

國立臺灣大學管理學院資訊管理學研究所



碩士論文

Graduate Institute of Information Management

College of Management

National Taiwan University

Master Thesis

建構生物醫學文獻與知識庫探勘技術以尋找舊藥之新用途

Mining Biomedical Literature and Ontologies for Drug

Repositioning Discovery

陳奎安

Kuei-an Chen

指導教授：魏志平 博士

Advisor: Chih-Ping Wei, Ph.D.

中華民國 102 年 7 月

July, 2013

國立臺灣大學碩士學位論文  
口試委員會審定書

建構生物醫學文獻與知識庫探勘技術以尋找舊  
藥之新用途

Mining Biomedical Literature and Ontologies for  
Drug Repositioning Discovery

本論文係 陳奎安 君（學號 R00725020）在國立臺灣  
大學資訊管理學所完成之碩士學位論文，於民國 102 年 7 月  
16 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

曾新怡

吳怡瑾

魏志平

所 長：

李瑞庭



## 誌謝

兩年的碩士生生涯就這麼晃眼而過，終要畫下句點。我並非一個自我要求甚高的學生，能夠走到今日，自是需要感謝許多師長同儕的指導、激勵與協助的。

本論文的完成，和我在碩士班兩年的日子裡，最要感謝的人莫過於我的指導教授——魏志平老師。魏老師嚴謹的治學態度，邏輯清晰、言詞暢達地教學、指導、與解惑，無論在身教、言教上，都為我們樹立了一個師範。在研究上，老師總是慈藹地循循善誘，引領我們突破思考盲點，刺激我們有系統地進一步解決問題；在人生上，老師也時常關心我們的生活、生涯規劃、做人處事、以及人生大事，並啟發我們如何成熟、積極、負責地面對人生路。不論在各方面，魏老師都對我有莫大的啟迪，能蒙老師的教導，學生實在粉身碎骨、感激不盡。此外，也特別感謝兩位口試委員：曾新穆老師與吳怡瑾老師，在口試時給予許多寶貴的意見，使本論文更臻完備。

接著，我要感謝虹鈞，為我們這些小蘿蔔頭張羅各種大小事務，並不時賞賜許多異國美食，微臣在此叩恩。感謝陳連進博士，在研究上給予許多建議和協助。感謝魏門的各位好夥伴：感謝尹安，一起沒爬成玉山和沒看成蘭嶼的星光；感謝冠宇，從入學第一天就比鄰而坐至今，祝早日成家立業；感謝細心又緊張的鴻英，幫了我很多忙，祝回澳門後事業愛情兩得意；感謝泰頤，同窗兩年新認識了你的許多不同面向，別再向太陽怒吼了，你一定會找到幸福的！感謝小蛇、牛奶、兔子、柏勳、好人等學長姐給予我的指點，以及學弟妹們：光昇、采璇、黃蕙、蓓妤的歡樂氣氛令我們挺過許多壓力。感謝我的好朋友們：感謝韻如，你的支持與包容給我極大的鼓舞；感謝其他篇幅不足無法詳列的鼓勵我、聽我抱怨的朋友們。

最後，要壓軸感謝的，當然是我的父母與家人，雖然我總是沒講清楚我到底在忙些什麼，但你們始終支持我做的決定，給予我鼓勵與方向，我愛你們。

陳奎安 謹識

于台大資訊管理學研究所

西元二零一三年七月

## 中文摘要

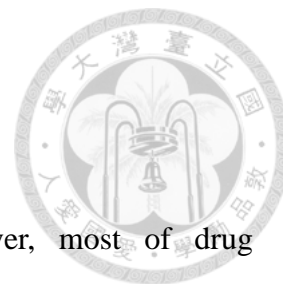


新藥物的推出能為藥品公司帶來相當的收入。然而，典型的新藥開發需要耗費大量資金與時間，且通過實驗、順利上市的成功率相當低，因此大部分的投資無法回收。近年來，許多藥品公司逐漸引入「舊藥新用」作為藥品開發的替代研發方法。「舊藥新用」是從既有的藥物，在其原本設計標的之外，尋找新適應症的藥物開發方式；由於既有藥物已經有許多前置研究基礎，可以省去許多臨床前評估與測試，藥品公司因而可以減少開發的時間與資金成本。

本研究基於 Swanson 提出的文獻探勘方法，分析超過 15,000,000 篇生物醫學文獻、以及藥物與疾病之知識庫，以自動化尋找尚未被發現且可能有直接關聯的既有藥物與疾病關係。我們建立三個實驗情境以評估本研究所提出之方法效能，其結果顯示，本研究提出之方法與所建構之綜合生物醫學概念網路能有效較既有方法有效提供潛在的藥物與疾病關係給研究者，以幫助研究者尋找可能的舊藥之新用途。

關鍵字：舊藥新用、文獻探勘、醫學文獻探勘

## ABSTRACT



Drug development is time-consuming and costly. However, most of drug development projects fail before they ever enter into clinical trials. To reduce the high risk of failure for drug development, pharmaceutical companies are exploring the drug repositioning approach for drug development. Previous studies have shown the feasibility of using computational methods to help extract plausible drug repositioning candidates, but they all encountered some limitations. We thus propose a novel drug-repositioning discovery method that takes into account multiple information sources, including more than 15,000,000 biomedical research articles and existing ontologies that cover detailed information about drugs, proteins and diseases, and follow the ABC model derived from Swanson's literature-based discovery works. We design three experiments to evaluate our proposed drug repositioning discovery method. The results show that our proposed method and our proposed integrated information source can better help researchers sift plausible drug-disease relationships in comparison with existing techniques.

Keywords: Drug repositioning, Drug repurposing, Literature-based discovery, Medical literature mining

# CONTENTS



口試委員會審定書 .....	i
誌謝 .....	ii
中文摘要 .....	iii
ABSTRACT .....	iv
CONTENTS .....	v
LIST OF FIGURES .....	vii
LIST OF TABLES .....	viii
<b>Chapter 1 Introduction .....</b>	<b>1</b>
1.1 Background .....	1
1.2 Research Motivation and Objective .....	5
<b>Chapter 2 Literature Review .....</b>	<b>7</b>
2.1 Literature-based Approach .....	7
2.2 Ontology-based Approach .....	10
<b>Chapter 3 Design of Drug Repositioning Discovery Method .....</b>	<b>13</b>
3.1 Literature-based Concept Network Construction .....	14
3.1.1. Data Collection .....	15
3.1.2. Link Extraction and Filtering .....	16
3.2 Ontology-based Concept Network Construction.....	17
3.2.1. Data Collection .....	17
3.2.2. Concept Mapping .....	19
3.2.3. Link Extraction and Filtering .....	19

3.3 Related Concept Retrieval .....	20
3.4 Link Weighting .....	22
3.5 Target Term Ranking .....	23
3.5.1 Single Intermediate Level Scenario.....	24
3.5.2 Multiple Intermediate Levels Scenario.....	25
<b>Chapter 4 Evaluation and Results .....</b>	<b>26</b>
4.1 Evaluation Design.....	26
4.2 Experiment 1: Comprehensive Network and Link Weighting Algorithm	27
4.3 Experiment 2: Target Term Ranking Algorithms for Single Intermediate Level Scenario.....	29
4.4 Experiment 3: Multiple Intermediate Levels Scenario.....	31
4.4.1 Parameter Tuning.....	31
4.4.2 Experiment Result .....	34
<b>Chapter 5 Conclusion and Future Work.....</b>	<b>36</b>
<b>References .....</b>	<b>38</b>



## LIST OF FIGURES

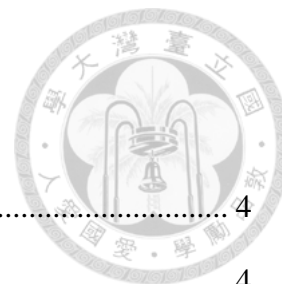


Figure 1.	Swanson’s literature-based discovering methodology.....	4
Figure 2.	Cheng et al.’s ontology-based network-based inference .....	4
Figure 3.	Graphical representation of Swanson’s ABC model. ....	8
Figure 4.	Overall Process of Our Drug Repositioning Discovery Method .....	14
Figure 5.	Illustration of Constraint to Category of Intra-intermediate Terms.....	21
Figure 6.	Illustration of Threshold to Number of Neighbors .....	21



# LIST OF TABLES



Table 1.	Selected Examples of Repositioned Drugs.....	3
Table 2.	Excluded Publication Types.....	15
Table 3.	Selected MeSH Subcategories.....	17
Table 4.	Top 10 Most Connected Intermediate Terms in Our Evaluation .....	21
Table 5.	Semantic Groups Selected for Our Experiments .....	27
Table 6.	Evaluation Results of Our Link Weighting and Comprehensive Network..	28
Table 7.	Comparison of Target Term Ranking Algorithms (Literature-only).....	29
Table 8.	Comparison of Target Term Ranking (Multiple Sources) .....	30
Table 9.	Tuning of <i>max_neighbor</i> ( $2 \leq \ell \leq 3$ ).....	32
Table 10.	Tuning of $\beta$ for Katz measurement ( $2 \leq \ell \leq 3$ ) .....	32
Table 11.	Tuning of <i>max_neighbor</i> ( $2 \leq \ell \leq 4$ ).....	33
Table 12.	Tuning of $\beta$ for Katz measurement ( $2 \leq \ell \leq 4$ ) .....	33
Table 13.	Comparison of Single and Multiple Intermediate Levels.....	34
Table 14.	Comparison of Using Different Information Sources ( $2 \leq \ell \leq 3$ ).....	35

# Chapter 1 Introduction



## 1.1 Background

Drug development is time consuming and costly. As United States Food and Drug Administration (FDA) regulated, the process of drug development can broadly be divided into two major stages: discovery and preclinical stage and clinical stage. In the discovery and preclinical stage, the pharmaceutical company or sponsor performs laboratory and animal tests to discover how the drug works and whether it is likely to be safe and work well in humans. After obtaining promising data, the candidate drug shall enter the clinical stage. It must pass all three phases of clinical trials (phase 1 studies typically involve 20 to 80 people; phase 2 studies typically involve a few dozen to about 300 people; and phase 3 studies typically involve several hundred to about 3,000 people), to determine whether the drug is safe when used to treat a disease and whether it provides a real health benefit. The whole process requires about 10-15 years, and costs between 500 million and 2 billion U.S. dollars to bring a new drug to market (Adams & Brantner, 2006; Pharmaceutical Research and Manufacturers of America, 2007).

For *de novo* drug development, about a half of the time and one-third of the total cost spend on discover and preclinical stage (DiMasi & Grabowski, 2007). Moreover, 80 to 90 percent of research projects fail before they ever get tested in human, according to the U.S. National Institutes of Health (National Institutes of Health, 2009). It is estimated that only 5 out of 5,000-10,000 tested compounds will qualify for clinical trials (Pharmaceutical Research and Manufacturers of America, 2007). To reduce the

high risk of failure for *de novo* drug development, pharmaceutical companies have been evaluating alternative paradigms for drug development, one of them being drug repositioning.



Drug repositioning is the process of finding new indications (i.e., treatment for diseases), other than original purposes, for existing drugs. Since the existing drugs already have their preclinical properties and established safety profiles, several experiments, analysis and tests can therefore be bypassed. Companies may thus reduce significant time and spending in the discovering and preclinical stage. Another advantage of drug repositioning is to make full use of company's intellectual property portfolio. By developing new uses of drugs, it is possible to extend their old, expiring patents, or get new method-of-use patents (Ashburn & Thor, 2004).

One notable example of repositioned drug is Thalidomide. It was originally marketed as a sedative and antiemetic for pregnant women to treat morning sickness, but was completely removed from the market after the drug was found responsible for severe birth defects (McBride, 1961). However, the banned drug was later discovered that it can effectively treat erythema nodosum leprosum (ENL), an agonizing inflammation in leprosy patients (Stephens & Brynner, 2001). After Celgene Corporation's repositioning works, FDA approved thalidomide for use in the treatment of ENL in 1998. The company further discovered that the drug is highly effective against several other diseases including multiple myeloma, a type of blood cell cancer that affects the bones and kidney. Accordingly, Celgene gets several utility patents for the repositioned thalidomide, and it brings in over 300 million U.S. dollars in revenue annually since 2004 (Celgene Corporation, 2006; 2009; 2013). In addition to

Thalidomide, several other repositioned drugs have been identified and reported. Table 1 shows some selected examples of repositioned drugs.

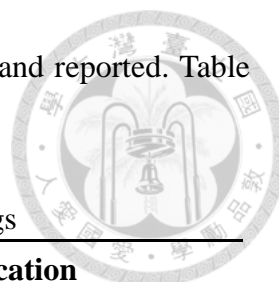


Table 1. Selected Examples of Repositioned Drugs

<b>Drug</b>	<b>Original Indication</b>	<b>New (Potential) Indication</b>
<b>Aspirin</b>	Pain, inflammation	Heart attack Antiplatelet Colon cancer
<b>Bromocriptine</b>	Parkinson's disease	Diabetes
<b>Finasteride</b>	Prostatic hypertrophy	Hair loss
<b>Mifepristone (RU486)</b>	Abortion	Cushing's syndrome Breast cancer
<b>Minoxidil (Rogaine)</b>	Hypertension	Hair loss
<b>Sildenafil (Viagra)</b>	Chest pain (expected)	Erectile dysfunction Pulmonary hypertension Altitude sickness (pulmonary edema)
<b>Thalidomide</b>	Morning sickness	ENL (severe inflammation) Multiple myeloma (blood cancer)

Reference: Ashburn & Thor, 2004; Thomson Reuters, 2012

Several *in silico* methods for drug repositioning have been developed to help medical researchers sift the most plausible drug-disease pairs from a wide range of combinations. Existing methods can broadly be classified into two approaches: literature-based and ontology-based. The literature-based approach analyzes a great size of biomedical literature, e.g., from MEDLINE database, to uncover new, potentially meaningful relationships between drugs and disease. For example, assume that, in biomedical articles, a drug frequently co-occurs with some biomedical concepts (such as enzymes, genes, pathological effects, and proteins) and many of these concepts also frequently co-occur with a disease, where the disease is not the known indication of the focal drug. In this case, it is likely that the disease is a new indication of the focal drug. The above-described methodology is developed by Swanson (1986), who successfully

discovered that fish oil is a treatment for Raynaud's syndrome. In contrast, the ontology-based approach relies on existing ontologies and knowledge bases to discover hidden relationships between drugs and diseases on the basis of the relations between the focal drug and relevant biomedical concepts, and those between these concepts and diseases recorded in the existing ontologies and knowledge bases. For example, Cheng et al. (2012) extracted drug-target interaction network from DrugBank database, and used network topology similarity to infer new targets for known drugs. Figure 1 and Figure 2 are illustrations of the above-mentioned methods.

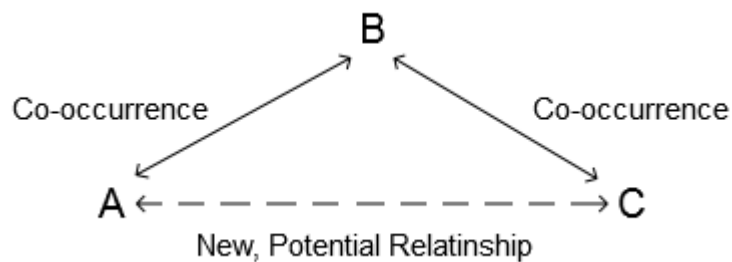


Figure 1. Swanson's literature-based discovering methodology

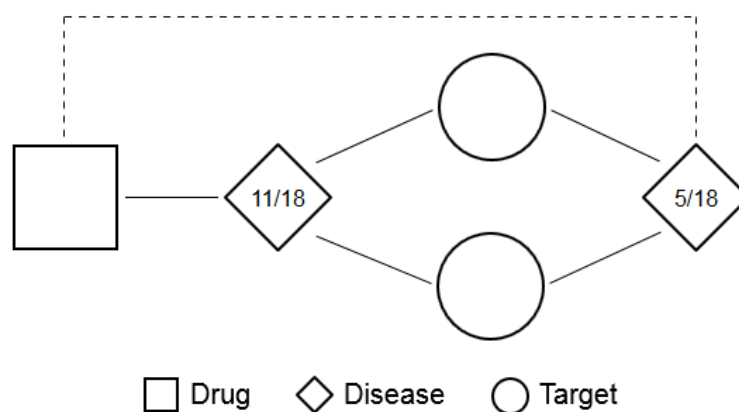
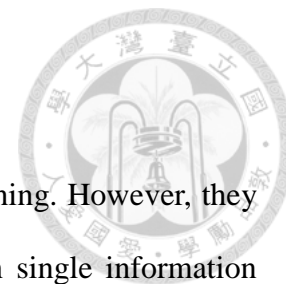


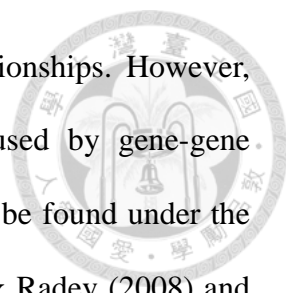
Figure 2. Cheng et al.'s ontology-based network-based inference

## 1.2 Research Motivation and Objective



Existing methods have shown its feasibility for drug repositioning. However, they incur some limitations. First, most previous methods rely only on single information source. The literature-based approach uses only biomedical literature to infer new, potentially meaningful relationships between drugs and diseases, whereas the ontology-based approach depends solely on existing ontologies and knowledge bases. Each information source has its own pros and cons. For example, the biomedical literature has a wider coverage on co-occurrence connections between drugs and relevant concepts and those between concepts and diseases. In contrast, the quality of relations between drugs and relevant concepts and those between concepts and diseases recorded in ontologies and knowledge bases should be higher than that of the co-occurrence connections derived from the biomedical literature. A case in point of this is Thalidomide. Not until 2006 has FDA approved Thalidomide for the treatment for multiple myeloma, which means, they stayed unrelated in most ontologies until then; however, the drug has been highly discussed in literature (and also been marketed) for myeloma treatment since late-1990s. Because existing methods for drug repositioning rely only on single information source for inferences, they cannot have the benefits of different information sources and, at the same time, cannot attempts to mitigate the inherent disadvantages of each information source.

Second, existing literature-based methods, of which follow Swanson's ABC model (Swanson, 1986), consider only single intermediate level, in other words, paths of length 2. For example, take drug concepts as starting terms, genes as intermediate terms, and diseases as target terms, we may find some plausible indirect drug-disease



relationships through combining drug-gene and gene-disease relationships. However, there may be interesting unknown drug-disease relationship caused by gene-gene relationships (i.e., a path of drug-gene-gene-disease), which cannot be found under the original ABC model. Previous studies, such as Özgür, Vu, Erkan, & Radev (2008) and Li, Zhu & Chen (2009), have suggested that gene-gene or protein-protein interactions are important in drug discovery.

In response to the limitations of existing methods, we propose to construct a comprehensive network of biomedical concepts through literature, ontologies and knowledge bases. We then adapt Swanson's undiscovered public knowledge model, also known as the ABC model, for our proposed network to extract plausible drug-disease relationships. Because the nature of links from literature and ontologies are greatly different (the former means co-occurrence while the later means meaningful relation), existing measurements in the ABC model cannot fulfill our need for weighting links from both literature and ontologies, since previous studies mostly based on single information source. We thus propose several algorithms to better assess relationships over our proposed network. Furthermore, we propose to extend the original ABC model to consider paths whose length longer than 2.

The remainder of this thesis is organized as follows. Chapter 2 reviews existing techniques relate to this study, and discuss their limitations to justify our research motivation. In Chapter 3, we describe the design of our proposed drug repositioning discovery method. Chapter 4 reports on our evaluation of proposed techniques. Finally, we conclude our study in Chapter 5 as well as some future research directions.

## Chapter 2 Literature Review



In this chapter, we review existing computational methods related to drug repositioning, which can be classified into two categories: literature-based approach and knowledge-based approach. The literature-based approach identifies plausible drug-disease links by extracting information from academic publications. The ontology-based approach uses existing ontologies or knowledge bases instead, to infer plausible drug-disease links. We briefly summarize the current progress and issues of these existing methods as follows.

### 2.1 Literature-based Approach

Swanson (1986) first introduced the idea of discovering hidden relationships from biomedical literatures in the mid-1980s. He examined across disjoint literatures, manually identified the plausible new connections, and found fish oil might be beneficial to the treatment of Raynaud's syndrome. It was validated by pharmaceutical chemists later. Swanson and Smalheiser (1997) further developed the model he used into a computational method. Figure 3 shows a graphical representation of this model. The basic assumption of Swanson's model is: if a biomedical concept  $A$  relates to concept  $B$ , and concept  $B$  relates to another concept  $C$ , there is a logically plausible relation between  $A$  and  $C$ . For example, if  $A$  is a chemical, and  $C$  is an illness, we may infer a potential new indication of drug  $A$  through this model. It is thus called "ABC model" or "undiscovered public knowledge (UPK) model", and this approach is often referred as literature-based discovery.



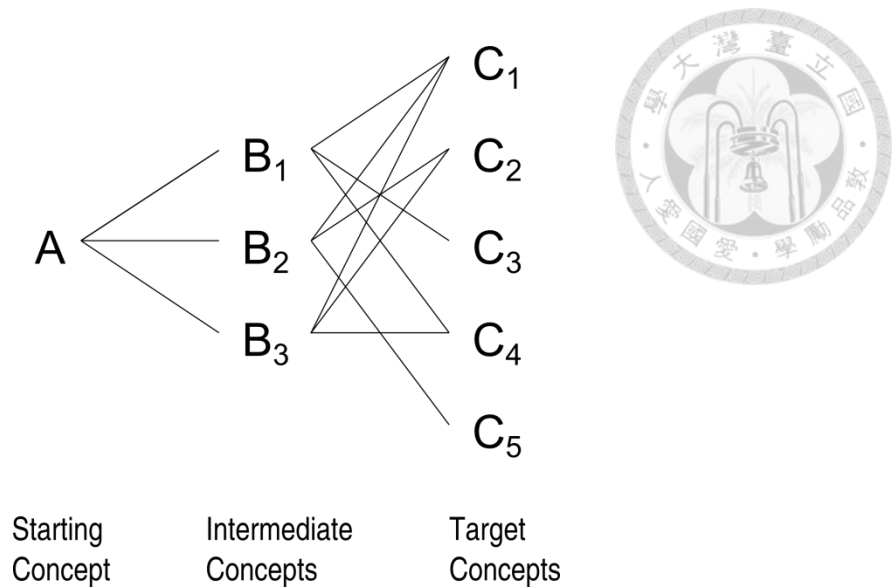
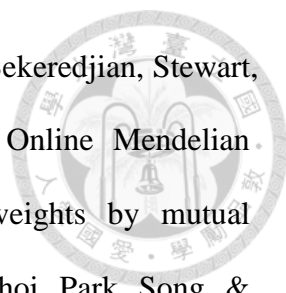


Figure 3. Graphical representation of Swanson's ABC model.

The process of Swanson's methodology can be divided into several steps. First, in *term selection* step, it defines which body of literature shall be extracted as terms (concepts), either words from title, abstract, annotation, or entire document. The second step, *link extraction and filtering*, is to identify relations between concepts. For example, Swanson used co-occurrence analysis to extract relations of concepts. Then, each link is assigned a weight through *link weighting* algorithm, which is often processed along with link extraction. Finally, the system ranks target concepts so that those which are highly relevant to the given starting concept will receive higher ranks in *target term ranking* step. The term selection, link weighting, and target terms ranking are three major research issues of literature-based discovery.

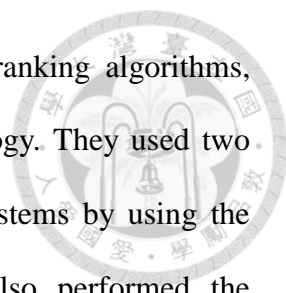
Ever since Swanson's efforts, many other researchers have adapted the ABC model, and developed several improving algorithms and concept extraction techniques. Weber, Klein, de Jong-van den Berg, & Vos (2001) followed Swanson's idea of co-occurrence analysis, while they translated words from titles and abstracts extracted from MEDLINE articles to Unified Medical Language System (UMLS) concepts to filter link



candidates with the help of semantic information. Similarly, Wren, Bekeredjian, Stewart, Shohet, & Garner (2004) mapped full text from articles into Online Mendelian Inheritance in Man (OMIM) concepts. They measured link weights by mutual information between concepts in replace of co-occurrence. Lee, Choi, Park, Song, & Lee (2012) further combined multiple thesauruses to better translate text into biomedical concepts. These researches suggested using full text as the corpus of concept extraction with the help of thesauruses. On the other hand, Srinivasan (2004), Hristovski, Peterlin, Mitchell, & Humphrey (2005), and Yetisgen-Yildiz and Pratt (2006) used Medical Subject Headings (MeSH), keywords annotated to each article in MEDLINE, instead of free text. They applied tf-idf, association rules, and z-score as the measurement of link weights, respectively. All of them reported the metadata-only approach is feasible, though Hristovski et al. noted some shortcoming of using MeSH such as insufficient information of involving genes.

As mentioned, the ABC model has successfully discovered some unknown chemical-disease relationships, including fish oil and Raynaud's syndrome, and magnesium and migraine (Swanson, 1986; 1988). Thus, researchers have suggested applying this approach to drug repositioning. Weeber et al. (2001), Wren et al. (2004), Frijters et al. (2010), and S. Lee et al. (2012) used it to find undiscovered relations between drugs and diseases through selecting different semantic groups of intermediate terms such as adverse effects, genes, and proteins.

As for evaluating the performance of literature-based approach, most researchers use case studies, for instance, replicate historical discoveries, or apply laboratory experiments, to conclude the improvement of their studies. To automatically and

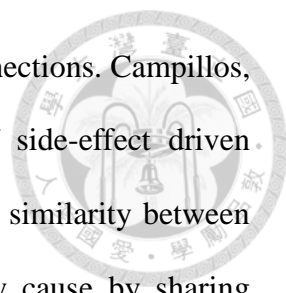


systematically compare different link weighting and target term ranking algorithms, Yetisgen-Yildiz & Pratt (2009) developed an evaluation methodology. They used two literature sets collected from separated time spans, and trained systems by using the older set to predict novel relations in the newer set. They also performed the performance comparison between most of the above-mentioned link weighting algorithms. According to their study, association rules mining seems to have the best performance over tf-idf, mutual information measure, and z-score. They also compared some target terms ranking algorithms and suggested using link term count with average minimum weight.

As mentioned in Chapter 1, most literature-based drug repositioning methods rely on single source. We may improve the performance by considering validated information in ontologies and knowledge bases. Besides, there is still much room for improving the performance of the ABC model itself. There may be other link weighting and target terms ranking algorithms that can boost the accuracy. Previous researches also considered only single intermediate level, while intra-intermediate relations may be important.

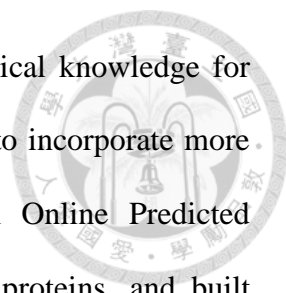
## **2.2 Ontology-based Approach**

Instead of using text from academic publications, ontology-based approach uses several existing ontologies and knowledge bases to help reduce the noisy relations extracted from free text. For example, DrugBank database contains much information of drugs like their indications, mechanisms, adverse effects, related genes and proteins, etc. With such kind of validated information, we can infer undiscovered relations based on their known connections.



Researchers have used different sources to extract possible connections. Campillos, Kuhn, Gavin, Jensen, & Bork (2008) constructed a network of side-effect driven drug-drug relations from UMLS ontology by measuring side-effect similarity between drugs. Assuming that similar side effects of unrelated drugs may cause by sharing common targets, they can be used to predict new drug-target interactions. They experimentally validated some of their results, and thus reported the feasibility of using phenotypic information to infer unexpected biomedical relations. Yang & Agarwal (2011) also based on side effect likelihood between drugs, but they constructed Naïve Bayes models to make predictions. They also took PharmGKB and SIDER knowledge bases, rather than phenotype database, as their information sources. Cheng et al. (2012) built a bipartite network by extracting known drug-target interaction data from DrugBank, and used the network similarity to predict new target of drugs. Li & Lu (2012) built a network similar to Cheng et al.'s work, but added the similarity of drug chemical structure into consideration.

These researches have done many efforts to display the effectiveness to discover unexpected relations based on ontologies and knowledge bases. Nevertheless, due to the carefulness of adding relations, the data set retrieved from ontologies is relatively small. As suggested by Qu, Gudivada, Jegga, Neumann, & Aronow (2009), the prediction of potential new therapeutic indication for drugs requires deep and broad pharmacological and biological knowledge. Therefore, it is important to incorporate more ontologies and knowledge bases together to better predict novel drug-disease relations. In respond to that, Qu et al. and H. S. Lee et al. (2012) both attempted to increase the size and scope of semantic data by constructing integrated network or database of ontologies. However, to our knowledge, few researchers took both ontologies and literature into account,



which may be a good way to acquire deeper and broader biomedical knowledge for making predictions of drug-disease relations. Li et al. (2009) tried to incorporate more knowledge by using protein-protein interactions extracted from Online Predicted Human Interaction Database (OPHID) to expand disease-related proteins, and built disease-specific drug-protein connectivity maps based on literature mining. His work inspires us to build a network over multiple information sources.

As mentioned in Chapter 1, our proposed technique is based on constructing comprehensive network over literature and ontologies, and applies Swanson's ABC model over our network to extract plausible drug-disease relationships and thus taking both literature and ontologies into account.

## Chapter 3

# Design of Drug Repositioning Discovery Method



As mentioned previously, we propose a drug repositioning discovery method based on Swanson's hidden relationship discovering model (ABC model) that takes both biomedical literature and existing ontologies and knowledge bases into account by constructing a comprehensive network of biomedical concepts. As Figure 4 illustrates, our method consists of five main phases: literature-based concept network construction, ontology-based concept network construction, related concept retrieval, link weighting, and target term ranking. The literature-based concept network construction phase extracts and filters biomedical concepts from the literature database (i.e., MEDLINE) and constructs the network via association rules mining, as suggested by Yetisgen-Yildiz & Pratt (2009). The ontology-based concept network construction phase extracts known relations between biomedical concepts from existing ontologies and knowledge bases as concept network links. Subsequently, we construct a comprehensive network of biomedical concepts. Given a specific drug, we retrieve a subgraph of related concepts from our comprehensive network. Depending on single or multiple intermediate levels, we apply different constraints in the related concepts retrieval phase. We then employ *Extended Normalized MEDLINE Similarity* algorithm to weight each link in the retrieved subgraph, either the link is from literature-based or ontology-based network. Finally, we rank target terms extracted through our discovering model in order to identify plausible novel drug-disease relationships. For single intermediate level scenario, we propose and employ two target term ranking algorithms, *Summation of*

*Minimum Weight and Summation of Average Weight*; for multiple intermediate levels scenario, we employ the *Katz measure*.

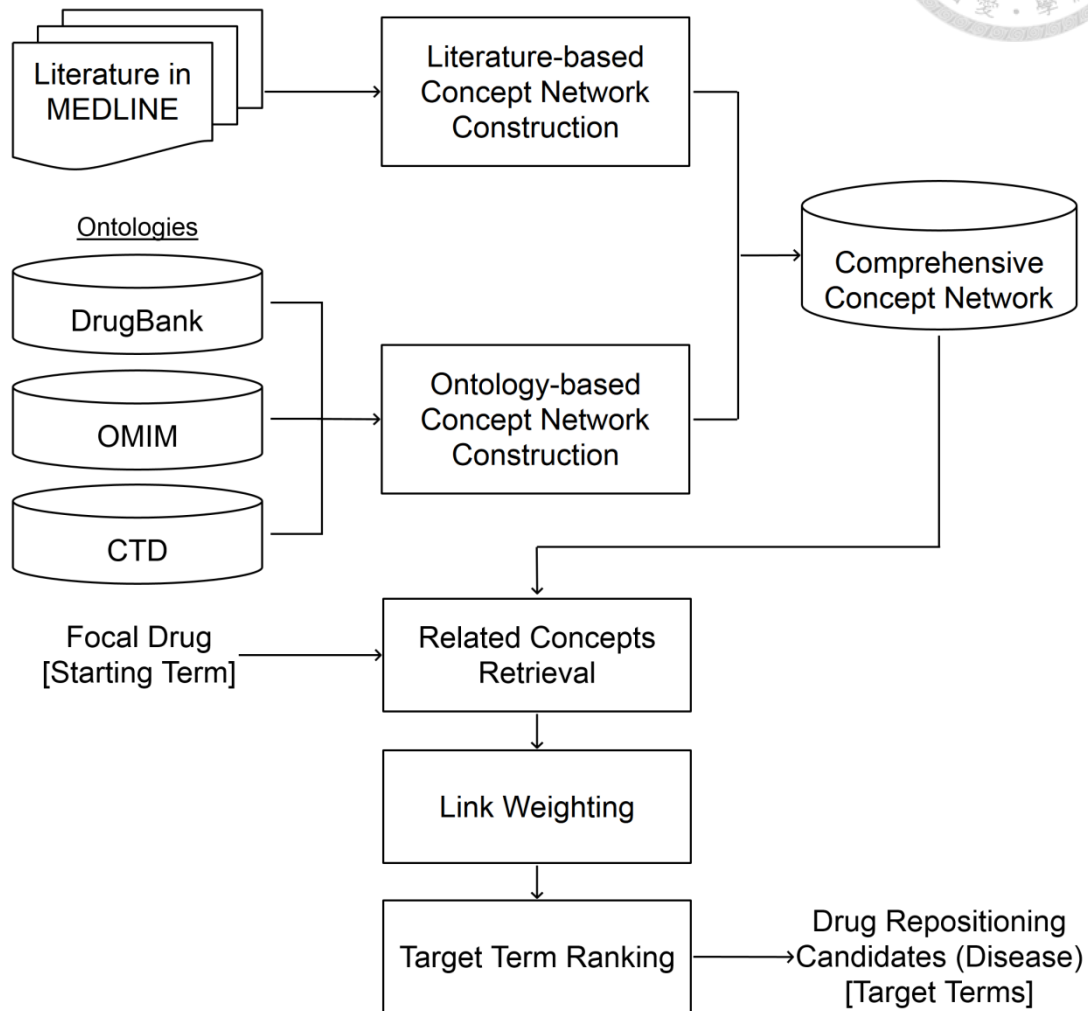


Figure 4. Overall Process of Our Drug Repositioning Discovery Method

### 3.1 Literature-based Concept Network Construction

As mentioned, the purpose of the literature-based concept network construction phase is to extract biomedical concepts from the literature and to construct the literature-based concept network via association rules mining.



### 3.1.1. Data Collection

We use MEDLINE database as our literature data source, which is constructed by U.S. National Library of Medicine (NLM). Specifically, the database we adopt is MEDLINE 2011 baseline. It contains 19,680,423 biomedical articles until 2010. For each article, NLM indexes the publication type of each document, such as newspaper, clinical trial report, journal article or guideline. Among the 61 publication types shown in MEDLINE 2011 baseline, we remove the publication types that are suggested as less relevant to literature-based discovery in previous studies (Yetisgen-Yildiz & Pratt, 2009), as shown in Table 2. As a result, our literature database consists of 18,712,338 biomedical articles.

Table 2. Excluded Publication Types

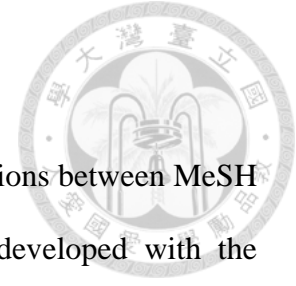
Addresses	Directory	Letter
Bibliography	Editorial	News
Biography	Guidelines	Newspaper article
Comment	Lectures	Patient education handout
Congresses	Legal ceases	Periodical index
Dictionary	Legislation	Practical guideline

NLM also indexes representative medical terms discussed in each biomedical article in MEDLINE into corresponding MeSH terms, which are controlled vocabulary maintained by NLM for the purpose of annotation. Except some articles that are not indexed with any annotation, the number of MeSH terms per MEDLINE article range from 1 to 97, and its average is 9.44.

In this study, we use MeSH terms rather than words from title or abstract of each biomedical article as input to the construction of literature-based concept network, as suggested by previous studies (Srinivasan, 2004; Yetisgen-Yildiz & Pratt, 2009).



### 3.1.2. Link Extraction and Filtering



We apply association rules mining approach to extract the relations between MeSH terms in MEDLINE articles. Association rules were originally developed with the purpose of market basket analysis, which is, to find two sets of items that are tend to be purchased together. Hristovski et al. (2005) first adapted association rules to identify the correlated biomedical terms. In their application, transactions are documents and items are terms. Thus, the two important measures for an association rule are defined as follows:

*Support:* Two biomedical terms  $A$  and  $B$  are correlated if they co-occur together in many documents.

$$s = |D_A \cap D_B|$$

*Confidence:*  $A$  and  $B$  are correlated if the percentage of documents containing  $B$  within all documents containing  $A$  is high.

$$c = \frac{|D_A \cap D_B|}{|D_A|}$$

where  $D_A$  is the set of documents in which term  $A$  appears and  $D_B$  is the set of documents that include term  $B$ .

Hristovski et al. suggested setting thresholds on support and confidence for limiting the number of related concepts and improving the effectiveness of mining. We follow Yetisgen-Yildiz & Pratt's experiment (2009) by setting the minimum support threshold as 2.6 and the minimum confidence threshold as 0.0055. Accordingly, the set of rules that pass the threshold test are used to extract the relations between MeSH terms in literature.

Additionally, since we focus on drug repositioning, we limit terms to be identified as correlated must be within some specific semantic groups, such as drugs, genes, proteins, enzymes, pathological effects, and diseases. Each MeSH term is organized in 16 categories, and each category is further divided into subcategories. We select several MeSH subcategories that can represent our specified semantic meanings. The subcategories we select are shown in Table 3. After filtering, we extract 12,278 MeSH terms and 2,623,222 relations from literature.

Table 3. Selected MeSH Subcategories

<b>Semantic Group</b>	<b>Corresponding MeSH Subcategories</b>
Drugs	D01-D05, D09, D10, D20, D26, D27
Genes, Proteins, and Enzymes	D06, D08, D12, D13, D23
Pathological Effects	G03-G16
Diseases	C01-C23

## 3.2 Ontology-based Concept Network Construction

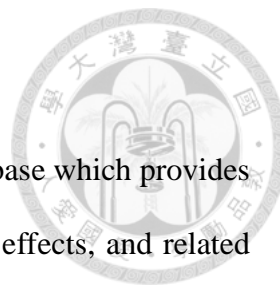
The purpose of this phase is to extract relations from ontologies. There are several ontologies and knowledge bases which record semantic relations between biomedical concepts. The relations depicted in these ontologies and knowledge bases are known and validated. Therefore, the credibility of the ontology-based network should be higher than literature-based network.

### 3.2.1. Data Collection

In this study, the ontologies and knowledge bases we adopt are DrugBank, Online Mendelian Inheritance in Man (OMIM), and Comparative Toxicogenomics Database (CTD).

## **DrugBank**

DrugBank (<http://www.drugbank.ca/>) is a richly annotated database which provides extensive information about targets, pathways, indications, adverse effects, and related proteins of various drugs (Knox, et al., 2011). It contains 6,811 drugs entries including 1,678 FDA-approved drugs and 5,080 experimental drugs. We use its drug-target interactions data to build our ontology-based network. The number of drug-target interactions we collect from DrugBank is 14,542.




## **OMIM**

OMIM is a comprehensive and authoritative knowledgebase of human genes and genetic phenotypes (Hamosh, Scott, Amberger, Bocchini, & McKusick, 2005). It is written and edited by scientists and physicians around the world. OMIM is freely available at <http://www.omim.org/>. The knowledgebase contains 4,380 manually annotated gene-disease relations, which we use as inputs to the construction of our ontology-based concept network.

## **CTD**

CTD (<http://ctdbase.org/>) is a database that integrates data from scientific literature to describe chemical interactions with genes and proteins, and diseases and genes/proteins, and others (Davis, et al., 2013). These relationships are manually curate by biocurators. According to CTD's own statistical report, the database contains 869,902 curated chemical-gene interactions and 27,397 gene-disease associations with direct evidences. We extract these two categories of relations.

### 3.2.2. Concept Mapping



There are several problems in combining multiple databases. First, different ontologies have different terminology and codification. DrugBank use UniProt as its protein name thesaurus, while OMIM has his own naming; besides, not all database providers offer mapping between themselves and external databases. Second, the definition for a same biomedical concept may be various in different ontologies. For example, Alzheimer disease is defined to 17 concepts in OMIM, but not in other ontologies.

In order to unify concepts from different ontologies and to integrate their relations with literature-based concept network, we decide to map all retrieved terms into MeSH terms. NLM provides MeSH Supplementary Concept Records (SCRs), which are designed to extend the search terms for NLM's PubMed search engine. SCRs contain mapping between some OMIM terms and MeSH terms. Also, UniProt database provides mapping between UniProt protein entries and OMIM terms. Furthermore, CTD contains a vast amount of chemical, disease, and gene synonyms and their mapping between OMIM and MeSH terms. With the help of above-mentioned information, we can translate terms from different codifications into MeSH terms; however, due to the complication of translation, some relations lost within the translating process.

### 3.2.3. Link Extraction and Filtering

We consider the relations retrieved from ontologies as credible. Thus, we only limit that relations must be between two MeSH terms within our specified MeSH subcategories, as shown in Table 5 previously. As a result, we retrieved 7,808 relations

for translated MeSH terms from DrugBank, 2,404 relations from OMIM, and 195,033 relations from CTD database.



### 3.3 Related Concept Retrieval

In the previous two phases, we have constructed our comprehensive biomedical concept network. Given a focal drug, we can retrieve all related concepts from the comprehensive network, including intermediate terms and target terms (in other words, plausible related diseases). These concepts form a subgraph which becomes the input to the following two phases. In response to our research objective, we consider two different scenarios in the related concept retrieval phase: single intermediate level (i.e., considering only paths of length 2), and multiple intermediate levels (i.e., considering paths of length longer than 2).

The single intermediate level scenario is to simply retrieve concepts related to the given drug in the network as *intermediate terms*. Then, we extract the disease concepts that related to these intermediate terms but not related to the given drug. These disease concepts are defined as *target terms*. This scenario is similar to the original ABC model which we have shown in Figure 3 in Section 2.1.

The multiple intermediate levels scenario is to consider paths of length longer than 2. For those longer paths, we add two constraints to our retrieval model. First, we limit the *intra-intermediate* relations must be between two terms of the same semantic group (as we defined in Table 3), such as protein-protein relationships. The purpose of this constraint is to make intra-intermediate relations more meaningful since we are only interested in drug repositioning candidates. The second constraint is that the number of

neighbors of intra-intermediate terms should be less than a threshold. This is meant to avoid popular terms that may lead to noisy long paths. Table 4 shows the top 10 most connected intermediate terms in our evaluation (described in Section 4.4). As it shows, these terms are likely to be general terms. Figure 5 and Figure 6 are illustrations of our constraints of multiple intermediate levels scenario.

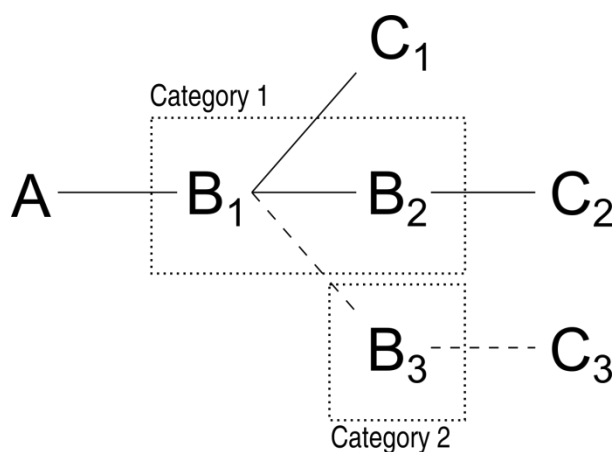


Figure 5. Illustration of Constraint to Category of Intra-intermediate Terms

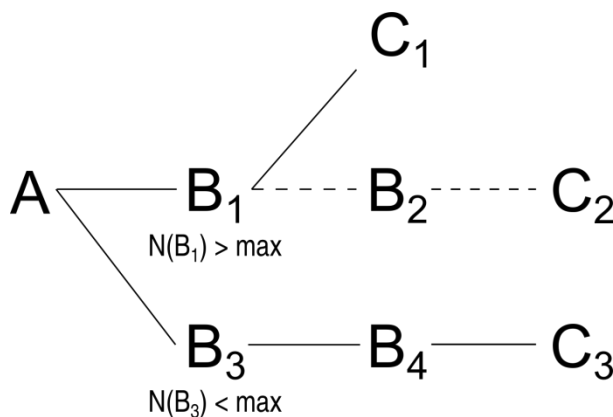


Figure 6. Illustration of Threshold to Number of Neighbors

Table 4. Top 10 Most Connected Intermediate Terms in Our Evaluation

Carrier Proteins	(4631)	DNA Primers	(4138)
Tissue Distribution	(4526)	Biological Transport	(3993)
Membrane Proteins	(4261)	Recombinant Fusion Proteins	(3877)
Drug Synergism	(4181)	Antibodies	(3791)
Peptide Fragments	(4156)	Cell Survival	(3789)

### 3.4 Link Weighting

The purpose of link weighting phase is to weight each link in the retrieved subgraph of related concepts, either the link is from literature or ontologies. Our weighting should reflect the following facts: first, links from ontologies are validated as related; second, weights of links from literature are correlated to its degree of co-occurrence in literature. Therefore, we develop a similarity measure on the basis of *Normalized Google Distance* (Cilibrasi & Vitányi, 2007) as our link weighting algorithm.

Normalized Google Distance (NGD) is an approximation to Normalized Information Distance (NID). NID expresses the similarity between two terms on a scale from 0 to 1, in which 0 being the same and 1 being completely different. Cilibrasi & Vitányi developed NGD as an implementation of NID by using pages indexed by Google as text corpus. The NGD is computed as follows:

$$NGD(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

where  $f(x)$  denotes the number of pages containing  $x$ ,  $f(x, y)$  denotes the number of pages containing both  $x$  and  $y$ , and  $M$  is total number of pages indexed by Google. The range of NGD is in between 0 and infinity, where  $NGD > 1$  is semantically identical to  $NID = 1$ .

Since the purpose of NGD is to measure the similarity of two terms from the given corpus, we can adapt NGD to MEDLINE as our similarity measurement. Lu and Wilbur (2009) also adapted NGD to MEDLINE and showed the feasibility to identify related queries in PubMed search engine. Besides, Lindsey, Veksler, Grintsveyg, & Gray (2007)

compared NGD and Pointwise Mutual Information, and reported NGD has better performance under different corpora.



We thus define our similarity measure as follows:

$$NMD(A, B) = \frac{\max\{\log|D_A|, \log|D_B|\} - \log|D_A \cap D_B|}{\log M - \min\{\log|D_A|, \log|D_B|\}}, NMD = 1 \text{ if } NMD > 1$$

$$NMS(A, B) = 1 - NMD(A, B)$$

$$Sim(A, B) = \begin{cases} NMS(A, B), & \text{if } (A, B) \in \text{Literature} \wedge \notin \text{Ontology} \\ 1, & \text{if } (A, B) \in \text{Ontology} \end{cases}$$

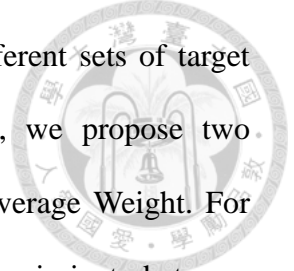
where  $D_A$  is the set of articles that include MeSH term  $A$ , and  $M$  denotes the total number of articles in MEDLINE. This similarity measure range from 0 to 1, in which 0 being completely unrelated and 1 being credibly related. If link  $(A, B)$  is from literature-based network and is not find in ontology-based network, we weight it by calculating its *Normalized MEDLINE Distance* and subtracting from 1, called *Normalized MEDLINE Similarity (NMS)*; otherwise, if the link is from the ontology-based network, we assign its weight as 1 since the relation is validated. We called our weighting as *Extended Normalized MEDLINE Similarity (Extended NMS)*. Accordingly, we weight each link in the retrieved subgraph of related concepts by Extended NMS.

### 3.5 Target Term Ranking

In this phase, we rank target terms extracted through our discovering model, according to our retrieved subgraph of related concepts and weights of links, in order to identify plausible novel drug-disease relationships. As described in Section 3.3, we consider two scenarios, single intermediate level and multiple intermediate levels, and



apply different constraints over them. Accordingly, we employ different sets of target term ranking algorithms. For single intermediate level scenario, we propose two algorithms, Summation of Minimum Weight and Summation of Average Weight. For multiple intermediate levels scenario, we apply Katz measure to discriminate between longer and shorter paths.



### 3.5.1 Single Intermediate Level Scenario

Yetisgen-Yildiz & Pratt (2009) suggested using *Link Term Count* with *Average Minimum Weight (LTC-AMW)* to have the best performance. In short, LTC-AMW takes the number of intermediate terms between starting term and target term as the major measurement, which is, the number of paths. The average minimum weight of paths only used when two target terms are same in their number of paths. The assumption of LTC-AMW is that all paths are equally important, which may not be precise if we have proper measurement for weights of paths. Since we have developed Extended NMS to measure the degree of relative for each link in Section 3.4, we wish to consider both number of paths and weights of paths in a same measure. Therefore, we propose two target term ranking algorithms as follows:

Summation of Minimum Weight (Sum\_MW): the information in each path is measured by the least information of internal edges in path.

$$Score(A, C) = \sum_{B \in N(A) \cap N(C)} \min\{Wt(A, B), Wt(B, C)\}$$

Summation of Average Weight (Sum\_AW): the information in each path is measured by the average information of internal edges in path.

$$Score(A, C) = \sum_{B \in N(A) \cap N(C)} \frac{Wt(A, B) + Wt(B, C)}{2}$$

where  $N(A)$  denotes the neighbor concepts of term  $A$ , and  $Wt(A, B)$  is the weight of link between  $A$  and  $B$ . The above algorithms differentiate the importance of each path according to their minimum or average weight of internal edges, and assign ranking score to each target term according to the cumulative information of all paths between the starting term and the target term. We then order target terms according to their scores.

### 3.5.2 Multiple Intermediate Levels Scenario

As we consider paths of length longer than 2, previous studies suggest the longer the transitivity inference is, the less likely the source concept is related to the target concept (Liben-Nowell & Kleinberg, 2007). Katz (1953) defines a measure that sums over the paths between two nodes, exponentially weighted by length and thus gives more weights to shorter paths. We apply its *Katz measurement* as our target term ranking algorithm for multiple intermediate levels scenario. Accordingly, the Katz measurement is defined as:

$$Score(A, C) = \sum_{\ell=2}^L \beta^{\ell} \cdot |paths_{A,C}^{\langle \ell \rangle}|$$

where  $paths_{A,C}^{\langle \ell \rangle}$  is the set of all length- $\ell$  paths between  $A$  and  $C$ , and  $\beta > 0$  is a parameter of the predictor.

## Chapter 4 Evaluation and Results



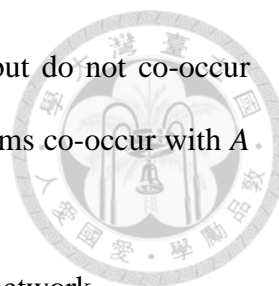
In this chapter, we describe the design of our evaluation, and then discuss our evaluation results. We design three experiments. The first experiment is to evaluate our proposed comprehensive network and link weighting algorithm. The second experiment is to evaluate our proposed target term ranking algorithms for single intermediate level scenario. Finally, we evaluate the performance of multiple intermediate levels scenario.

### 4.1 Evaluation Design

We follow the evaluation procedure proposed by Yetisgen-Yildiz & Pratt (2009). Specifically, we describe our experiment procedure step by step as follows:

Given a starting term (i.e., drug)  $A$ :

1. We set cut-off date as January 1, 2000 and divide MEDLINE 2011 baseline into:
  - a. *Pre-cut-off set* ( $S_{t1}$ ) which includes documents prior to 1/1/2000.
  - b. *Post-cut-off set* ( $S_{t2}$ ) which includes documents after 1/1/2000.
2. We use documents in the pre-cut-off set along with ontologies as the input to construct our comprehensive concept network.
3. We create a gold-standard set  $G_A$ , which contains terms that satisfied following rules:
  - a. Terms are within our specified target semantic group, i.e., disease.



- b. Terms that co-occur with  $A$  in the post-cut-off set, but do not co-occur with  $A$  in the pre-cut-off set. In other words, these terms co-occur with  $A$  in literature only after the cut-off date.
  - c. Terms are not related with  $A$  in our ontologies-based network.
4. We calculate the overall performance using the information retrieval metrics:

$$\text{Precision: } P_A = \frac{|T_A \cap G_A|}{|T_A|}$$

$$\text{Recall: } R_A = \frac{|T_A \cap G_A|}{|G_A|}$$

where  $T_A$  is the set of target terms generated by our discovery method.

Table 5 includes the list of semantic groups that we used in our experiments. For performance benchmark, we randomly select 100 terms from the semantic group of drugs as starting terms, i.e., focal drugs.

Table 5. Semantic Groups Selected for Our Experiments

Intermediate Term Selection	Target Term Selection
Drugs	Diseases
Genes, Proteins, and Enzymes	
Pathological Effects	
Diseases	

## 4.2 Experiment 1:

### Comprehensive Network and Link Weighting Algorithm

In this experiment, we compare: (1) the performance of our proposed link weighting algorithm, Extended NMS, with the algorithm suggested by previous researches, and (2) the result extract through our proposed comprehensive biomedical concept network with information sources used in previous studies.

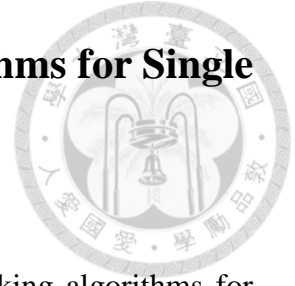
We employ single intermediate level scenario in this experiment. The performance benchmark is the original ABC model over only literature which uses association rules as link weighting algorithm. We evaluate three sets of result for our discovering model, one is over only literature-based network, another one is over only ontology-based network, and the third one is over the comprehensive network. Both our model and benchmark model apply LTC-AMW as target term ranking algorithm. Table 6 shows the evaluation results.

Table 6. Evaluation Results of Our Link Weighting and Comprehensive Network

<b>Recall</b>	<b>Association Rules (Literature)</b>	<b>NMS (Literature)</b>	<b>Extended NMS (Ontology)</b>	<b>Extended NMS (Integrated)</b>
<b>0%</b>	62.61%	57.72%	39.01%	59.33%
<b>10%</b>	29.72%	29.93%	20.16%	30.54%
<b>20%</b>	22.07%	23.75%	15.96%	23.89%
<b>30%</b>	17.80%	18.95%	14.22%	19.01%
<b>40%</b>	15.13%	16.27%	12.58%	16.26%
<b>50%</b>	11.73%	13.80%	12.25%	13.62%
<b>60%</b>	9.52%	11.69%	13.77%	11.53%
<b>70%</b>	7.61%	9.66%	0%	9.61%
<b>80%</b>	7.17%	7.76%	0%	7.69%
<b>90%</b>	2.31%	6.01%	0%	5.97%
<b>100%</b>	0.60%	3.80%	0%	3.84%
<b>AUC-PR</b>	<b>15.47%</b>	<b>16.86%</b>	<b>10.84%</b>	<b>16.97%</b>

As shown above, our proposed Extended NMS outperforms the benchmark link weighting algorithms, association rules, in both literature and integrated information sources. Also, using both literature and ontologies as information sources would improve overall performance, especially precisions on higher ranks. This would better help researchers sift plausible drug-disease relations for the purpose of drug repositioning.

## 4.3 Experiment 2: Target Term Ranking Algorithms for Single Intermediate Level Scenario



In this experiment, we evaluate our proposed target term ranking algorithms for single intermediate level scenario, Summation of Minimum Weight (Sum\_MW) and Summation of Average Weight (Sum\_AW).

The benchmark algorithm we use is LTC-AMW. We apply Extended NMS as our link weighting algorithm in this experiment since we have shown that our Extended NMS outperforms association rules. To detail the performances under different information sources, we employ two sets of evaluation, one being using only literature, and another using comprehensive network, which is, both literature and ontologies.

Table 7. Comparison of Target Term Ranking Algorithms (Literature-only)

<b>Recall</b>	<b>LTC-AMW</b>	<b>Sum_MW</b>	<b>Sum_AW</b>
<b>0%</b>	57.72%	55.85%	58.70%
<b>10%</b>	29.93%	31.70%	32.32%
<b>20%</b>	23.75%	23.97%	24.45%
<b>30%</b>	18.95%	20.30%	20.80%
<b>40%</b>	16.27%	17.29%	17.28%
<b>50%</b>	13.80%	14.66%	14.84%
<b>60%</b>	11.69%	12.58%	12.79%
<b>70%</b>	9.66%	10.43%	10.41%
<b>80%</b>	7.76%	8.36%	8.40%
<b>90%</b>	6.01%	6.30%	6.33%
<b>100%</b>	3.80%	3.80%	3.80%
<b>AUC-PR</b>	<b>16.86%</b>	<b>17.54%</b>	<b>17.89%</b>

Table 8. Comparison of Target Term Ranking (Multiple Sources)

<b>Recall</b>	<b>LTC-AMW</b>	<b>Sum_MW</b>	<b>Sum_AW</b>
<b>0%</b>	59.33%	59.14%	61.70%
<b>10%</b>	30.54%	33.46%	33.16%
<b>20%</b>	23.89%	24.86%	24.52%
<b>30%</b>	19.01%	20.91%	20.69%
<b>40%</b>	16.26%	17.32%	17.01%
<b>50%</b>	13.62%	14.74%	14.56%
<b>60%</b>	11.53%	12.42%	12.17%
<b>70%</b>	9.61%	10.50%	10.29%
<b>80%</b>	7.69%	8.41%	8.22%
<b>90%</b>	5.97%	6.35%	6.22%
<b>100%</b>	3.84%	3.84%	3.84%
<b>AUC-PR</b>	<b>16.97%</b>	<b>18.05%</b>	<b>17.96%</b>

Table 7 shows the result of using only literature as the information source, and Table 8 shows the result of using our comprehensive network as the information source. Both Sum\_MW and Sum\_AW outperform the benchmark algorithm, LTC-AMW. These results show that our link weighting algorithm, Extended NMS, is a more effective measure to weight paths, and considering both number and weights for paths between starting term and target terms can improve the effectiveness of discovery.

We further compare Sum\_MW with Sum\_AW. Sum\_AW performs better in using only literature-based network as the information source, while Sum\_MW performs slightly better when using the comprehensive network. We think this may lead by some parsing error in our concept mapping process (as we described in Section 3.2.2).

Overall, as we show in experiment 1 and experiment 2, using our proposed comprehensive concept network as information source can improve the effectiveness of predicting plausible drug-disease relations. Furthermore, our proposed link weighting

and target term ranking algorithms all outperform existing algorithms.



## 4.4 Experiment 3: Multiple Intermediate Levels Scenario

In this experiment, we evaluate our discovering model under multiple intermediate levels scenario. We consider two settings of path length,  $2 \leq \ell \leq 3$  and  $2 \leq \ell \leq 4$ , and compare both of them with the benchmark setting,  $\ell = 2$ . As suggested in the previous experiments, we use our comprehensive network as information source, and Extended NMS as link weighting algorithm. Our target term ranking algorithm is Katz measurement. There are two parameters require tuning: the threshold number of neighbors of intra-intermediate terms (*max\_neighbor*), and  $\beta$  of Katz measure.

### 4.4.1 Parameter Tuning

We set  $\beta = 0.05$  as default for tuning *max\_neighbor*, a value suggested in previous study (Liben-Nowell & Kleinberg, 2007). After that, we examine if our default  $\beta$  is optimal. We apply different sets of parameters to  $2 \leq \ell \leq 3$  and  $2 \leq \ell \leq 4$ , and show their tuning processes as follows.



Table 9. Tuning of  $max\_neighbor$  ( $2 \leq \ell \leq 3$ )

Recall	$N(B) < 1250$	$N(B) < 1000$	$N(B) < 750$	$N(B) < 500$	$N(B) < 250$
0%	60.82%	60.95%	63.39%	62.17%	59.38%
10%	29.88%	31.16%	31.81%	32.71%	30.84%
20%	22.29%	23.27%	24.34%	24.79%	24.00%
30%	18.26%	19.35%	19.99%	19.89%	19.14%
40%	15.27%	16.12%	16.83%	17.03%	16.40%
50%	12.70%	13.40%	13.96%	14.11%	13.65%
60%	10.88%	11.27%	11.84%	11.95%	11.58%
70%	9.07%	9.49%	9.86%	9.94%	9.63%
80%	7.29%	7.53%	7.76%	7.89%	7.70%
90%	5.71%	5.80%	5.96%	6.02%	5.96%
100%	3.58%	3.58%	3.59%	3.66%	3.78%
<b>AUC-PR</b>	<b>16.36%</b>	<b>16.97%</b>	<b>17.58%</b>	<b>17.72%</b>	<b>17.05%</b>

Table 10. Tuning of  $\beta$  for Katz measurement ( $2 \leq \ell \leq 3$ )

Recall	$\beta = 0.05$	$\beta = 0.01$	$\beta = 0.005$
0%	62.17%	60.27%	59.47%
10%	32.71%	31.00%	30.66%
20%	24.79%	24.17%	24.01%
30%	19.89%	19.24%	19.12%
40%	17.03%	16.52%	16.37%
50%	14.11%	13.71%	13.63%
60%	11.95%	11.63%	11.57%
70%	9.94%	9.68%	9.62%
80%	7.89%	7.73%	7.70%
90%	6.02%	5.95%	5.95%
100%	3.66%	3.66%	3.66%
<b>AUC-PR</b>	<b>17.72%</b>	<b>17.16%</b>	<b>17.02%</b>

As Table 9 shows, when  $max\_neighbor$  is set to 500, the performance is the best among others when considering paths of length no longer 3. We further examine the  $\beta$  of Katz measure as Table 10 shows, and find that 0.05 is the optimal value for  $\beta$ .

Table 11. Tuning of  $max\_neighbor$  ( $2 \leq \ell \leq 4$ )

Recall	$N(B) < 750$	$N(B) < 500$	$N(B) < 250$
0%	60.32%	60.20%	59.22%
10%	28.90%	30.62%	30.73%
20%	21.70%	23.12%	23.94%
30%	17.55%	18.90%	19.20%
40%	14.84%	15.84%	15.96%
50%	12.40%	13.03%	13.32%
60%	10.51%	11.03%	11.43%
70%	8.65%	9.13%	9.47%
80%	7.02%	7.35%	7.62%
90%	5.46%	5.66%	5.85%
100%	3.58%	3.58%	3.62%
<b>AUC-PR</b>	<b>15.90%</b>	<b>16.66%</b>	<b>16.90%</b>

Table 12. Tuning of  $\beta$  for Katz measurement ( $2 \leq \ell \leq 4$ )

Recall	$\beta = 0.05$	$\beta = 0.01$	$\beta = 0.005$
0%	59.22%	59.25%	58.84%
10%	30.73%	30.31%	30.42%
20%	23.94%	23.86%	23.84%
30%	19.20%	19.09%	19.00%
40%	15.96%	16.13%	16.17%
50%	13.32%	13.47%	13.53%
60%	11.43%	11.52%	11.51%
70%	9.47%	9.51%	9.53%
80%	7.62%	7.64%	7.66%
90%	5.85%	5.92%	5.93%
100%	3.62%	3.62%	3.62%
<b>AUC-PR</b>	<b>16.90%</b>	<b>16.89%</b>	<b>16.88%</b>

For considering paths of length no longer than 4, the performance top at  $N(B) < 250$ . We also examine the  $\beta$  of Katz measure, and 0.05 remains optimal for  $\beta$ . Table 11 and Table 12 show our tuning processes and results.

Accordingly, we set  $max\_neighbor$  as 500 and  $\beta$  as 0.05 when considering paths of length no longer than 3, and set  $max\_neighbor$  as 250 and  $\beta$  as 0.05 when considering paths of length no longer than 4.



#### 4.4.2 Experiment Result

Table 13. Comparison of Single and Multiple Intermediate Levels

<b>Recall</b>	$\ell = 2$	$2 \leq \ell \leq 3$	$2 \leq \ell \leq 4$
<b>0%</b>	59.33%	62.17%	59.22%
<b>10%</b>	30.54%	32.71%	30.73%
<b>20%</b>	23.89%	24.79%	23.94%
<b>30%</b>	19.01%	19.89%	19.20%
<b>40%</b>	16.26%	17.03%	15.96%
<b>50%</b>	13.62%	14.11%	13.32%
<b>60%</b>	11.53%	11.95%	11.43%
<b>70%</b>	9.61%	9.94%	9.47%
<b>80%</b>	7.69%	7.89%	7.62%
<b>90%</b>	5.97%	6.02%	5.85%
<b>100%</b>	3.84%	3.66%	3.62%
<b>AUC-PR</b>	<b>16.97%</b>	<b>17.72%</b>	<b>16.90%</b>

Table 13 shows our evaluation result to multiple intermediate levels scenario. As shown, considering paths of length no longer than 3 does improve the performance, while further considering paths of length 4 do not improve. This result is consistent to that of previous studies which suggest that intra-intermediate relations, such as gene-gene and protein-protein interactions, are important in drug repositioning discovery; meanwhile, the transitive inference decays more than our assumption. As a result, we conclude that appropriately considering paths of length longer than 2 can make better inferences in drug repositioning discovery.

We further compare the performances between using our comprehensive network as information source and using only literature-based network in order to justify our assumption of this experiment. We set  $\beta$  as 0.05 and consider paths of length no longer than 3, as suggested above. In previous sections, we conclude that using both literature and ontologies as information sources outperforms using only literature.

The comparison is shown in Table 14, and the result under multiple intermediate levels scenario is consistent to our previous conclusion, that is to say, using our comprehensive network as information source under both single and multiple intermediate levels scenarios can improve the effectiveness of predicting plausible drug-disease relations.

Table 14. Comparison of Using Different Information Sources ( $2 \leq \ell \leq 3$ )

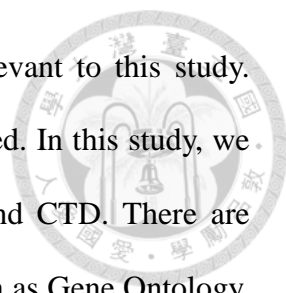
<b>Recall</b>	<b>Integrated (<math>N(B) &lt; 500</math>)</b>	<b>Literature (<math>N(B) &lt; 250</math>)</b>	<b>Literature (<math>N(B) &lt; 500</math>)</b>	<b>Literature (<math>N(B) &lt; 750</math>)</b>
<b>0%</b>	62.17%	58.64%	60.07%	61.31%
<b>10%</b>	32.71%	31.75%	32.24%	31.55%
<b>20%</b>	24.79%	24.44%	24.59%	23.87%
<b>30%</b>	19.89%	19.63%	19.97%	20.07%
<b>40%</b>	17.03%	16.74%	17.03%	16.71%
<b>50%</b>	14.11%	14.11%	14.27%	13.96%
<b>60%</b>	11.95%	11.94%	12.02%	11.79%
<b>70%</b>	9.94%	9.86%	9.97%	9.85%
<b>80%</b>	7.89%	7.88%	7.94%	7.78%
<b>90%</b>	6.02%	6.03%	6.03%	5.92%
<b>100%</b>	3.66%	3.63%	3.60%	3.56%
<b>AUC-PR</b>	<b>17.72%</b>	<b>17.35%</b>	<b>17.59%</b>	<b>17.39%</b>

## Chapter 5 Conclusion and Future Work



Drug repositioning can reduce significant time and spending in comparison with *de novo* drug development and can also create opportunities for pharmaceutical companies to make full use of their intellectual property portfolio. Researchers have developed several automated methods to help discover these hidden drug-disease relationships. However, previous studies mostly rely on single information source, either literature or ontologies. Also, previous proposed methods that rely on literature do not consider multiple intermediate levels.

In this study, we develop a drug repositioning discovery method that uses both biomedical literature and ontologies as information sources by constructing a comprehensive network of biomedical concepts. Based on Swanson's ABC model, we extend it to consider multiple intermediate levels, and propose several algorithms for better assessing relations in our network. We experimentally evaluate our proposed method, and show that taking both literature and ontologies into account can improve the effectiveness of predicting novel drug-disease relationships. Also, we develop a similarity measurement, Extended NMS, that can assign unified weight to links from literature and ontologies, and it outperforms existing link weighting techniques. Besides, our proposed target term ranking algorithms can better infer plausible drug-disease relations over our weighting and integrated information source. Furthermore, we show that considering paths of length no longer than 3 can make better predictions in comparison with considering only single intermediate level. Overall, our technique can help researchers sift most plausible unknown drug-disease relationships, i.e., potential drug repositioning candidates.

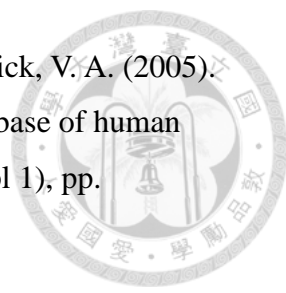


There are some limitations and future research directions relevant to this study. First, the quality of relations from ontologies can be further improved. In this study, we only introduce three biomedical ontologies: DrugBank, OMIM, and CTD. There are plenty of other ontologies and knowledge bases can be adopted, such as Gene Ontology, PharmGKB, OPHID, etc. Incorporating more ontologies can widen the coverage and cross-validate these relations. Besides, some relations lost while translating concepts from ontologies into MeSH terms. Some mapping may also be incorrect. Researchers may develop a better methodology for concept mapping in future. Second, our method, as same as most previous drug repositioning approach, do not leverage known plausible or implausible links. It is possible to apply supervised learning for drug repositioning purposes. There may be several structural characteristics that differentiate between known plausible indirect links and known implausible ones. This may improve the effectiveness for predicting possible drug repositioning candidates. Third, our method can be applied for other purposes, such as discovering unknown drug-drug interactions or adverse drug reactions.



## References

- Adams, C. P., & Brantner, V. V. (2006). Estimating the cost of new drug development: Is it really \$802 million? *Health Affairs*, 25(2), pp. 420-428.
- Ashburn, T. T., & Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8), pp. 673-683.
- Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J., & Bork, P. (2008). Drug target identification using side-effect similarity. *Science*, 321(5886), pp. 263-266.
- Celgene Corporation. (2006). *2005 Annual Report*. Summit, NJ: Celgene Corporation.
- Celgene Corporation. (2009). *2008 Annual Report on Form 10-K*. Summit, NJ: Celgene Corporation.
- Celgene Corporation. (2013). *2012 Annual Report on Form 10-K*. Summit, NJ: Celgene Corporation.
- Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., . . . Tang, Y. (2012). Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Computational Biology*, 8(5), p. e1002503.
- Cilibrasi, R. L., & Vitányi, P. M. (2007). The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3), pp. 370-383.
- Davis, A., King, B. L., Mockus, S., Murphy, C. G., Saraceni-Richards, C., Rosenstein, M., . . . Mattingly, C. J. (2013). The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Research*, 39(suppl 1), pp. D1067-D1072.
- DiMasi, J. A., & Grabowski, H. G. (2007). The cost of biopharmaceutical R&D: Is biotech different? *Managerial and Decision Economics*, 28(4-5), pp. 469-479.
- Frijters, R., van Vugt, M., Smeets, R., van Schaik, R., de Vlieg, J., & Alkema, W. (2010). Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Computational Biology*, 6(9), p. e1000943.

- 
- Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(suppl 1), pp. D514-D517.
- Hristovski, D., Peterlin, B., Mitchell, J. A., & Humphrey, S. M. (2005). Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics*, 74, pp. 289-298.
- Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1), pp. 39-43.
- Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., . . . Wishart, D. S. (2011). DrugBank 3.0: A comprehensive resource for 'omics' research on drugs. *Nucleic Acids Research*, 39(suppl 1), pp. D1035-D1041.
- Lee, H., Bae, T., Lee, J.-H., Kim, D., Oh, Y., Jang, Y., . . . Kim, S. (2012). Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug. *BMC Systems Biology*, 6(1), p. 80.
- Lee, S., Choi, J., Park, K., Song, M., & Lee, D. (2012). Discovering context-specific relationships from biological literature by using multi-level context terms. *BMC Medical Informatics and Decision Making*, 12(Suppl 1), p. S1.
- Li, J., & Lu, Z. (2012). A new method for computational drug repositioning. *2012 IEEE International Conference on Bioinformatics and Biomedicine*, (pp. 1-4). Philadelphia, PA.
- Li, J., Zhu, X., & Chen, J. Y. (2009). Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. *PLoS Computational Biology*, 5(7), p. e1000450.
- Liben-Nowell, D., & Kleinberg, J. (2007). The link prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7), pp. 1019-1037.



Lindsey, R., Veksler, V. D., Grintsvayg, A., & Gray, W. D. (2007). Be wary of what your computer reads: The effects of corpus selection on measuring semantic relatedness. *Proceedings of ICCM 2007: Eighth International Conference on Cognitive Modeling* (pp. 279-284). Oxford, UK: Taylor & Francis/Psychology Press.

Lu, Z., & Wilbur, W. (2009). Improving accuracy for identifying related PubMed queries by an integrated approach. *Journal of Biomedical Informatics*, 42(5), pp. 831-838.

McBride, W. G. (1961). Thalidomide and congenital abnormalities. *Lancet*, 278(7216), p. 1358.

National Institutes of Health. (2009, 5 20). *NIH Announces New Program to Develop Therapeutics for Rare and Neglected Diseases*. (National Institutes of Health) Retrieved 12 2, 2012, from NIH News:  
<http://www.nih.gov/news/health/may2009/nhgri-20.htm>


Özgür, A., Vu, T., Erkan, G., & Radev, D. R. (2008). Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, 24(13), pp. i277-i285.

Pharmaceutical Research and Manufacturers of America. (2007). *Drug Discovery and Development: Understanding the R&D Process*. Washington, DC: Pharmaceutical Research and Manufacturers of America.

Qu, X. A., Gudivada, R. C., Jegga, A. G., Neumann, E. K., & Aronow, B. J. (2009). Inferring novel disease indications for known drugs by semantically linking drug action and disease mechanism relationships. *BMC Bioinformatics*, 10(Suppl 5), p. S4.

Srinivasan, P. (2004). Text mining: Generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology*, 55(5), pp. 396-413.

Stephens, T. D., & Brynner, R. (2001). *Dark Remedy: The Impact of Thalidomide and Its Revival as a Vital Medicine*. Cambridge, MA: Perseus Publishing.

- 
- Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1), pp. 7-18.
- Swanson, D. R. (1988). Migraine and magnesium: eleven neglected connections. *Perspectives in biology and medicine*, 31(4), pp. 526-557.
- Swanson, D. R., & Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence*, 91(2), pp. 183-203.
- Thomson Reuters. (2012, 9). *White Paper: Knowledge-based Drug Repositioning to Drive R&D Productivity*. Philadelphia, PA: Thomson Reuters. Retrieved 6 14, 2013, from <http://ip-science.thomsonreuters.com/info/drugrepositioning/>
- Weeber, M., Klein, H., de Jong-van den Berg, L. T., & Vos, R. (2001). Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. *Journal of the American Society for Information Science and Technology*, 52(7), pp. 548-557.
- Wren, J. D., Bekereditian, R., Stewart, J. A., Shohet, R. V., & Garner, H. R. (2004). Knowledge discovery by automated identification and ranking for implicit relationships. *Bioinformatics*, 20(3), pp. 389-398.
- Yang, L., & Agarwal, P. (2011). Systematic drug repositioning based on clinical side-effects. *PLoS ONE*, 6(12), p. e28025.
- Yetisgen-Yildiz, M., & Pratt, W. (2006). Using statistical and knowledge-based approaches for literature-based discovery. *Journal of Biomedical Informatics*, 39(6), pp. 600-611.
- Yetisgen-Yildiz, M., & Pratt, W. (2009). A new evaluation methodology for literature-based discovery systems. *Journal of Biomedical Informatics*, 42(4), pp. 633-643.