis using graph theory. By using information extraction and network analysis methods, we intended to extract list of disease associated genes and prioritize them.

The rest of the chapter is organized as follows. An overview of related works on co-occurrence based information extraction and biological network analysis is given in 5.2, followed by description of our gene-disease association extraction system architecture and methods in 5.3. The results is discussed in 5.4, and conclusion and future work are given in section 5.5.

5.2 Related works

Gene-disease association extraction has been attracting much attention in recent years since the completion of human genome project and with the rapid development of proteomic technology. In (Adamic et al. 2002) a method to extract gene sets relevant to query of disease was presented using statistical analysis of gene-disease co-occurrence. Intuitively, if a gene mention occurs at the same frequency in a small disease-focused document collection as in a large non-disease focused document collection, based on normal approximation to the binomial distribution one would conclude the gene is not statistically associated with the disease. By using

a large number of PubMed collection (3 million abstracts) the authors could rank genes that are statistically correlated with leukemia and breast cancer. In (Singh and K 2005) a statistical method using term frequency and document frequency for co-occurrence of the diseases and proteins was presented. Similar to (Adamic et al. 2002), this study also used both positive collection (specific to disease of interest) and negative collection, but the negative collection is considerably smaller (40,000-45,000). In another study term co-occurrence based method has also been used together with rule-based association relation mining to rank gene-gene, gene-disease associations in (Cheng et al. 2008).

Gene network study under graph theory framework has been widely applied in finding genetic linkage of variety of phenotypes (Mason and Verwoerd 2007b)(Lu et al. 2011)(Lage et al. 2007). A general overview on information extraction using gene interaction network was given in chapter 2 literature review. In (Gonzalez et al. 2007) the authors first obtained a list of atherosclerosis associated genes or gene products from CBioC (Collaborative Bio Curation) database, which contains automatically extracted facts including protein-protein interaction, gene-disease and gene-bioprocess relations and their accuracy were rated by a social network of biomedical researchers. The variants of gene names were then normalized to HUGO nomenclature (HUGO Gene Nomenclature Committee) and used as initial set to expand the network by nearest-neighbor algorithm using CBioC dataset. The extended set of genes were ranked by a heuristic scoring algorithm to predict the most likely disease associated genes. This approach integrated human annotation with the automated text mined gene-disease association relations. However their database are not publicly available. In another study (Chen and Sivachenko 2006) the authors created initial set of 70 Alzheimer's disease associated genes from OMIM (NCBI) and HUGO database. The set was extended by nearest-neighbor expansion to

construct a network consisting of 657 human proteins and 775 interactions using OPHID database which contains total ~9000 human proteins and ~40,000 interactions (Brown and Jurisica 2005). The authors then conducted statistical analysis to find all significant sub-networks that form higher connectivity among protein nodes than those randomly selected protein nodes from OPHID. Proteins in sub-network are scored using a heuristic relevance score function that takes into consideration of their overall role in the network and contribution to the sub-network. However the interactions among neighbors are not considered thus leads to bias towards the seed genes in their final ranked gene list.

In addition to heuristic scoring methods, centrality measurement of gene network is another important way to rank the node in the graph. Several studies have successfully applied degree, betweenness, and essentiality measurements from graph theory to rank importance of the genes in the interaction network (Joy et al. 2005)(Jeong et al. 2001)(Goh et al. 2007). In (Ozgür et al. 2008) a prostate cancer specific gene interaction network was built around a list of seed genes known to be related to the disease. Instead of expanding the initial gene set using protein-protein interaction dataset from database, the genes associated with the initial gene set were mined from literature by syntactic dependency tree parsing and SVM classification. Specifically, sentences were filtered using a set of manually created interaction words to retain those sentences containing at least one seed gene and an interaction word. Syntactic dependency parse tree was applied to each sentence to extract the shortest path between two gene pairs. The similarity between extracted paths was measured by word-edit distance and was used as SVM kernel function to train a classification model on two golden standard corpus. The trained system was then applied on new sentences from PubMedCentral database to build gene interaction network related to prostate cancer. The text mined gene network was analyzed by node centrality

in the graph. Their results suggest that betweenness, eigenvector, and degree centrality perform best in ranking top 10 and 20 genes associated with prostate cancer.

In addition to homogeneous gene-gene or protein-protein interaction (PPI) networks, recently a holistic view of phenotype-gene interaction was also proposed to study the molecular mechanisms of common human disease. In this heterogeneous phenotype-gene network that combines both genetically similar phenotypes with their associated genes, it is possible to infer candidate disease associated genes by network topology measurement taking into consideration of phenotype-phenotype, phenotype-gene, and gene-gene associations (Yao et al. 2011). In (Lage et al. 2007)(Wu et al. 2008)(Lee et al. 2011) human disease associated candidate genes are inferred from the heterogeneous phenotype-gene network by their topological closeness to the disease based on the assumption that phenotypically similar human diseases are likely caused by functionally related genes.

5.3 Experiments and results

5.3.1 System design architecture

Figure 5-1 shows the system architecture of mining disease-associated genes from literature and ranking them based on network analysis. Our assumption is that the co-occurrence of the gene with the disease in the literature indicates its association likelihood with the disease, even though the association type may vary. The system integrated four major steps to 1) collect disease focused corpora for concept extraction and indexing, 2) perform union operation on gene co-concepts, followed by gene name disambiguation and normalization, 3) create initial seed of known disease-associated genes and expand seed genes to construct gene interaction network for the given disease, 4) build disease-gene heterogeneous network based on literature mined and

seed gene extended interaction network for network analysis and candidate disease-associated gene ranking.

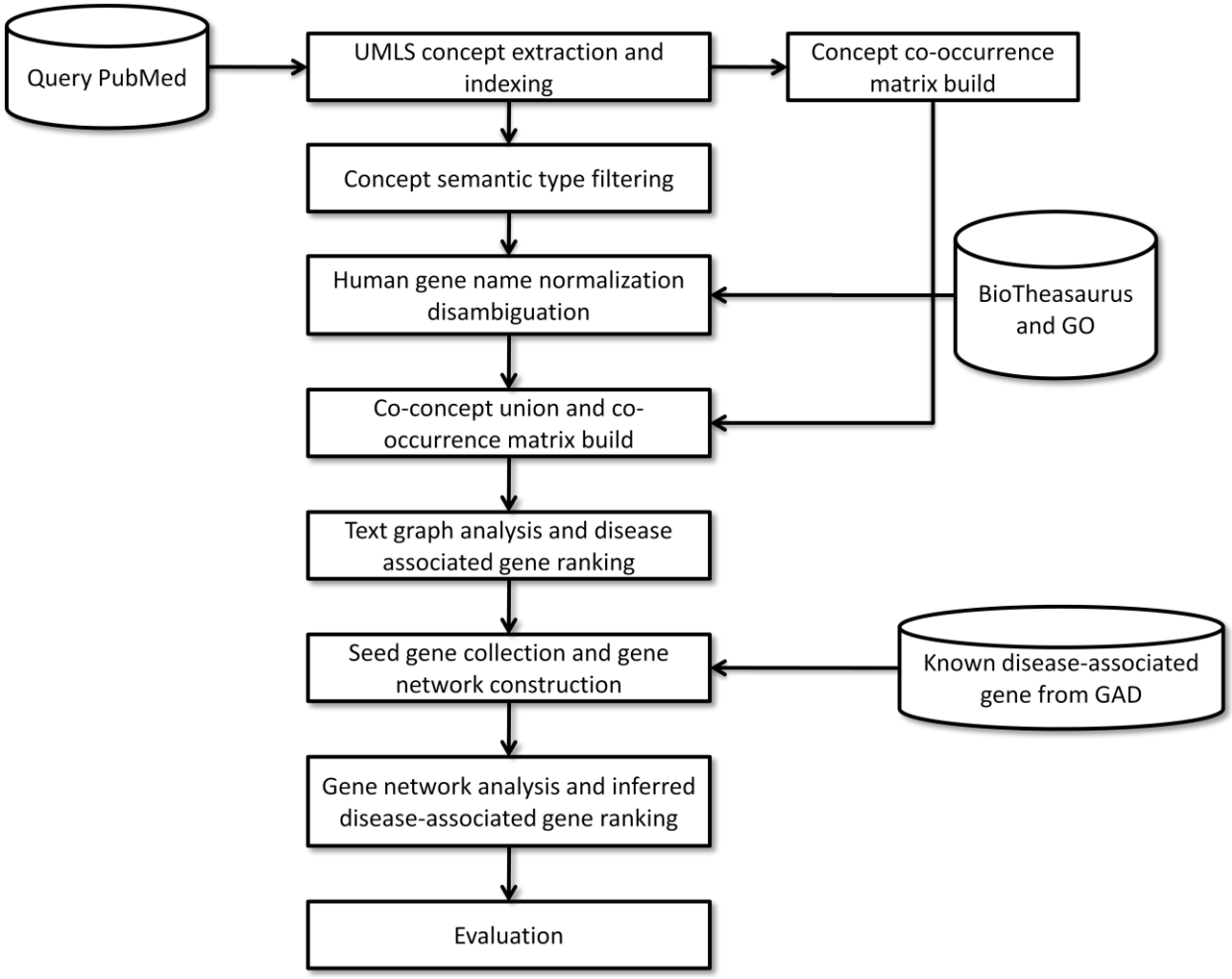


Figure 5-1. System architecture for disease-associated gene mining.

5.3.2. Corpora preparation and indexing

We continued to choose Huntington disease as case study in our experiment. To prepare the corpora related to this disease, we queried PubMed literature database, currently containing more than 23 million citations for biomedical literatures, using MeSH term of the disease. MeSH (Medical Subject Headings) (NCBI) terms is set of controlled vocabulary thesaurus for indexing

PubMed abstracts and it allows us to collect abstracts with the disease as their major topic. Total 8843 abstracts concerning Huntington disease were downloaded from PubMed using Entrez eUtil web service. We used Dragon toolkit (Zhou et al. 2007) to extract and index biomedical concepts from document collection using UMLS meta-thesaurus. UMLS meta-thesaurus is one of the most comprehensive meta-thesaurus containing millions of biomedical and health related concepts, synonymous names, and their relations organized in a semantic network. Therefore we reasoned it is the most suitable resource from which the biomedical related concepts can be extracted. However, the limitation of concept based disease-associated gene extraction is that it is often ambiguous and redundant due to variations of biomedical named entities, which will introduce noise and result in high dimensionality problem. To address this concern, we performed co-concept union and disambiguation which we will discuss later. As mentioned earlier in this chapter, the concept extraction and indexing was performed at abstract level instead of sentence level to avoid losing any gene mention that are not co-localized with the disease mention in the same sentence, but do carry the information of disease association relation. The concept co-occurrence sparse matrix was build using the dragon toolkit during indexing process.

5.3.3. Semantic context analysis of Huntington disease

We analyzed all concepts with semantic type "Disease or Syndrome" (UMLS TUI: T047) extracted from the corpora. The Huntington disease concept were extracted as UMLS concept C0020179 for Huntington Chorea (HD) and C0751208 for Juvenile Huntington Disease. According to wikipedia, "*HD is the most common genetic cause of abnormal involuntary writhing movements called chorea, which is why the disease used to be called Huntington's*

chorea". Juvenile Huntington Disease is an early onset HD at age before 20 and accounts for 5-10% of HD cases (Warby, Simon C, Rona K Graham 2010). All disease/syndrome concepts co-occurrence counts were normalized against Huntington Chorea (C0751208). Figure 5-2 shows the top 25 correlated disease/syndromes with C0751208 (HD), indicating its close association with syndromes including motor disorder, undernutrition, senile dementia, movement disorder, dystonia disorder, circadian dysregulation, late-onset disorder and early disease onset, gastric motor dysfunction, dyskinesia, and Parkinson like disorders. The disease is also closely related to other neurodegenerative diseases including Parkinson disease and Gehrig disease. The extensive spectrum of disease/syndrome concepts gave the semantic context annotation for the disease from another angle.

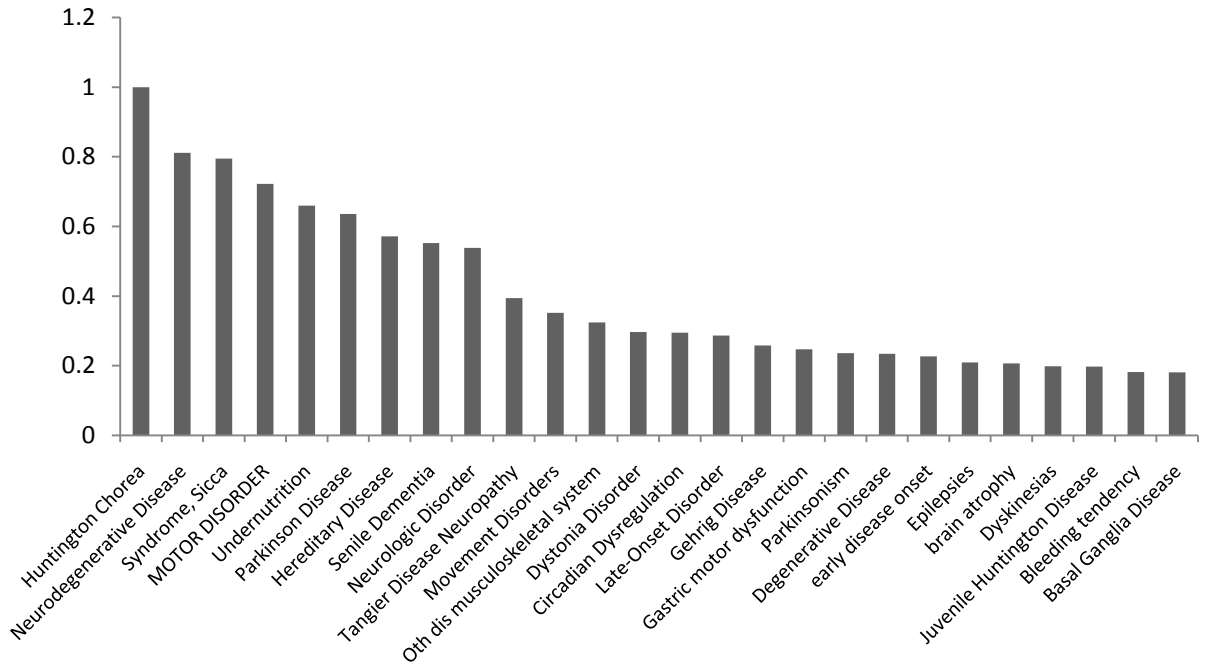


Figure 5-2. Disease and symptom concept co-occurrence pattern. Co-occurrence (y-axis) was normalized against Huntington Chorea (C0020179). Two concepts for Huntington

disease (Huntington Chorea C0020179 and Juvenile Huntington Disease C0751208) were extracted from document collection.

5.3.4. Co-concept union and human gene name normalization

We first filtered the extracted concepts to include all possible genes and gene products using selected UMLS semantic types. Considering the ambiguity in the UMLS meta-thesaurus (Lang et al. 2009) we expanded semantic types to include not only gene concept, but also nucleotide, molecular sequences, peptide, enzyme, receptor and protein related concepts. Table 5-1 shows the expanded semantic types used for filtering. Although non gene or gene products will be introduced with the expanded semantic type, further gene name disambiguation step will exclude them from the final human gene list.

Table 5-1. Expanded UMLS semantic types related to gene and gene products used for concept semantic filtering.

Semantic type group	Category	TUI	Description
GENE	**Genes & Molecular Sequences**	**T028**	**Gene or Genome**
GENE	Genes & Molecular Sequences	T087	Amino Acid Sequence
GENE	Genes & Molecular Sequences	T088	Carbohydrate Sequence
GENE	Genes & Molecular Sequences	T085	Molecular Sequence
GENE	Genes & Molecular Sequences	T086	Nucleotide Sequence
CHEM	Chemicals & Drugs	T192	Receptor
CHEM	Chemicals & Drugs	T116	Amino Acid, Peptide, or Protein
CHEM	Chemicals & Drugs	T126	Enzyme
CHEM	Chemicals & Drugs	T125	Hormone

CHEM	Chemicals & Drugs	T129	Immunologic Factor
CHEM	Chemicals & Drugs	T114	Nucleic Acid, Nucleoside, or Nucleotide

Gene mention normalization is an important step to map gene names and their variants to unique and officially approved symbols. For example, human GABBR1 gamma-aminobutyric acid (GABA) B receptor is also known as GB1, GPRC3A, GABABR1, and GABBR1-3. It can be normalized to Entrez Gene id 2550 referencing to its official name and all synonyms. If a gene name is mapped to more than one unique database entry, a further disambiguation step is required. Dictionary based and machine learning based methods are two commonly used approaches for gene mention disambiguation. Many studies have been devoted to this area of study. In order to bring community effort in this research area, BioCreative II has organized gene normalization (GN) task in conjunction with the gene mentioning (GM) task since 2006 (Hirschman et al. 2005a).

We utilized the dictionary based method implemented in Moara java toolkit which stores BioThesaurus and Gene Ontology (GO) (Gene Ontology Consortium 2001) in a local MySQL database to normalize and disambiguate gene names (Neves et al. 2010). The BioThesaurus (Liu et al. 2006) is a comprehensive collection of protein and gene names with more than 2.8 million names extracted from different databases using cross-references provided by iProClass (Wu et al. 2004). The disambiguation method takes input of the extracted gene mention and its text context, searches BioTheasaurus for the exact match. If more than one match is found it will generate an representative document for the gene in question using information from Entrez Gene and Gene Ontology (GO) databases to compare (1) the cosine document similarity between the text context

and the generated gene representative document, (2) the common tokens between the two documents, or (3) both document cosine similarity and common tokens between two documents. Cosine similarity is a well established document similarity measurement in IR based on vector space model, in which text is modeled by a vector of terms. Formally, it is represented as:

$$Sim\left(\underset{D}{\rightarrow},\underset{Q}{\rightarrow}\right) = \frac{\vec{D}\cdot\vec{Q}}{\left\|\vec{D}\right\|\left\|\vec{Q}\right\|} = \frac{\sum_{i=1}^{n} D_i \times Q_i}{\sqrt{\sum_{i=1}^{n}(D_i)^2} \times \sqrt{\sum_{i=1}^{n}(Q_i)^2}} \qquad (5\text{-}1)$$

where $\underset{D}{\rightarrow}\ and\ \underset{Q}{\rightarrow}$ denotes the document and query vector respectively, $\underset{D}{\rightarrow}.\underset{Q}{\rightarrow}$ denotes their dot product, $\left\|\underset{D}{\rightarrow}\right\|\ and\ \left\|\underset{Q}{\rightarrow}\right\|$ denotes the length (or norm) of the document and query vectors.

The information used for representative gene document include gene aliases, symbols, description and summary, phenotypes, relations etc, all stored in local MySQL database. Using aforementioned disambiguation step, the predicted gene with highest score is selected as the normalized gene.

Total 15654 UMLS concepts were extracted from 8843 abstracts. After semantic type filtering and gene normalization step, 3416 human genes and gene products were mapped to unique Entrez Gene id. Among normalized human genes, total 336 were found to be associated with more than one UMLS concepts. For example, 6 extracted UMLS concepts with CUI C0806318 (HD gene.CAG repeats), C0252274 (HD protein, human), C0872189 (Huntington gene), C0872190 (Huntington protein), C0247953 (IT15 gene product, human), and C1415504 (HTT gene) were normalized to the Entrez gene identifier 3064 (HD gene or HTT).

5.3.5. Text graph and disease-associated genes extraction

Driven by big data collected from large scale genome-wide gene expression measurement (microarray) and proteins/metabolites identification (mass spectrometry), graph theory has been applied on biomedical network study widely in recent years. For a literature overview of graph theory and biological networks please refer to chapter 2. In this section we focus on a special type of graph network, e.g. text graph. It has been well exploited in linguistic studies that correlation of words, terms, concepts can be represented and modeled as text graph, where vertices denote words or terms, and edges denote co-occurrence, syntactic, semantic, or orthographic relations between terms (Blanco and Lioma 2011). Early work conducted in (Minsky 1968) applied graph theoretic approaches on information retrieval (IR) and follow up studies have extended it to web search (Lawrence Page) and variety of IR applications to improve the retrieval performance.

We extracted gene-related biomedical concepts from literature and further disambiguated and mapped them to human genes. We next analyzed the correlation data between gene-gene and gene-disease in an aim to identify candidate HD associated genes. The correlation networks constructed in 5.3.4 are based on concept co-occurrence and no gene or protein expression data from external database is involved, therefore is treated as an undirected text graph, where nodes are normalized genes or Huntington disease, edges are their interactions and the co-occurrence counts are edge weights. Formally, the graph is defined as undirected graph $G\{V,E\}$, where G denotes the graph with set of vertices V, and set of edges E between pair of vertices u and v $(u, v \in V)$. Undirected graph G can be represented as an edgelist containing all edges in the graph, where $uv \in E$ and $u, v \in V$. If the connection between two vertex u and v is either 0 (no connection) or 1 (connected), the graph is an un-weighted binary graph. Otherwise it is a weighted graph. Another way of network representation is a n x n symmetric matrix with vertices

as rows and columns, their binary relation or weighted relation as cell values. Equation 5.2 shows a symmetric binary adjacency matrix A containing element A_{ij} for un-weighted binary graph.

$$Aij = \begin{cases} 1 & if\ vertex\ i\ and\ j\ are\ connected \\ 0 & otherwise \end{cases} \qquad (5.2)$$

In equation 5.3 for a weighted undirected graph, A_{ij} equals to the connection strength between pair of vertices:

$$A_{ij} = weight\ of\ connection\ between\ vertex\ i\ and\ j \qquad (5.3)$$

The topological properties of network was pioneered by works in (P. Erdos 1960) on random graph and was later generalized to non random graphs. Random graph, by its name, is randomly generated graph for study under graph theory and probability theory. Major undirected and un-weighted graph topological properties used in this chapter are formally defined below:

1). Degree distribution

For each node $u \in V$, the degree is defined as number of edges connected to u. The degree distribution measures the probability of vertex u, e.g. $P(u)$ having degree of k.

2). Average cluster coefficient distribution

The cluster coefficient C_u of node u is defined as:

$$c_u = \frac{2e_u}{k_u(k_u - 1)} \qquad (5.4)$$

where k_u is the number of neighbors of node u, e_u is the number of connected pairs between all neighbors of u (Watts and Strogatz 1998). It measures the tendency of nodes

clustering together. The average cluster coefficient distribution measures the average cluster coefficient of all nodes in the graph with *k* neighbors *(k=2...n)*.

3). Neighborhood connectivity distribution

The neighborhood connectivity of node u is the average connectivity of all neighbors of *u* and connectivity is defined as number of its neighbors (Maslov and Sneppen 2002). The neighborhood connectivity distribution measures the average of the neighborhood connectivity of all nodes with k neighbors *(k = 0, 1, ... n)*.

4). Topological coefficient

The topological coefficient T_u of node u is defined as:

$$Tu = \frac{avg\,(J\,(u,m))}{k_u} \qquad (5.5)$$

where k_u is neighbors of node *u* and *J(u,m)* is the number of neighbors that shared between node *u* and *m* (Stelzl et al. 2005). It is a measurement of extend to which a node shares neighbors with others in the network.

5). Closeness centrality and its distribution

The closeness centrality of node *u* is defined as:

$$Cc(u) = \frac{1}{avg\,(L(u,m))} \qquad (5.6)$$

where *L(u,m)* denotes the length of the shortest path between node *u* and *m* (Newman 2003). The shortest path is also used to compute the network diameter which equals to the maximum length

of shortest paths between two nodes in the graph. The closeness centrality of all nodes against its neighbors is given as the network closeness centrality distribution.

6). Betweenness centrality and its distribution

The betweenness centrality of node *u* is defined as:

$$C_b(u) = \sum_{s \neq u \neq t} \left(\frac{\sigma_\pi(u)}{\sigma_\pi} \right) \qquad (5.7)$$

s, *u*, and *t* are nodes ($s \neq u \neq t$). σ_π denotes the number of shortest paths between *s* and *t*, and $\sigma_\pi(u)$ denotes number of shortest paths between *s* and *t* that *u* lies on (Brandes 2001). Compared to the global connectivity measurement, it is an more important local centrality measurement for a network node that equals to the number of shortest paths between all nodes that pass through the node. Similar to closeness centrality distribution, we plot the betweenness centrality distribution as betweenness centrality of all nodes against its neighbors.

We developed BasicGraphCreation program using JUNG graph toolkit to construct the initial correlation network. The pseudo code is shown in figure 5.3. Figure 5.4 shows the generated correlation network with 3416 vertices and 47892 edges.

For networks generated by Cytoscape, we used two widely cited plug-ins, Network Analyzer (Assenov et al. 2008) and CentiScaPe (Scardoni et al. 2009), to compute the network parameters. For network generated by JUNG Java toolkit, we computed network parameters using its scoring algorithms and R package.

```
Add all normalized human genes as vertices v={g1, g2, ..., gn}
For each gene pair gp belong to v
        Get co-occurrence count ct_gp (as edge weight)
        If (ct_gp) > 0
                Add edge Egp to the graph
End forloop
Display network using selected layout algorithm
```

Figure 5-3. Pseudo code for building initial gene correlation network.

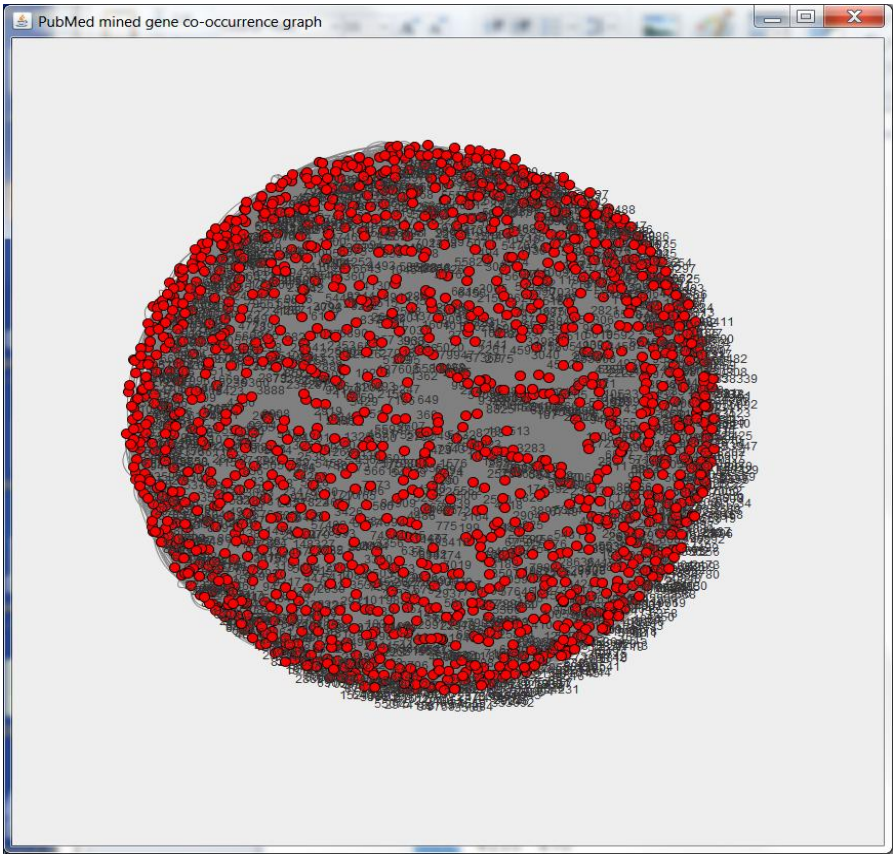


Figure 5-4. The dense sub network of gene correlations. Each gene vertex is labeled with its Entrez gene id.

Since we are interested in prioritizing genes that most likely associated with the disease, we filtered the network by including only the top n gene correlations ranked by their weights, e.g. co-occurrence between gene pairs. The resulting network consists of 310 connected vertices and was imported into Cytoscape (Shannon et al. 2003) for enhanced visualization. Figure 5-5-A shows that most genes are connected to the central hub gene (Entrez id 3064: HD gene or HTT) and the highly connected gene cluster is shown in 5-5-B.

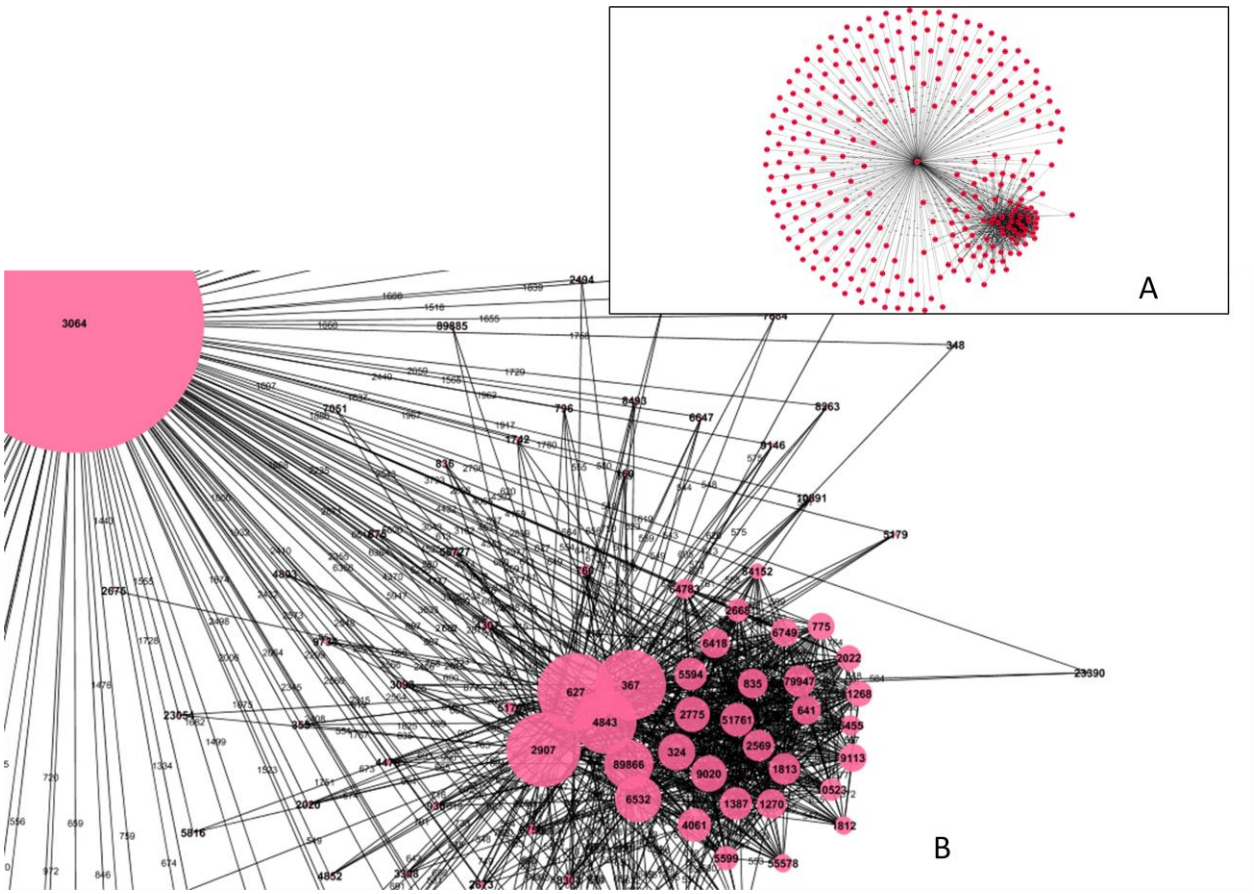


Figure 5-5. A) the network consisting of top n gene correlations. B) the highly connected gene cluster. The node is labeled with the Entrez gene id and the node size is mapped to its degree.

After filtering network using its co-occurrence weight, we transformed the network to a undirected un-weighted binary graph. We computed the degree centrality, closeness centrality, and betweenness centrality for each gene and the top 20 ranked human genes are listed in table 5-2.

Table 5-2. The top 20 human genes ranked by degree, closeness, and betweeness centrality. The number in the parenthesis indicates Entrez gene unique id.

Rank	Degree centrality	Closeness centrality	Betweeness centrality
1	HTT (3064)	HTT (3064)	HTT (3064)
2	BDNF (627)	APC (324)	BDNF (627)
3	GRINA (2907)	SLC6A4 (6532)	GRINA (2907)
4	AR (367)	CREBBP (1387)	AR (367)
5	NOS2 (4843)	SEC16B (89866)	NOS2 (4843)
6	APC (324)	GABRR1 (2569)	SEC16B (89866)
7	SLC6A4 (6532)	MAP3K14 (9020)	SLC6A4 (6532)
8	CREBBP (1387)	MAPK1 (5594)	APC (324)
9	MAP3K14 (9020)	TS (775)	MAP3K14 (9020)
10	SEC16B (89866)	SET (6418)	MAPK1 (5594)
11	GNAO1 (2775)	FACT (6749)	LY6E (4061)
12	GABRR1 (2569)	CASP2 (835)	GNAO1 (2775)
13	LY6E (4061)	ENG (2022)	CREBBP (1387)
14	MAPK1 (5594)	ATP8A2 (51761)	DHDDS (79947)
15	WTS (9113)	BLM (641)	GABRR1 (2569)
16	CASP2 (835)	CHERP (10523)	ATP8A2 (51761)

17	DRD2 (1813)	PPP1R1B (84152)	DRD2 (1813)
18	FACT (6749)	DRD1 (1812)	CASP2 (835)
19	TS (775)	JNK (5599)	BLM (641)
20	ATP8A2 (51761)	OTT (64783)	CNTF (1270)

Several studies have taken the phenotype into gene network construction (Yao et al. 2011)(Wu et al. 2008)(Lage et al. 2007)(Köhler et al. 2008) which leads to a heterogenic gene-phenotype network. It provided a new way to study gene-gene interaction, phenotype-gene interaction (disease-gene interaction), and phenotype-phenotype interaction in the same heterogeneous network. To study the Huntington disease and gene correlation, we added HD disease to form a literature mined heterogeneous disease-gene network. Figure 5-5 shows the heterogeneous network for HD disease and its correlated genes. The network is filtered using edge weight cutoff of 500 (A) and 1000 (B) respectively. It shows a network with clustering coefficient of 0.31, which is significantly higher than the corresponding random network of 0.0138, indicating its non-random network characteristics. The network analysis is given in figure 5-7. The degree distribution log-log plot (5-7-A) shows a scale-free network following a weak power law of degree distribution which fits function $y = 15.388x^{-0.696}$ (correlation = 0.778, R-squared = 0.564). Further analysis on the average clustering coefficient distribution (figure 5-7-B) suggests a highly clustered network with small number of neighborhood and then follows a power law cluster coefficient distribution with larger number of neighbors. The neighborhood connectivity distribution (figure 5-7-C) shows a decreasing function of node neighbors, suggesting edges between low connected and high connected nodes dominate the network. The topological coefficient distribution (figure 5-7-D) is the tendency measurement for

the node in our text graph to have shared neighbors, which indicates a power law decreasing of tendency for nodes with large number of neighbors. Two centrality distributions (figure 5-7-E and figure 5-7-F) suggest that few nodes with high centrality interconnect with majority of low centrality nodes. Among the high centrality nodes, the HTT (HD gene) forms the central hub for the network as indicated in table 5-2 and figure 5-4. Indeed, previous studies have demonstrated HTT plays the pivotal role in Huntington disease and the HTT-interactome, e.g. the interaction network around HTT, has been extensively studied using proteomic approaches (Shirasaki et al. 2012).

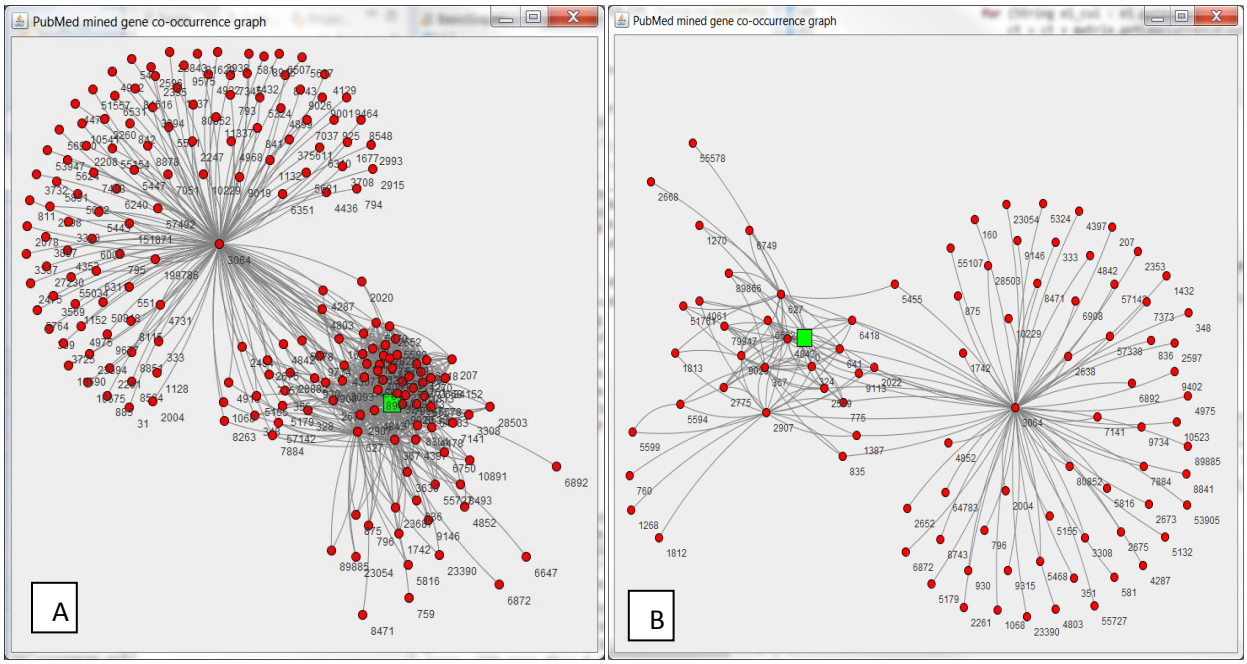


Figure 5-6. the heterogeneous disease-gene correlation network with edge weight cutoff of 500 (A) and 1000 (B). The red circle indicates the human genes and green rectangle indicates the HD disease. Entrez gene id is used to label gene node, "0" is used as the HD disease label. The network for A contains 292 node, 1243 edges, with cluster coefficient of 0.31, diameter of 2, average path length of 1.971, and average number of neighbors of 8.514.

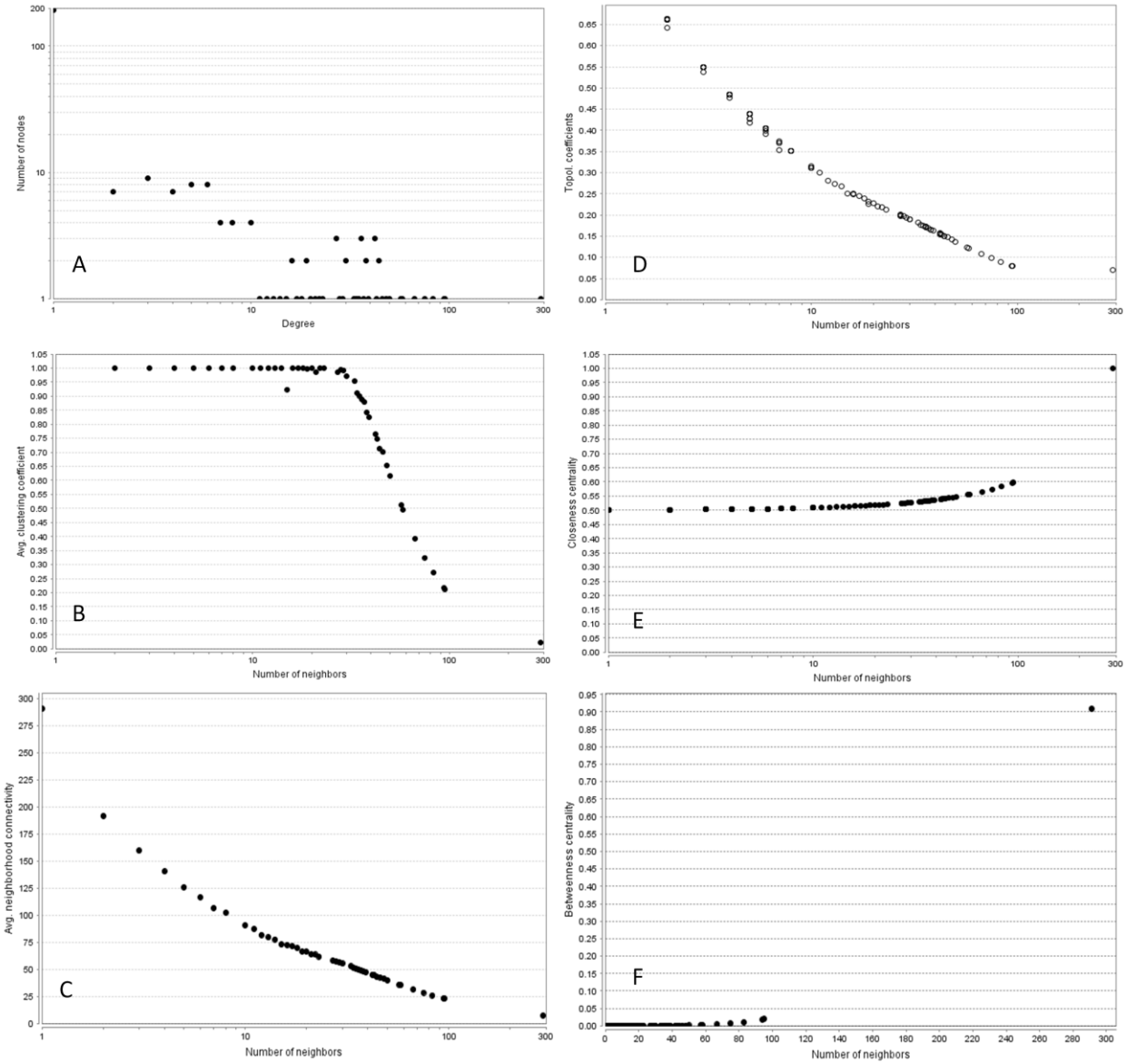


Figure 5-7. network analysis for degree distribution (A), average clustering coefficient distribution (B), neighborhood connectivity distribution (C), topological coefficients (D), closeness centrality (E), and betweenness centrality (F).

5.3.6. HD disease associated gene network construction using seed genes

To construct a gene network related to Huntington disease using empirical protein-protein interaction dataset, we started by collecting known disease-associated genes from

Genetic Association Database (GAD) (Kevin Becker, Kathleaen Barnes, Tiffani Bright 2004) and use them as seed genes. The GAD is a public repository collecting genetic associated data related to human diseases and is manually curated by National Institute on Aging. Total 41 genes associated with Huntington disease were compiled as seed genes to build gene interaction network using protein-protein interaction data from NCBI Entrez database (Table 5-3). The resulting network shown in figure 5-8 is a scale-free network with 1890 nodes and diameter of 6, cluster coefficient of 0.085, characteristic path length of 3.538, and average number of neighbors of 2.599. Further network analysis shown in figure 5-9 indicates the network follows a strong power-law degree distribution. The network topology follows a similar pattern with our text mined network but with more diverse distributions.

We analyzed the concept extraction and normalization for the 41 seed genes from GAD. Among them, 36 (87.8%) were extracted by our conceptual based information extraction and normalized to Entrez gene id that were validated manually. Table 5-3 shows the 41 seed genes and their extracted corresponding UMLS concepts and normalized Entrez id/symbols. 5 genes that do not have their corresponding UMLS concepts extracted were not normalized. It could be due to lack of related abstracts in our PubMed collection, lack of UMLS concept coverage, or not being extracted during text mining process. All 41 genes, however, were used for extended network construction detailed below.

Table 5-3. List of 41 seed genes compiled from GAD database and their corresponding UMLS concept CUI. Concepts and symbols marked with * were not extracted or normalized during text mining process.

Seed genes	Entrez id	Normalized Entrez symbol	Extracted UMLS CUI

ADORA2A	28882	2 a a	C0255998
APOE	348	apoe	C0003595, C1412481, C1370077
ATG16L1	55054	*	*
ATG3	64422	*	*
ATG5	9474	5 atg	C0531514
ATG7	10533	7 atg	C0758241
ATN1	1822	drpla	C1414155
ATXN1	6310	1 ataxin	C0380755, C0297041, C0807868, C1419828
BDNF	627	bdnf	C0084873, C1332408, C0966355
BECN1	8678	1 becn	C1412785, C1453431
CBS	875	cbs	C1439329
CREBBP	1387	cbp	C0056695, C1337090, C1454863, C0256079, C1455376
FEN1	2237	1 fen	C0541280, C0525494, C0252912, C1414583
FMR1	2332	1 fmr	C0806150, C1414649
FTL	2512	ftl	C1414852
GRIK1	2897	5 glur	C0536091
GRIK2	2898	6 glur	C0385096, C1415294
GRIN2A	2903	2 a grin	C1415299
GRIN2B	2904	2 b grin	C1415300
GSTO1	9446	1 gsto	C1421933

GSTO2	119391	2 gsto	C1427885
HAP1	9001	1 hap	C1455520, C1455519
HD	3064	hd	C0252274, C0872190, C1456457, C1415504, C0247953, C0872189, C0806318
HDAC1	3065	1 hd	C1333891, C1333892, C1334032
HEXA	3073	*	*
HIP1	9026	1 hip	C1415546, C1310518
JPH3	57338	3 jph	C1422484
MAP2K6	5608	6 mek	C1334475
MAP3K6	9064	*	*
MAPT	4137	protein tau	C0085401
MTHFR	4524	mthfr	C0919427
MTR	4548	mtr	C1417453
MTRR	4552	mtrr	C1417458
OGG1	4968	1 ogg	C1335081, C1313359, C0050091, C0167195
PEX7	5191	*	*
POU3F2	5454	2 3 f pou	C0250353, C1418762
PRNP	5621	prnp	C1418941, C0291825, C0285899
STH	246744	saitohin	C1137121
TBP	6908	tbp	C1337106
UCHL1	7345	thiolesterase ubiquitin	C0164005, C1436157, C1421309,

			C1435231
ZDHHC17	23390	14 hip	C1175645

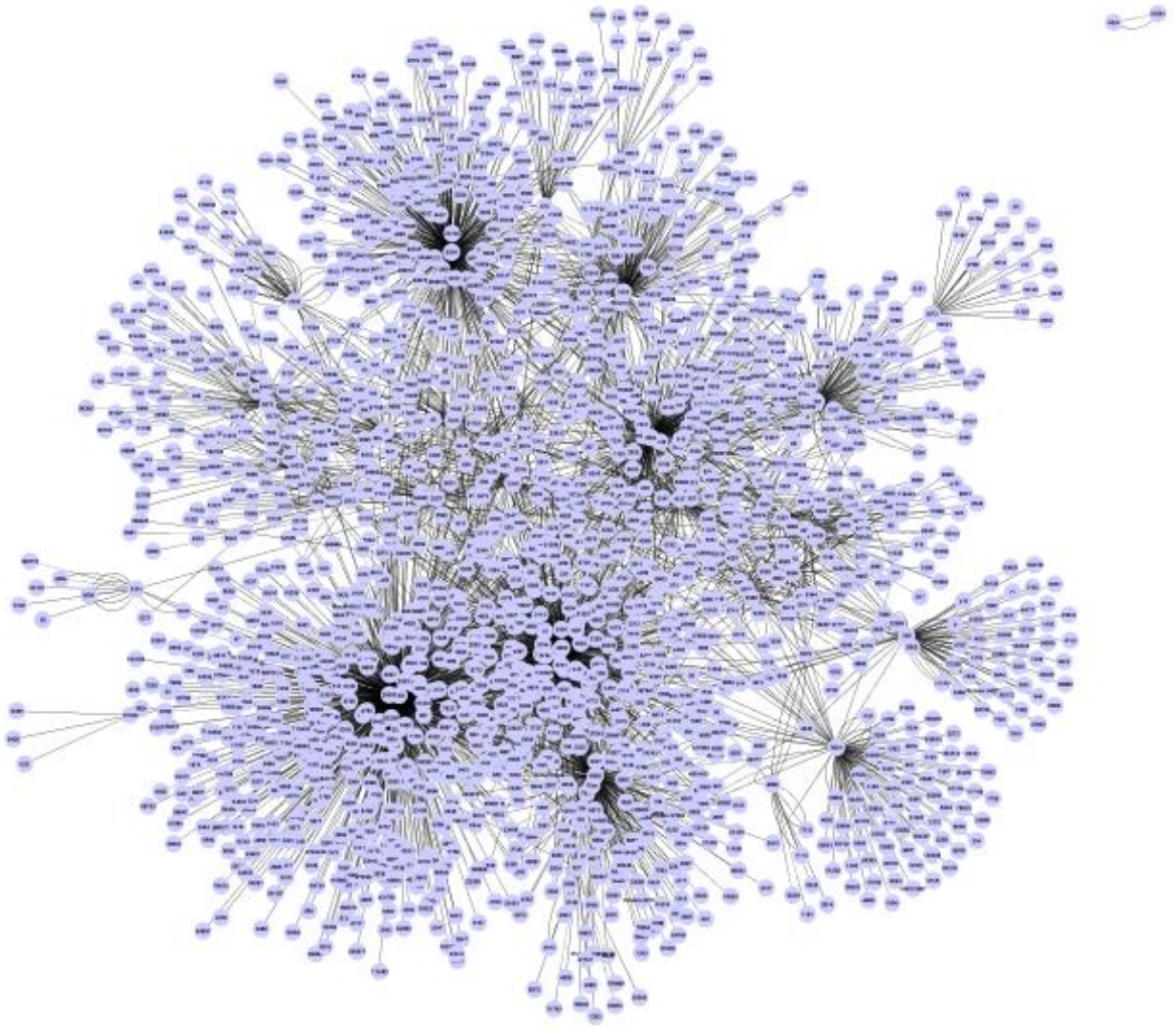


Figure 5-8. Hunting disease associated gene network using 41 seed genes compiled from GAD.

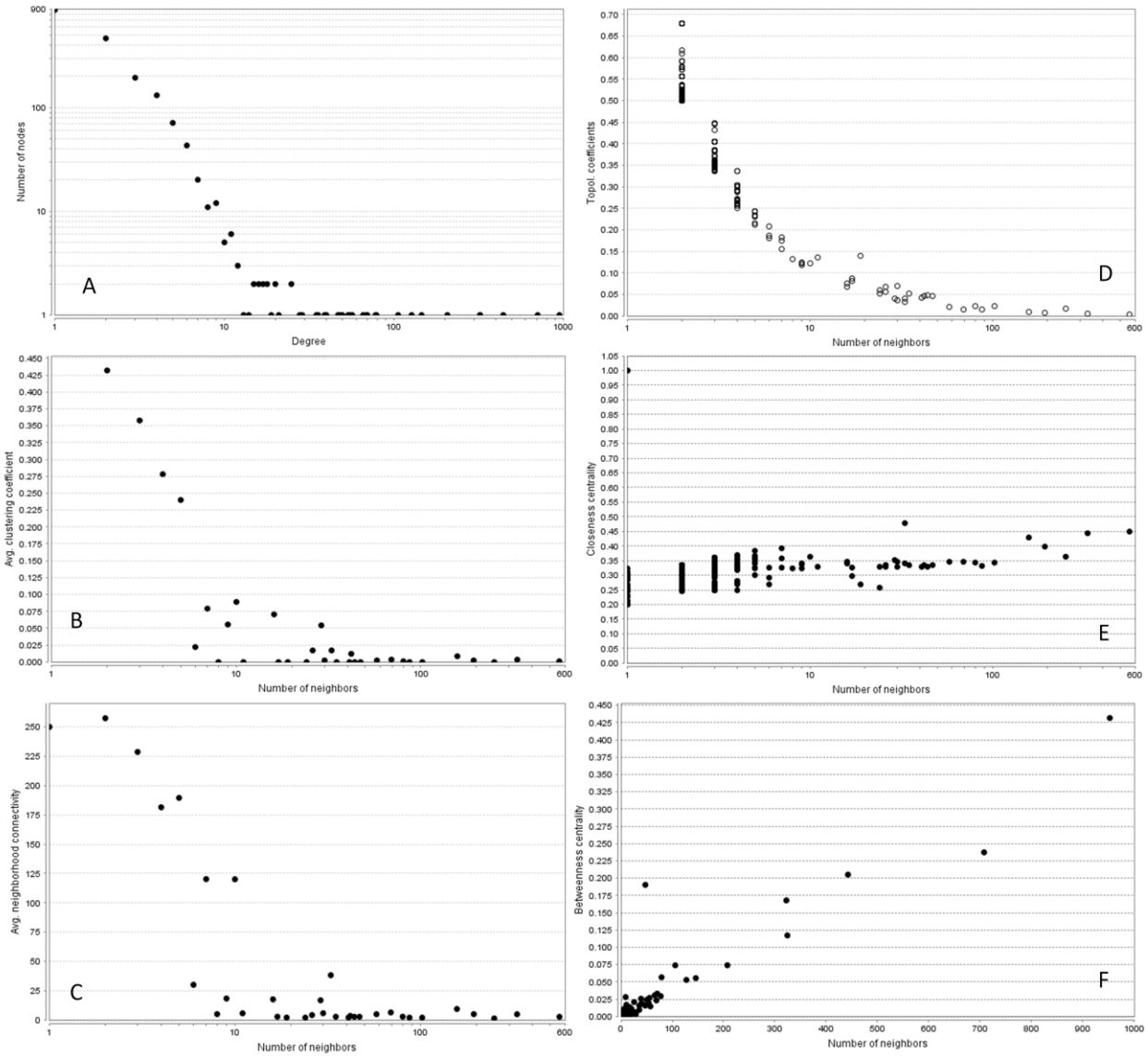


Figure 5-9. network analysis of Huntington disease associated gene network expanded from 41 seed genes.

5.3.7. Merge of literature mined HD disease-gene network with seed gene expanded network

We merged the seed gene expanded interaction network with the disease-gene heterogeneous network mined from literature to form a super graph by network union. Since

both networks utilized Entrez gene id as node identifier therefore identifier translation step is not needed, and the super graph is merged by Entrez identifier matching. As described in section 5.3.4, the Huntington disease is identified by unique id 0 in the merged network. The union process can be defined as follows:

For graphs *G1{V1, E1}* and *G2{V2, E2}* the union of both graphs is *G{V, E}*, where $V = V1 \cup V2 \; and \; E = E1 \cup E2$.

Figure 5-10 shows the merged heterogeneous disease-associated gene network. The center located HD disease is colored in red and the seed genes in green. The merged network is the neighborhood of the 41 seed genes and Huntington disease node.

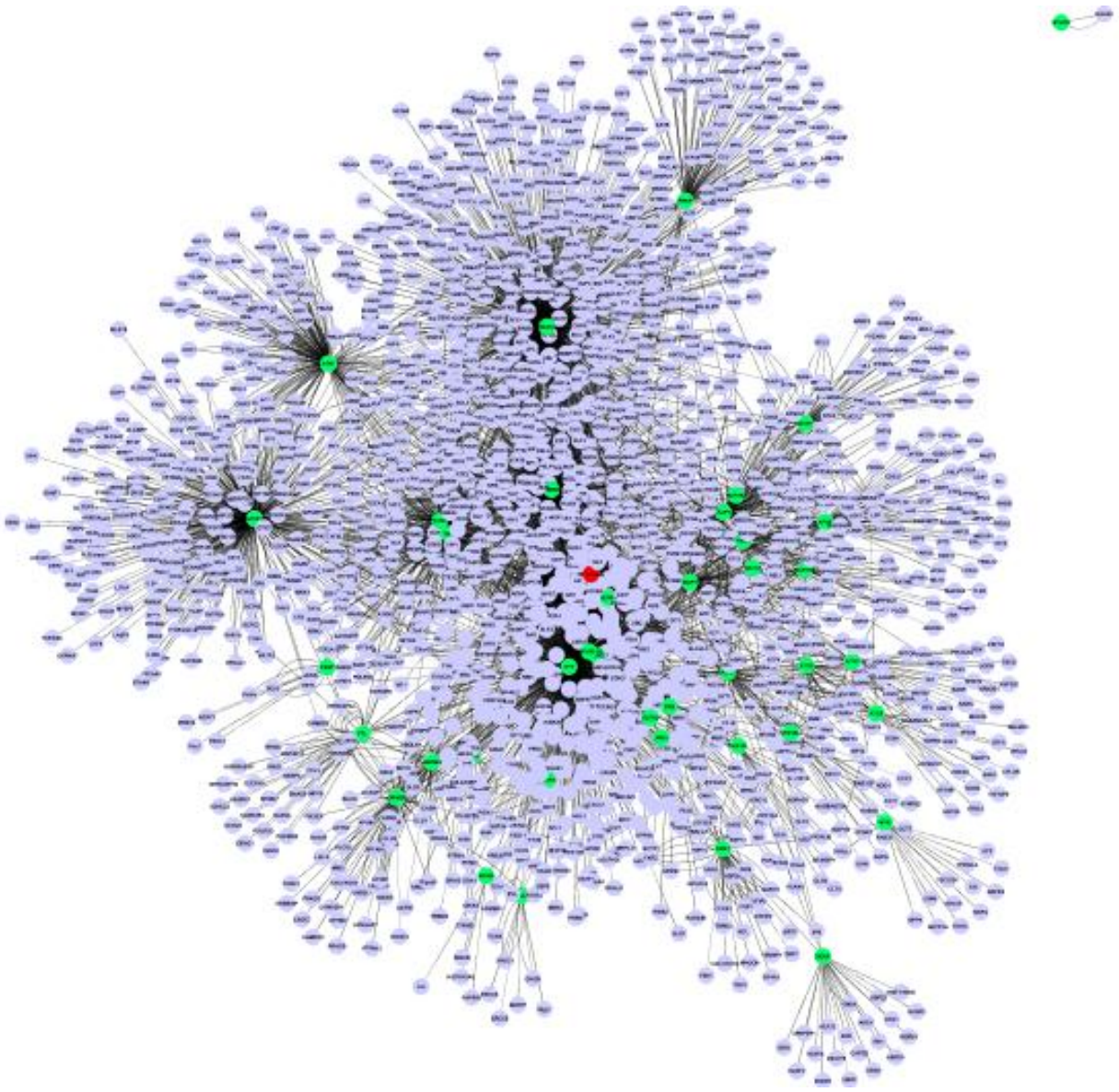


Figure 5-10. Merge of text graph with the GAD 41 gene network. The HD disease is highlighted in red color, and seed genes are in green color. Each gene node is labeled by its Entrez official name and the Huntington disease node is labeled by its official name.

We computed the 3 centrality measurements for gene nodes in the merged network and the top 25 most central genes is presented in table 5-4. In table 5-5 the precision of centrality ranking for top 10 and 25 genes are given as percentage of the top ranked 10 or 25 genes that are from the 41 seed gene set, which we use it as golden standard for evaluation. The betweenness

centrality performs best for top 10 (92% precision) and top 25 (92.2% precision) disease associated gene ranking. The degree centrality achieves 70% and 44.4% precision for top 10 and 25 genes ranking respectively. The closeness centrality achieves 60% and 22.4% for top 10 and 25 genes ranking.

Table 5-4. Top 25 genes ranked by centrality of the merged disease-gene heterogeneous network. Genes in bold font are seed genes and others are inferred disease associated genes.

Rank	Degree centrality	Betweeness centrality	Closeness centrality
1	**HDAC1**	**HDAC1**	**MTHFR**
2	**HTT**	**HTT**	NAA38
3	**CREBBP**	**CREBBP**	**HTT**
4	**ATXN1**	**ATXN1**	UBC
5	**TBP**	UBC	**HDAC1**
6	**BDNF**	**TBP**	**CREBBP**
7	**ATN1**	**ATN1**	**TBP**
8	GRINA	**PRNP**	HSPA4
9	AR	**HAP1**	AR
10	NOS2	**MAPT**	PIAS1
11	**HAP1**	**APOE**	EP300
12	**MAPT**	**UCHL1**	TP53
13	SEC16B	**BECN1**	MAPK8
14	SLC6A4	**GRIN2B**	SP1
15	**PRNP**	**CBS**	CTBP1

16	APC	**ATG7**	MYC
17	MAP3K14	**FTL**	JUN
18	GNAO1	**FEN1**	SET
19	LY6E	**HEXA**	CTNNB1
20	ERK	HSPA4	**BDNF**
21	**APOE**	**FMR1**	PPARG
22	DHDDS	**ATG3**	SUMO1
23	ATP8A2	**BDNF**	HEY2
24	GABRR1	**ATG16L1**	ACACA
25	BLM	**MTR**	CREB1

Table 5-5. Percentage of top 10 and 25 genes associated with Huntington disease based on 41 seed genes.

Top *n*	Degree centrality	Betweeness centrality	Closeness centrality
10	0.7	0.9	0.6
25	0.44	0.92	0.24

In figure 5-11 we plotted the scatter chart for degree and betweenness centrality of top 20 ranked genes. Their biological significance to the Huntington disease will be discussed in section 5.4.

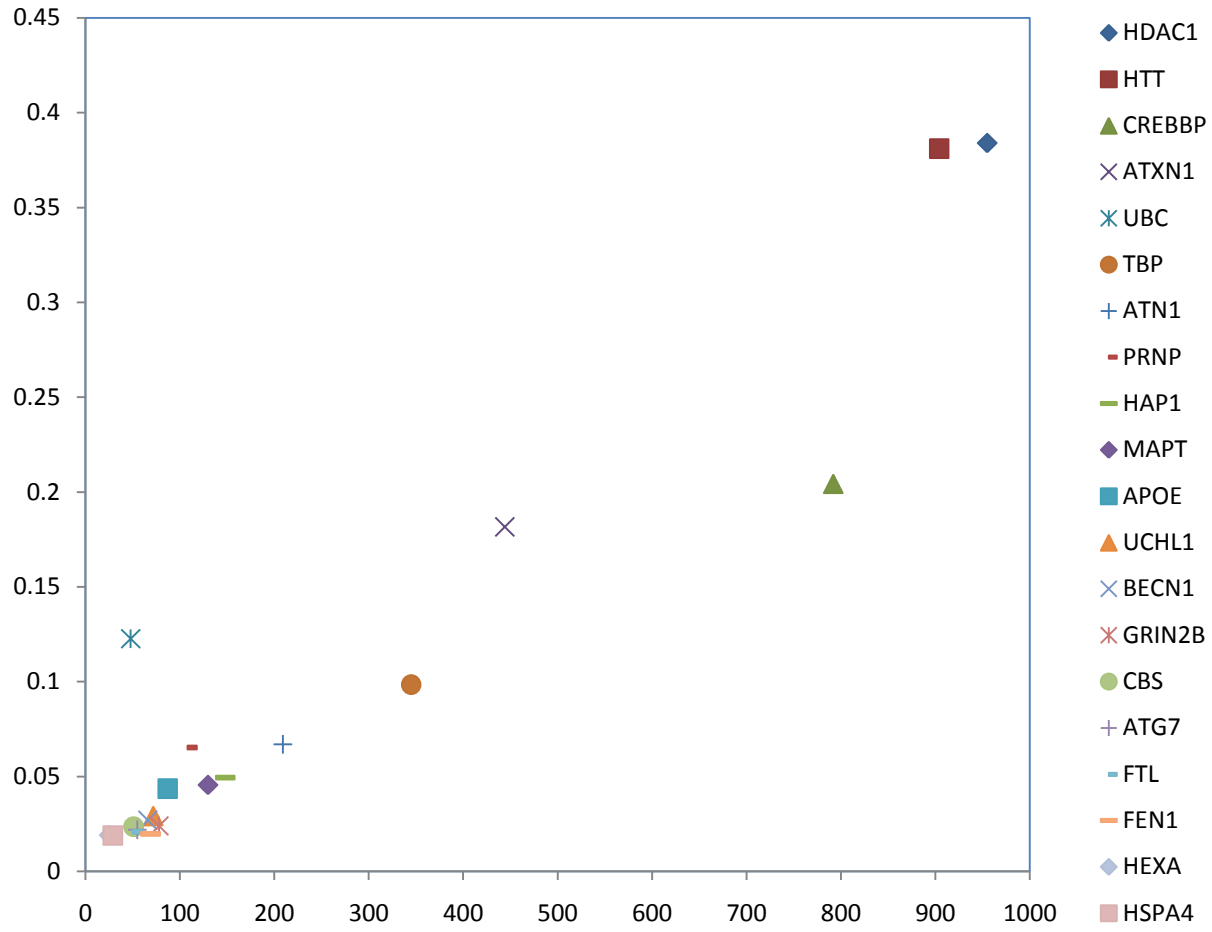


Figure 5-11. Scatter plot for degree (x-axis) and betweenness (y-axis) centrality of top 20 Huntington disease associated genes ranked by betweenness centrality.

5.4. Discussion and conclusion

In this chapter we presented a novel method to extract and rank disease associated genes through information extraction and network analysis. By using concept based information extraction approach we were able to build co-occurrence matrix between biomedical concepts using UMLS metathesaurus. We analyzed the correlation of extracted syndromes and diseases with HD disease to gain deeper understanding of semantic context about HD in our document collection. We also showed that concepts related to genes and gene products can be filtered and

joined by gene normalization, which allow us to construct a text graph containing concepts of interest, e.g. genes and disease, in order to analyze their associations under graph theory. Topological analysis of networks constructed from literature mining, seed gene expanding from database, and the heterogeneous network combining two, provided an integrated approach to study gene-disease associations. By taking global network topology into account, the hidden relations between gene-gene and gene-disease that do not occur in the same document can potentially be identified.

Our experiments suggest concepts network mined from literature forms a scale-free network with HTT gene as its central hub, consistent with the finding that biological networks are scale-free with power law distribution of connectivity around network hubs. As shown in figure 5-1 system architecture, our experiment starts with topic-focused PubMed document collection on Huntington disease, on which many empirical studies have been carried out to exploit its genetic association factors in an aim to find novel diagnostic and prognostic biomarkers. By using known HD associated genes from GAD as golden standard, we have shown our concept based extraction and gene normalization approach could retrieve 87.8% of known HD associated genes from 8843 PubMed abstracts out of millions of PubMed citations. We used the normalized genes and the HD disease to construct the heterogeneous disease-gene network. Network analysis using centrality of graph theory indicates the HTT is the central hub which is in well agreement with existing clinical and basic researches on HD (Bordelon 2013) (Shirasaki et al. 2012). Our integrated approach also revealed several predicted HD associated genes with high network centrality property. Among top ranked predicted genes in table 5-5 and figure 5-10, UBC is the polyubiquitin gene involved in ubiquitin proteolytic process and it has been shown presence of ubiquitin-positive neuronal inclusion bodies in HD brains (Li and Li

2011) as well as involvement of UBC gene (Bett et al. 2009). NOS2 or iNOS has been reported to be inhibited by Minocycline which delays Huntington disease progression in mouse model R6/2 (Thomas et al. 2004). HSPA4 is reported in NCBI AceView database to be functionally associated with Huntington disease (Cornett et al. 2005). The ERK is recently proposed as a novel target for Huntington disease in (Bodai and Marsh 2012). The NMDA receptors (GRIN glutamate receptor, ionotropic, N-methyl-d-aspartate) may also influence the variability in age of onset (AO) of HD (Arning et al. 2005). HD Research Crossroads database contains over 800 extensively curated genes relevant to HD (Kalathur et al. 2012). In (Kalathur et al. 2012) a global profiling based on this database have predicted 24 candidate genetic modifier of HD disease. Among top 25 of our predicted genes, SUMO1, CREB1, and HSPA4 are on their genetic modifier list.

To the same problem of finding candidate genes associated with human diseases from literature there exists different methods including co-occurrence based, rule based, and machine learning based relation extraction, with each having its own advantages and disadvantages. It is noted however, the co-occurrence based approach has its limitations on ranking newly identified disease associated genes, as that not many papers have been published to support the finding which leads to lower co-occurrence counts. By integrating bag-of-words co-occurrence, rule-based syntactic analysis and statistical method, and graph theory will likely to boost the novel disease-associated gene identification in terms of extraction precision and recall. It will be a subject of our future work.

CHAPTER 6. CONCLUSION AND FUTURE WORK

Information extraction (IE) in biomedical domain concerns itself with extraction of entities and their relationships for concise and precise data representation as well as decision making. Inspired by the growing interests on exploiting disease-gene relations and its application on future personalized medicine, in this thesis we focused our work on mining specific disease associated genes using IE methods ranging from the disease and gene entity recognition, to relation extraction and graph representation of their interaction networks.

6.1. Contributions of this thesis

6.1.1 Disease and gene named entity recognition

As the first step towards information extraction of disease associated genes, in chapter 3 a statistical machine learning approach (CRF method) was utilized to formulate the disease NER as a sequential prediction problem. The key to success of this approach lies on how document features are presented and how domain knowledge are utilized for the conditional statistical modeling. To this end, we explored rich set of textual features, and analyzed effect of domain specific POS tagging , domain specific dictionary, and entity encoding schemas on NER system performance. The results show that they are important factors contributing to the performance improvement of our disease NER system. We then utilized the sentence level semantic concept information as one of discriminative features for disease named entity recognition. Our method takes advantage of semantic types related to disease concept in UMLS metathesaurus by fuzzy dictionary lookup. We developed a new algorithm to engineer semantic concept feature into feature space for CRF training. The results show significant improvement for the performance of current disease NER methods with this new feature. To our knowledge, this is the first time the

semantic concept type is used as an important feature to improve biomedical named entity recognition.

Regarding the first research question, e.g. how to better represent text with concept features to improve disease NER using machine learning based approach, experimental results show UMLS semantic concept can be effectively incorporated into machine learning based NER to improve the overall disease NER performance and the concept feature and domain knowledge base enhanced NER outperforms state-of-the-art systems on this specific problem.

6.1.2. Disease-gene relation extraction

In chapter 4 we focused on another aspect of IE, e.g. relation extraction for disease-gene associations. We attempt to answer the question on how to develop efficient relation extraction machine learning model for disease associated gene mining. In this study we constructed an annotated corpora with human diseases and gene entities, annotated by NER system described in chapter 3 followed by manual curation. The relation extraction system uses the string kernel based SVM classification method to learn the global and local contextual information surrounding the two entities. Experimental results show that the global tri-grams 'bag-of-words' feature is more effective than local contextual features for disease-gene relation extraction, suggesting this shallow linguistic kernel based machine learning is a feasible and efficient approach to extract disease-gene relations from large text corpora.

6.1.3. Mining disease associated genes using IE and graph theory

Machine learning based relation extraction for mining disease-associated genes suffers from one major limitation, which is its reliance on annotated training corpora. In many real-

world applications the quantity and quality of annotated corpora are not guaranteed. Moreover, biomarker information is often dispersed in the entire abstract, thus making the machine learning based relation extraction at sentence level not an ideal solution to this problem.

In chapter 5 we presented a novel approach to identify and prioritize disease associated genes using concept co-occurrence and graph theory. In order to address research question on how to represent gene-gene and gene-disease network in concept space and achieve dimension reduction for the concept text graph, as well as how to incorporate concepts mined from literature with empirical data from protein interaction database to reveal and prioritize disease-associated genes by network topology analysis, we constructed text graph to represent disease gene network based on concept co-occurrence matrix. We demonstrated the feasibility of creating such text graph from large document collection by filtering the semantic types of concept and further gene name disambiguation and normalization. We expanded the network using a set of experimentally validated seed genes and protein-protein interaction dataset. The topology of resulting disease-gene heterogeneous network is analyzed, and important gene nodes are ranked by network centrality measurements. In consistent with findings on topology of most biological networks, the expanded heterogeneous network shows scale-free property, power law degree distribution, and connected by central hub genes. We demonstrated that centrality measurements not only retrieved known disease associated genes, but also revealed novel disease associated genes. These results provide us some useful insights into graph representation of concept space in biomedical information extraction field.

6.2. Future work

Mining biomarker genes from literature, as an increasing important subject of information extraction, involves multiple steps and each step can be approached by different methods. In this thesis we applied state-of-the-art machine learning methods, graph theory, and to some extent the concept space framework, to address key issues related to biomedical entity recognition as well as relation and knowledge extraction.

There are several aspects need further works with regards to above steps.

Firstly, in addition to concept type feature, semantic feature space need to be further exploited for machine learning based NER. A richer semantic feature space is expected to reduce the feature sparsity and noise commonly seen with bag-of-words feature space. Meanwhile high dimensionality of feature space can be addressed by feature induction method, which iteratively reduce high dimensional feature set by only preserving those features with information gain during training process. It is also interesting to utilize large silver-standard corpora such as CALBC for machine learning based NER, especially for certain biomedical subdomains that lack golden-standard corpora.

Secondly, for our preliminary work on machine learning based relation extraction, it is necessary to expand the annotated corpora size by including documents referenced by other major genetic association databases. Alternatively, phrases that describe disease-gene associations can be utilized to replace the manually annotated corpora. Source of such phrases can come from GeneRIF database, and potentially other online databases including OMIM, EntrezGene, and other metathesaurus and ontologies. Although shallow linguistic based contextual features have been shown in our results to be efficient for kernel based relation

learning, other features including parse tree and semantic features may also be considered in our future work.

Finally, in chapter 5 the text window between co-location of conceptual entities can be defined to distinguish co-occurrence at phrasal, sentence, paragraph, and section level. This granulate of constraint applied on the discoursed concept entities will be useful to improve computing of the association measurement and to better represent the associated entities in the text graph. Additionally, rule-based syntactic analysis and statistical method can be incorporated into our graph theory based solution to boost the recall and precision of disease associated gene identification and ranking. It is noted that newly published disease-associated genes with co-occurrence below cutoff threshold may not be ranked high. By utilizing large control negative corpora and statistical co-occurrence based methods their ranking may be boosted to reveal more predicted novel genes.

LIST OF REFERENCES

Adamic L a, Wilkinson D, Huberman B a, Adar E (2002) A literature based method for identifying gene-disease connections. Proc IEEE Comput Soc Bioinform Conf 1:109–17.

Agichtein E, Gravano L (2000) Snowball : Extracting Relations from Large Plain-Text Collections.

Ando RK, Zhang T (2005) A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. J Mach Learn Res 6:1817–1853.

Arning L, Kraus PH, Valentin S, et al. (2005) NR2A and NR2B receptor gene variations modify age at onset in Huntington disease. Neurogenetics 6:25–8. doi: 10.1007/s10048-004-0198-8

Assenov Y, Ramírez F, Schelhorn S-E, et al. (2008) Computing topological parameters of biological networks. Bioinformatics 24:282–4. doi: 10.1093/bioinformatics/btm554

Aurenhammer F (1991) Voronoi diagrams---a survey of a fundamental geometric data structure. ACM Comput Surv 23:345–405. doi: 10.1145/116873.116880

Bach N, Badaskar S (2007) A survey on relation extraction. Lit. Rev. Lang. Stat. II

Barabási, A. RA (1999) Emergence of Scaling in Random Networks. Science (80-) 286:509–512. doi: 10.1126/science.286.5439.509

Baum LE, Petrie T, Soules G, Weiss N (1970) A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. Ann Math Stat 41:164–171. doi: 10.1214/aoms/1177697196

Bett JS, Benn CL, Ryu K-Y, et al. (2009) The polyubiquitin Ubc gene modulates histone H2A monoubiquitylation in the R6/2 mouse model of Huntington's disease. J Cell Mol Med 13:2645–57. doi: 10.1111/j.1582-4934.2008.00543.x

Biomarkers Definitions Working Group (2001) Review Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 69:89–95.

Björne J, Ginter F, Pyysalo S, et al. (2010) Complex event extraction at PubMed scale. Bioinformatics 26:i382–90. doi: 10.1093/bioinformatics/btq180

Blanco R, Lioma C (2011) Graph-based term weighting for information retrieval. Inf Retr Boston 15:54–92. doi: 10.1007/s10791-011-9172-x

Bodai L, Marsh JL (2012) A novel target for Huntington's disease: ERK at the crossroads of signaling. The ERK signaling pathway is implicated in Huntington's disease and its upregulation ameliorates pathology. Bioessays 34:142–8.

Bordelon YM (2013) Clinical neurogenetics: huntington disease. Neurol Clin 31:1085–94. doi: 10.1016/j.ncl.2013.05.004

Brandes U (2001) A faster algorithm for betweenness centrality*. J Math Sociol 25:163–177. doi: 10.1080/0022250X.2001.9990249

Brown KR, Jurisica I (2005) Online predicted human interaction database. Bioinformatics 21:2076–82. doi: 10.1093/bioinformatics/bti273

Bundschus Markus, Dejori Mathaeus, Stetter Martin, et al. (2008) Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics 9:207. doi: 10.1186/1471-2105-9-207

Bunescu R, Ge R, Kate RJ, et al. (2005) Comparative experiments on learning information extractors for proteins and their interactions. Artif Intell Med 33:139–55. doi: 10.1016/j.artmed.2004.07.016

Bunescu RC, Mooney RJ (2005) A shortest path dependency kernel for relation extraction. Proc Conf Hum Lang Technol Empir Methods Nat Lang Process pages:724–731. doi: 10.3115/1220575.1220666

Bunescu RC, Mooney RJ (2006) Subsequence kernels for relation extraction. Adv Neural Inf Process Syst 18:171.

Chang C-C, Lin C-J (2011) LIBSVM. ACM Trans Intell Syst Technol 2:1–27. doi: 10.1145/1961189.1961199

Chen JYUE, Sivachenko AY (2006) Mining Alzheimer Disease Relevant Proteins from Integrated Protein Interactome Data. Pacific Symp Biocomput 11:367–378.

Cheng D, Knox C, Young N, et al. (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic Acids Res 36:W399–405. doi: 10.1093/nar/gkn296

Chowdhury F, Lavelli A (2011) Assessing the practical usability of an automatically annotated corpus. Proc 5th Linguist Annot Work 101–109.

Chun H-W, Tsuruoka Y, Kim J-D, et al. (2006a) Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. Pacific Symp Biocomput 15:4–15.

Chun H-W, Tsuruoka Y, Kim J-D, et al. (2006b) Automatic recognition of topic-classified relations between prostate cancer and genes using MEDLINE abstracts. BMC Bioinformatics 7 Suppl 3:S4. doi: 10.1186/1471-2105-7-S3-S4

Ciravegna F (2001) Adaptive information extraction from text by rule induction and generalisation. 1251–1256.

Collier N, Nobata C, Tsujii J (2000) Extracting the names of genes and gene products with a hidden Markov model. Proc. 18th Conf. Comput. Linguist. -. Association for Computational Linguistics, Morristown, NJ, USA, pp 201–207

Cornett J, Cao F, Wang C-E, et al. (2005) Polyglutamine expansion of huntingtin impairs its nuclear export. Nat Genet 37:198–204. doi: 10.1038/ng1503

Cotton RG, McKusick V, Scriver CR (1998) The HUGO Mutation Database Initiative. Science 279:10–1.

Culotta A, Sorensen J (2004) Dependency tree kernels for relation extraction. Proc 42nd Annu Meet Assoc Comput Linguist ACL 04 4:423–es. doi: 10.3115/1218955.1219009

David Campos, Sérgio Matos JLO (2012) Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools. Theory Appl Adv Text Min. doi: 10.5772/3115

Dietterich TG (1998) Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Comput 10:1895–1923.

Ding J, Berleant D, Nettleton D, Wurtele E (2002) Mining MEDLINE: abstracts, sentences, or phrases? Pac Symp Biocomput 326–37.

Donaldson I, Martin J, de Bruijn B, et al. (2003) PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics 4:11. doi: 10.1186/1471-2105-4-11

Feldman R, Dagan I (1995) Knowledge Discovery in Textual Databases (KDT). Int. Conf. Knowl. Discov. DATA Min. pp 112–117

Feldman R, Sanger J (2007) The text mining handbook: advanced approaches in analyzing unstructured data. Casualty Actuar Soc E-Forum, Spring 2010 423.

Fukuda K, Tamura A, Tsunoda T, Takagi T (1998) Toward information extraction: identifying protein names from biological papers. Pac Symp Biocomput 707–18.

Fundel K, Küffner R, Zimmer R (2007) RelEx--relation extraction using dependency parse trees. Bioinformatics 23:365–71. doi: 10.1093/bioinformatics/btl616

Gärdenfors P (2004) Conceptual Spaces: The Geometry of Thought. 307.

Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. Genome Res 11:1425–33. doi: 10.1101/gr.180801

Giuliano C, Lavelli A, Romano L (2006) Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. EACL 18:401–408.

Goh K-I, Cusick ME, Valle D, et al. (2007) The human disease network. Proc Natl Acad Sci U S A 104:8685–90. doi: 10.1073/pnas.0701361104

Gonzalez G, Uribe JC, Tari L, et al. (2007) Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures. Pac Symp Biocomput 39:28–39.

Grishman R, Sundheim B (1996) Message Understanding Conference-6. Proc. 16th Conf. Comput. Linguist. -. Association for Computational Linguistics, Morristown, NJ, USA, p 466

Harris Z (1958) Linguistic transformations for information retrieval. Proc Int Conf Sci Inf 158.

Hepple M (2000) Independence and Commitment: Assumptions for Rapid Training and Execution of Rule-based POS Taggers. Proc 38th Annu Meet Assoc Comput Linguist 277–278.

Hirschman L, Colosimo M, Morgan A, Yeh A (2005a) Overview of BioCreAtIvE task 1B: normalized gene lists. BMC Bioinformatics 6 Suppl 1:S11. doi: 10.1186/1471-2105-6-S1-S11

Hirschman L, Yeh A, Blaschke C, Valencia A (2005b) Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics 6 Suppl 1:S1. doi: 10.1186/1471-2105-6-S1-S1

Hoffmann R, Valencia A (2005) Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics 21 Suppl 2:ii252–8. doi: 10.1093/bioinformatics/bti1142

Hristovski D, Peterlin B, Mitchell J a, Humphrey SM (2005) Using literature-based discovery to identify disease candidate genes. Int J Med Inform 74:289–98. doi: 10.1016/j.ijmedinf.2004.04.024

HUGO Gene Nomenclature Committee HUGO Gene Nomenclature Committee Home Page | HUGO Gene Nomenclature Committee. http://www.genenames.org/.

Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 106:620–630.

Jeong H, Mason SP, Barabási AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411:41–2. doi: 10.1038/35075138

Jeong H, Tombor B, Albert R, et al. (2000) The large-scale organization of metabolic networks. Nature 407:651–4. doi: 10.1038/35036627

Jianfeng Gao HS (2005) Long Distance Dependency in Language Modeling: An Empirical Study. Lect Notes Comput Sci 3248:396–405. doi: 10.1007/b105612

Jiang J, Zhai C (2007) A Systematic Exploration of the Feature Space for Relation Extraction. HLT-NAACL

Jones BC (2002) Werner Kalow, Urs A. Meyer and Rachel F. Tyndale (eds): Pharmacogenomics. Genes, Brain Behav 1:194–194. doi: 10.1034/j.1601-183X.2002.103083.x

Jonnalagadda S, Cohen T, Wu S, et al. (2013) Using empirically constructed lexical resources for named entity recognition. Biomed Inform Insights 6:17–27. doi: 10.4137/BII.S11664

Joy MP, Brock A, Ingber DE, Huang S (2005) High-betweenness proteins in the yeast protein interaction network. J Biomed Biotechnol 2005:96–103. doi: 10.1155/JBB.2005.96

Jung J-Y, Deluca TF, Nelson TH, Wall DP (2013) A literature search tool for intelligent extraction of disease-associated genes. J Am Med Inform Assoc amiajnl–2012–001563–. doi: 10.1136/amiajnl-2012-001563

Kalathur RKR, Hernández-Prieto MA, Futschik ME (2012) Huntington's disease and its therapeutic target genes: a global functional profile based on the HD Research Crossroads database. BMC Neurol 12:47. doi: 10.1186/1471-2377-12-47

Kevin Becker, Kathleaen Barnes, Tiffani Bright AW (2004) The Genetic Associated Database. Nat Genet 36:431–432.

Kim J-D, Ohta T, Tateisi Y, Tsujii J (2003) GENIA corpus--a semantically annotated corpus for bio-textmining. Bioinformatics 19:i180–i182. doi: 10.1093/bioinformatics/btg1023

Kim J-D, Ohta T, Tsuruoka Y, et al. (2004) Introduction to the bio-entity recognition task at JNLPBA. 70–75.

Kim S, Yoon J, Yang J (2008) Kernel approaches for genic interaction extraction. Bioinformatics 24:118–26. doi: 10.1093/bioinformatics/btm544

Köhler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 82:949–58. doi: 10.1016/j.ajhg.2008.02.013

Kristin P. Bennett AD (1999) Semi-supervised support vector machines. Adv Neural Infor Process Syst 368–374.

Kulick S, Bies A, Liberman M, et al. (2004) Integrated Annotation for Biomedical Information Extraction. Production Linking Bi:61–68.

Lafferty J, McCallum A, Pereira F (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Mach. Learn. - Int. Work. pp 282–289

Lage K, Karlberg EO, Størling ZM, et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 25:309–16. doi: 10.1038/nbt1295

Lang F, Shooshan S, Mork J, Aronson A (2009) Ambiguity in the UMLS Metathesaurus. skr.nlm.nih.gov 2009:1–46.

Lawrence Page SBRMTW The PageRank Citation Ranking: Bringing Order to the Web.

Leaman R, Gonzalez G, Leaman Robert, Gonzalez Graciela (2008) BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput 652–63.

Lee I, Blom UM, Wang PI, et al. (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. 1109–1121. doi: 10.1101/gr.118992.110.7

Lee SH, Kim P, Jeong H (2009) Statistical properties of sampled networks.

Li X-J, Li S (2011) Proteasomal dysfunction in aging and Huntington disease. Neurobiol Dis 43:4–8. doi: 10.1016/j.nbd.2010.11.018

Li Y, Hu X, Lin H, Yang Z A framework for semisupervised feature generation and its applications in biomedical literature mining. IEEE/ACM Trans Comput Biol Bioinform 8:294–307. doi: 10.1109/TCBB.2010.99

Liu H, Hu Z-Z, Zhang J, Wu C (2006) BioThesaurus: a web-based thesaurus of protein and gene names. Bioinformatics 22:103–5. doi: 10.1093/bioinformatics/bti749

Lodhi H, Saunders C, Shawe-Taylor J, et al. (2002) Text classification using string kernels. J Mach Learn Res 2:419–444. doi: 10.1162/153244302760200687

Lu L, Li Y, Li S (2011) Computational identification of potential microRNA network biomarkers for the progression stages of gastric cancer. Int J Data Min Bioinform 5:519–531.

Mann G, McCallum A (2007) Efficient computation of entropy gradient for semi-supervised conditional random fields. Conf North Am Chapter Assoc Comput Linguist 109–112.

Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. Science 296:910–3. doi: 10.1126/science.1065103

Mason O, Verwoerd M (2007a) Graph theory and networks in Biology. IET Syst Biol 1:89–119.

Mason O, Verwoerd M (2007b) Graph theory and networks in biology. Syst Biol IET 52.

McCallum A (2003) Efficiently inducing features of conditional random fields. Proc Ninet Conf Uncertain Artif Intell 403–410.

McCallum A, Freitag D, Pereira FCN (2000) Maximum Entropy Markov Models for Information Extraction and Segmentation. 591–598.

McCallum AK (2002) MALLET: A Machine Learning for Language Toolkit.

Minsky M (1968) Semantic information processing. MIT Press, Cambridge Mass.

Nadeau D, Sekine S (2007) A survey of named entity recognition and classification. Lingvisticae Investig 1–20.

NCBI PubMed - NCBI. http://www.ncbi.nlm.nih.gov/pubmed/.

NCBI OMIM - NCBI. http://www.ncbi.nlm.nih.gov/omim/.

NCBI Entrez Gene - NCBI. http://www.ncbi.nlm.nih.gov/gene.

NCBI MeSH - NCBI. http://www.ncbi.nlm.nih.gov/mesh.

Nédellec C (2005) Learning Language in Logic - Genic Interaction Extraction Challenge. Proc. Learn. Lang. Log. 2005 Work. Int. Conf. Mach. Learn.

Neves ML, Carazo J-M, Pascual-Montano A (2010) Moara: a Java library for extracting and normalizing gene and protein mentions. BMC Bioinformatics 11:157. doi: 10.1186/1471-2105-11-157

Newman MEJ (2003) A measure of betweenness centrality based on random walks. 15.

Ozgür A, Vu T, Erkan G, Radev DR (2008) Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics 24:i277–85. doi: 10.1093/bioinformatics/btn182

P. Erdos AR (1960) On the Evolution of Random Graphs. Publ. Math. Inst. HUNGARIAN Acad. Sci. pp 17–61

Perez-Iratxeta C, Bork P, Andrade MA (2002) Association of genes to genetically inherited diseases using data mining. Nat Genet 31:316–9. doi: 10.1038/ng895

PETITOT J (1988) MORPHODYNAMICS AND THE CATEGORICAL PERCEPTION OF PHONOLOGICAL UNITS. Theor Linguist 15:25 – 72. doi: 10.1515/thli.1988.15.1-2.25

Pustejovsky J, Castaño J, Zhang J, et al. (2002) Robust relational parsing over biomedical literature: extracting inhibit relations. Pac Symp Biocomput 362–73.

Pyysalo S, Ginter F, Heimonen J, et al. (2007) BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinformatics 8:50. doi: 10.1186/1471-2105-8-50

Pyysalo S, Ohta T, Rak R, et al. (2012) Overview of the ID, EPI and REL tasks of BioNLP Shared Task 2011. BMC Bioinformatics 13 Suppl 1:S2. doi: 10.1186/1471-2105-13-S11-S2

Quinlan JR (1990) Learning logical definitions from examples. Mach Learn 5:239–266.

Raja K, Subramani S, Natarajan J (2013) PPInterFinder--a mining tool for extracting causal relations on human proteins from literature. Database (Oxford) 2013:bas052. doi: 10.1093/database/bas052

Ratnaparkhi A (1997) A Simple Introduction to Maximum Entropy Models for Natural Language Processing. IRCS Tech. Reports Ser.

Rebholz-Schuhmann D, Jimeno Yepes AJ, Van Mulligen EM, et al. (2010) CALBC silver standard corpus. J Bioinform Comput Biol 8:163–79.

Rebholz-Schuhmann D, Kirsch H, Arregui M, et al. (2006) Protein annotation by EBIMed. Nat Biotechnol 24:902–3. doi: 10.1038/nbt0806-902

Rebholz-Schuhmann D, Marcel S, Albert S, et al. (2004) Automatic extraction of mutations from Medline and cross-validation with OMIM. Nucleic Acids Res 32:135–42. doi: 10.1093/nar/gkh162

Rindflesch TC, Tanabe L, Weinstein JN, Hunter L (2000) EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput 517–28.

Rosario B, Hearst M (2004) Classifying semantic relations in bioscience texts. Proc. 42nd Annu. Meet. Assoc. Comput. Linguist.

Sarawagi S (2007) Information extraction. Found Trends Database 1:261–377. doi: 10.1561/1500000003

Sasaki Y, Tsuruoka Y, McNaught J, Ananiadou S (2008) How to make the most of NE dictionaries in statistical NER. BMC Bioinformatics 9 Suppl 11:S5. doi: 10.1186/1471-2105-9-S11-S5

Scardoni G, Petterlini M, Laudanna C (2009) Analyzing biological network parameters with CentiScaPe. Bioinformatics 25:2857–9. doi: 10.1093/bioinformatics/btp517

Sha F, Pereira F (2003) Shallow parsing with conditional random fields. Proc. 2003 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. NAACL 03. pp 134–141

Shannon P, Markiel A, Ozier O, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–504. doi: 10.1101/gr.1239303

Shawe-Taylor J, Cristianini N (2004) Kernel Methods for Pattern Analysis. 462.

Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet 31:64–8. doi: 10.1038/ng881

Shirasaki DI, Greiner ER, Al-Ramahi I, et al. (2012) Network organization of the huntingtin proteomic interactome in mammalian brain. Neuron 75:41–57. doi: 10.1016/j.neuron.2012.05.024

Singh HA-M, K R (2005) A New Text Mining Approach for Finding Protein-to-Disease Associations. Am J Biochem Biotechnol 1:145–152.

Smith L, Rindflesch T, Wilbur WJ (2004) MedPost: a part-of-speech tagger for bioMedical text. Bioinformatics 20:2320–2321. doi: 10.1093/bioinformatics/bth227

Stelzl U, Worm U, Lalowski M, et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122:957–68. doi: 10.1016/j.cell.2005.08.029

Stephen Soderland, Claire Cardie RM (1999) Learning Information Extraction Rules for Semi-structured and Free Text. Mach Learn 1:233–272.

Surdeanu M, Harabagiu S, Williams J, Aarseth P (2003) Using Predicate-argument Structures for Information Extraction. Proc. 41st Annu. Meet. Assoc. Comput. Linguist. pp xx–xx

Tanabe L, Xie N, Thom LH, et al. (2005) GENETAG: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics 6 Suppl 1:S3. doi: 10.1186/1471-2105-6-S1-S3

Thomas M, Ashizawa T, Jankovic J (2004) Minocycline in Huntington's disease: a pilot study. Mov Disord 19:692–5. doi: 10.1002/mds.20018

Thompson P, Iqbal SA, McNaught J, Ananiadou S (2009) Construction of an annotated corpus to support biomedical information extraction. BMC Bioinformatics 10:349. doi: 10.1186/1471-2105-10-349

Tong AHY, Lesage G, Bader GD, et al. (2004) Global mapping of the yeast genetic interaction network. Science 303:808–13. doi: 10.1126/science.1091317

Tsai RT-H, Lai P-T, Dai H-J, et al. (2009) HypertenGene: extracting key hypertension genes from biomedical literature with position and automatically-generated template features. BMC Bioinformatics 10 Suppl 1:S9. doi: 10.1186/1471-2105-10-S15-S9

Tsuruoka Y, Tsujii J (2004) Improving the performance of dictionary-based approaches in protein name recognition. J Biomed Inform 37:461–70. doi: 10.1016/j.jbi.2004.08.003

Viterbi A (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inf Theory 13:260–269. doi: 10.1109/TIT.1967.1054010

Wagner A (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol Evol 18:1283–92.

Wagner A, Fell DA (2001) The small world inside large metabolic networks. Proc Biol Sci 268:1803–10. doi: 10.1098/rspb.2001.1711

Wallach H (2004) Conditional random fields: An introduction. Tech Reports 1–9.

Warby, Simon C, Rona K Graham MRH (2010) Huntington Disease. GeneReviews

Watts DJ, Strogatz SH (1998) Collective dynamics of "small-world" networks. Nature 393:440–2. doi: 10.1038/30918

Wermter J, Tomanek K, Hahn U (2009) High-performance gene name normalization with GeNo. Bioinformatics 25:815–21. doi: 10.1093/bioinformatics/btp071

Wu CH, Huang H, Nikolskaya A, et al. (2004) The iProClass integrated database for protein functional analysis. Comput Biol Chem 28:87–96.

Wu X, Jiang R, Zhang MQ, Li S (2008) Network-based global inference of human disease genes. Mol Syst Biol 4:189. doi: 10.1038/msb.2008.27

Wuchty S, Stadler PF (2003) Centers of complex networks. J Theor Biol 223:45–53.

Yao X, Hao H, Li Y, Li S (2011) Modularity-based credible prediction of disease genes and detection of disease subtypes on the phenotype-gene heterogeneous network. BMC Syst Biol 5:79. doi: 10.1186/1752-0509-5-79

Yu H, Greenbaum D, Xin Lu H, et al. (2004) Genomic analysis of essentiality within protein networks. Trends Genet 20:227–31. doi: 10.1016/j.tig.2004.04.008

Zelenko D, Aone C, Richardella A (2003) Kernel methods for relation extraction. J Mach Learn Res 3:1083–1106.

Zhang M (2006) A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features. Proc 21st Int Conf Comput Linguist 44th Annu Meet ACL 825–832. doi: 10.3115/1220175.1220279

Zhang T, Johnson D (2003) A robust risk minimization based named entity recognition system. Proc. seventh Conf. Nat. Lang. Learn. HLT-NAACL 2003 -. Association for Computational Linguistics, Morristown, NJ, USA, pp 204–207

Zhang X, Gao Z, Rong Z, Zhu Y (2011) Two novel composite kernels for relation extraction. 2011 Int Conf Multimed Technol 5207–5210. doi: 10.1109/ICMT.2011.6002253

Zhou GD (2006) Recognizing names in biomedical texts using mutual information independence model and SVM plus sigmoid. Int J Med Inform 75:456–67. doi: 10.1016/j.ijmedinf.2005.06.012

Zhou X, Hu X, Zhang X (2007) Dragon Toolkit: Incorporating Auto-Learned Semantic Knowledge into Large-Scale Text Retrieval and Mining. 19th IEEE Int Conf Tools with Artif Intell 2007 2:197–201. doi: 10.1109/ICTAI.2007.117

Zhou X, Zhang X, Hu X (2006) MaxMatcher: Biological concept extraction using approximate dictionary lookup. PRICAI 2006 Trends Artif Intell 1145–1149.

Zhu S, Okuno Y, Tsujimoto G, Mamitsuka H (2005) A probabilistic model for mining implicit “chemical compound-gene” relations from literature. Bioinformatics 21 Suppl 2:ii245–51. doi: 10.1093/bioinformatics/bti1141

Zhu S, Okuno Y, Tsujimoto G, Mamitsuka H (2006) Application of a new probabilistic model for mining implicit associated cancer genes from OMIM and medline. Cancer Inform 2:361–71.

Zotenko E, Mestre J, O’Leary DP, Przytycka TM (2008) Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. PLoS Comput Biol 4:e1000140. doi: 10.1371/journal.pcbi.1000140

VITA

ZHONG HUANG

EDUCATION

Ph.D. Information Science & Technology (February 2014), Drexel University, PA

M.S. Information Systems (2006), Drexel University, PA

M.S. Physiology (1994), Beijing University Medical Center, Beijing, China

B.S. Medicine (1991), LuZhou Medical College, China

RESEARCH INTERESTS

Data mining, Bioinformatics, Information extraction in biomedical domain

SELECTED PUBLICATIONS

1. Zhong Huang, Xiaohua Hu: Disease Named Entity Recognition by Machine Learning Using Semantic Type of Metathesaurus. *International Journal of Machine Learning and Computing*, 3(6), 2013, Pages: 494-498
2. Zhong Huang: Mining Disease Associated Biomarker Networks from PubMed. *7th International Conference on Systems Biology (ISB 2013) on Issue*, 23-25 Aug, 2013, Pages:15-18, HuangShan, China
3. Zhong Huang: Image Retrieval for Open Access PubMedCentral Archive. *IEEE 2011 international Conference on Medical Information and Bioengineering (ICMIB 2011),* January 14-15, 2011, Pages: 111-114, Chengdu, China
4. Zhong Huang, Xuheng Xu, Xiaohua Hu: Machine Learning Approach for Human Mitochondrial Protein Prediction. Book: *Computational Intelligence in Bioinformatics,* IEEE CS Press/Wiley, 2008.
5. Zhong Huang, Xuheng Xu, Xiaohua Hu: Prediction Of Human Mitochondrial Proteins Using SVM And Neural Network, accepted in the *2007 Asia-Pacific Bioinformatics Conference*, poster paper, Jan 15-17, 2007, HongKong
6. Zhong Huang, Xiaohua Hu: Object Oriented Modeling of Protein Translation Systems, in the *Proceedings of the 2006 IEEE International Confernece on Granular Computing, (IEEE GrC 2006)*, Atlanta, GA, May 15-17, 2006, Pages: 353-356
7. Zhong Huang, Yun Li, Xiaohua Hu: SVM Classification to Predict Two Stranded Anti-parallel Coiled Coils based on Protein Sequence Data, in the *Proceedings of the 2005 International Conference on Computational Science and its Applications,* Pages: 374-382