SYSTEM BIOLOGY MODELING: THE INSIGHTS FOR COMPUTATIONAL DRUG DISCOVERY

Hui Huang

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University

May 2014

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Huanmei Wu, PhD, Chair

Jake Chen, PhD

Doctoral Committee

Mohammad Al Hasan, PhD

Yunlong Liu, PhD

February 21, 2014

Yaoqi Zhou, PhD

Dedication

To my father and my mother, without whom none of my success would be possible

To Iselin, whose constant support and encouragement have sustained me throughout my life

Acknowledgements

The journey of a Ph.D. is not a lonely experience and quite the opposite it relies on the help and encouragement of many. This work is no exception. It is my great pleasure to have an opportunity to thank many without whom this thesis is not possible.

First and foremost my sincere appreciation must go to my advisor, Dr. Jake Chen, with whom my association dates back to August 2008. He has been of immense support and an excellent mentor. Jake not only introduced me to computational drug discovery research but also gave me the necessary freedom to pursue my scientific interests. I consider myself fortunate to have an informal and friendly interaction with Jake during my PhD study. He has provided me with excellent research advice, great insights, enthusiasm and encouragement throughout the development of this project. Without him, this work would not be materialized.

My thanks need also be extended to all my research committee for supporting my thesis work and all the useful discussions and suggestions. They guidance and advice keep my research focused and on track, which has allowed me to accomplish this goal.

I am also grateful to other faculty from the School of Informatics. I thank Dr. Mohammad Al Hasan for teaching me how to apply different algorithms in bioinformatics field. I thank Dr. Yunlong Liu for introducing the state of art NGS researches. I thank Dr. Huanmei Wu for discussions about translational bioinformatics. I thank Dr. Sujuan Gao for making learning statistics that intuitive and clear. I thank Dr.

Pedro Romero for teaching me various machine learning skills which I could apply in my research. Last but not least, I thank Dr.Yaoqi Zhou, Dr. Karl MacDorman, Dr. Mathew Palakal for their academic and financial support during the PhD study.

One of the greatest things about doing research is the opportunities for collaborations. I thank Dr. Xiaogang Wu for the collaborations in various projects. His energy, broad knowledge and sense of humor have made the sharing of the lab in the past five years a pleasant experience and enable the synergies in our research. I thank Dr. Jiao Li for the collaboration in my first paper writing and helping me understand the computational connectivity map project. I thank Sara Ibrahim and Thanh Nguyen for closely working with me in the disease model and algorithm development in the past year. I thank Dr. Fan Zhang and Madhankumar Sonachalam in the CCE project and many other lab members in Dr. Chen's lab for different collaborative projects.

I also thank my supervisor and colleagues at the Systematic Drug Repositioning group at GlaxoSmithKline for all the support and guidance during my summer internships there. Special thanks must go to Dr. Lun Yang, who worked with me closely to advance the project. I also thank Dr. Pankaj Agarwal for the all the insightful discussions and critiques, Dr. Philippe Sanseau and Dr. Angela Qu for writing the manuscripts together, Dr. Vinod Kumar, Dr.Qing Xie and Dr.Lixia Yao for all the invaluable scientific discussions.

Finally I would like to express special thanks and appreciation to my family. I thank my father, Siwen Huang, and my mother, Jiacui Cui for their consistent encouragement and tolerance for me to pursue my career. I thank my young brother, Peng Huang, for being supportive in the family. I would like to conclude by extending my thanks to Iselin Tong who has stood by me through all these years with endless patience and understandings.

Hui Huang

SYSTEM BIOLOGY MODELING: THE INSIGHTS FOR COMPUTATIONAL DRUG DISCOVERY

Traditional treatment strategy development for diseases involves the identification of target proteins related to disease states, and the interference of these proteins with drug molecules. Computational drug discovery and virtual screening from thousands of chemical compounds have accelerated this process. The thesis presents a comprehensive framework of computational drug discovery using system biology approaches. The thesis mainly consists of two parts: disease biomarker identification and disease treatment discoveries.

The first part of the thesis focuses on the research in biomarker identification for human diseases in the post-genomic era with an emphasis in system biology approaches such as using the protein interaction networks. There are two major types of biomarkers: Diagnostic Biomarker is expected to detect a given type of disease in an individual with both high sensitivity and specificity; Predictive Biomarker serves to predict drug response before treatment is started. Both are essential before we even start seeking any treatment for the patients. In this part, we first studied how the coverage of the disease genes, the protein interaction quality, and gene ranking strategies can affect the identification of disease genes. Second, we addressed the challenge of constructing a central database to collect the system level data such as protein interaction, pathway, etc. Finally, we built case studies for biomarker identification for using Diabetes as a case study.

The second part of the thesis mainly addresses how to find treatments after disease identification. It specifically focuses on computational drug repositioning due to its low lost, few translational issues and other benefits. First, we described how to implement literature mining approaches to build the disease-protein-drug connectivity map and demonstrated its superior performances compared to other existing applications. Second, we presented a valuable drug-protein directionality database which filled the research gap of lacking alternatives for the experimental CMAP in computational drug discovery field. We also extended the correlation based ranking algorithms by including the underlying topology among proteins. Finally, we demonstrated how to study drug repositioning beyond genomic level and from one dimension to two dimensions with clinical side effect as prediction features.

Huanmei Wu, PhD, Chair

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction to Drug Discovery with System Approaches

The complexity of human biology makes it challenging for drug discovery. Traditionally drug discovery involves disease identification for the patients and treatment identification for the patients with that specific disease. The former corresponds to reliable disease biomarker development for the diagnostic purpose. The latter refers to develop effective medicines for the treatment. With the development of omics-based techniques, systems biology leverages the high-throughput data to connect molecular network and pathway information to build better disease models and help predict drug effects in patients.

1.1 Disease Biomarker Identification

Biomarkers are molecular signatures that enable early diagnosis, guide molecularly targeted therapy and monitor the activity and therapeutic responses across a variety of disease. They are increasingly important in both therapeutic and diagnostic processes. The hope of finding new biomarkers for assessing cancer risk, detecting cancer at an early stage, subtyping tumours, selecting optimal therapies, and monitoring therapeutic response is the motivation behind substantial current investments in biomarker research.

Biomarker can be classified according to its purpose. Diagnostic Biomarker is expected to detect and identify a given type of disease in an individual with both high sensitivity and specificity. Prognostic Biomarker is used to predict the probable course of the disease including it recurrence and progression once the disease status has been set. Predictive Biomarker serves to predict drug response before treatment is started. Generally, this marker classifies individuals as likely responders or non-responders to a particular treatment.

The identification of biomarker involved in human diseases has been a primary focus of post-genomic biomedicine for pursuing the clinical goals of diagnosis and therapeutic treatment. Recent advances in genomics, transcriptomics, proteomics, and metabolomics have begun to help unravel the disease molecular mechanisms. Gene expression profiling has revealed common gene fusions and expression ‘signatures’ in cancer patients. For example, two studies show that the common recurrent gene fusion between TMPRSS2

and ERG promotes prostate cancer in both mouse and humans, when the function of PTEN is concurrently lost(*1*).

From a systems perspective, one of the emerging themes today is to re-characterize a protein's biological function in their molecular interaction network and pathway context. Network Biomarker (*2-4*) is a new concept for biomarker discovery in systems biology. By integrating cancer susceptibility genes, gene expressions, and their protein interaction network, Marc Vidal's group at Harvard constructed a protein interaction network for breast cancer susceptibility and identified HMMR as a new susceptibility locus for the breast cancer (*5*). Later, Trey Ideker's group at UCSD integrated protein network and gene expression data to improve the prediction of metastasis formation in breast cancer patients (*4, 6*). The two studies suggest that protein interaction networks and pathway, although drafty and incomplete, can serve as a molecular-level conceptual roadmap to guide cancer biomarkers studies (*7*).

Pathway Biomarker (*8*) is a concept for biomarker discovery by integrating functional genomics and known signaling pathway data. Recent finding suggests that cancer is dysregulated at the pathway level. Coupling Omics results with molecular signaling pathways involved in cancer and studying how cancer cellular function is regulated at the pathway level have been a key topic in cancer systems biology.

1.2 Drug Discovery

Traditional treatment strategy development for diseases involves the identification of target proteins related to disease states, and the interference of these proteins with drug molecules. An explosion of high-throughput data has help measure the drugs' effect experimentally. For example, the experimental Connectivity Map (*15*) contains more than 7,000 expression profiles representing 1,309 compounds. It enables researches on pharmacology at gene expression levels. By correlating the disease gene expression with the pharmacology profiling from CMAP, Lamb (*15*) identified novel drug indications in diet-induced obesity or Alzheimer's disease. Other pharmacology databases like NCI60(*20*) and CCLE(*21*) haven also been developed to provide pharmacological profiles about drug sensitivity.

Computational drug discovery and virtual screening from thousands of chemical compounds have accelerated this process. The conventional “One disease, One gene, and One drug” paradigm (*9*) works effectively for simple genetic disorders. However, recent research studies show, in the case of both older psychiatric drugs and modern anticancer therapies, that drugs with multiple targets can contribute to the drug’s therapeutic efficacy(*11*). Thus the concept of network pharmacology (*12*)or network medicine (*13*) has been developed to understand the actions of drugs by considering targets in the context of the biological networks. Such a bioinformatics network analysis of high-throughput data sets offers an opportunity for integration of biological complexity and multilevel connectivity(*14*).

Machine learning or text mining based methods have been developed to overcome some of those limitations. Gottlieb (*22*) developed a logistic regression method called PREDICT to predict drug actions. The regression is mainly based on drug-drug similarity (i.e. chemical similarity, side effects, drug targets sequence similarity, PPI and GO distance) and disease-disease similarity (i.e. similarity from text mining and human phenotype ontology). The PREDICT achieves an AUC of 0.9 while the AUC of above mentioned CMAP method is only around 0.4. Another approach called The Connectivity Maps (C-Maps) web server (*23*) is an online bioinformatics resource that provides biologists with potential relationships between drugs and genes/proteins in specific disease contexts based on network mining and literature mining. Disease-specific protein-drug association profiles are computationally generated by mining bio-molecular interaction networks and PubMed literature (*24*). Despite of those advancements, statistical association based studies are hard to differentiate drug efficacy from toxicity and even harder to provide a mechanistic view about why drugs have certain actions in specific disease conditions.

Non-mechanistic approach contributes to a serious reliability and reproducibility issue in preclinical cancer drug research (*25*). A former cancer researcher at Amgen identified 53 "landmark" publications -- papers in top journals, from reputable labs and 47 of the 53

could not be replicated(*25*). Mechanistic model-based analysis(*26*) has been explored to aid in drug discovery to understand and predict the interaction of small molecule inhibitors with pathways. Take breast cancer for example. In a simple one-drug-one-target scenario, tamoxifen can treat ER positive patients by inhibiting its target--over-expressed estrogen receptor. However how to address the off-target effects (OTE) of a drug on the proteins downstream in the signaling pathways and then manipulate them for therapeutic purposes remains a big challenge.

Khatri (*27*) reviewed the ten year pathway analysis and pointed out that the challenges of pathway based studies lie in the incompleteness of the pathway data which causes trouble for the current Pathway Topology (PT)-Based Approaches. Though HPD(*28*) collects human pathways from various sources, a disease specific pathway is not available. KEGG(*29*) covers only a limited list of disease pathway. For example KEGG doesn't contain breast cancer pathway. Even with such a disease pathway, another challenge is to identify drug's effect on the pathway proteins. There are over 40 drug-target (protein-compound interaction) databases, according to Pathguide (*30*), (e.g. DrugBank(*31*), STITCH(*32*), and PharmGKB (*33*) et.al). DrugBank, for example, informs the researcher about interactions between drugs, physical drug target and proteins that metabolize the drug (*31*). However, these databases seldom directly inform the researcher about the directionality of a drug-target relation although this information may be scattered within a description or referenced text. Another difficulty is the inability to integrate the states of single proteins jointly into the higher level states of protein modules or pathway levels. The biological system is far from a homogenous one and thus makes a single general function, such as the sum of gene expression levels in a module, far from reality.

Despite of those challenges, the computational prediction of drug efficacy could be particularly rewarding, especially in drug repositioning for personalized medicine(*34*) applications. The traditional drug development pharmaceutical product development requires at least 10 to 15 years and costs between $500 million and $2 billion(*35*). According to the U.S. FDA, up to 90% of all experimental drug compounds going through clinical trials failed to gain FDA approvals and drug efficacy accounts for 25-

30%(*36*). Repositioning (*37*)refers to the identification and development of new uses for existing or abandoned drugs. It could greatly accelerate drug discovery because existing drugs have established clinical data(*38*). With recent advances in Omics and next generation sequencing techniques, elucidating the molecular basis of disease on a personalized level and thus tailing treatments accordingly has become an attainable goal. Drug failed at whole population may work at certain sub-population as personalized medicine(*39*). Drug repositioning for personalized medicine (*34*) aims to improve the productivity of current drug discovery pipelines and will benefit from the improved computational drug efficacy prediction.

1.3 Drug Repositioning

Discovering new indication for existing drugs, known as drug repositioning, is a hot topic in the translational bioinformatics field (*17, 22, 40*). Traditional drug discovery takes billions of dollars and an average of fifteen years to bring a new drug to the market (*41*). It's estimated 90% of the drugs fail in the early stage of drug development(*42*). Repositioning of drugs already approved for human use could alleviate the cost associated with early stages and offer a shorter path for new approval(*43*). Both academia and pharmaceutical companies have achieved a number of successes by using drug repurposing. For example, the drug sildenafil, initially developed for pulmonary hypertension and angina pectoris, has been repositioned for erectile dysfunction indication. Thalidomide, originally applied for treatment of morning sickness and withdrawn from the market after causing thousands of severe birth defects, has been approved for indication in severe erythema nodosum leprosum(*44*).

Current computational methods for drug repositioning include: (i) studying the structural similarity of each drug to their targets' ligand set using chemoinformatics tools (*45*) or drug–drug and disease–disease similarity with machine learning methods(*22*), (ii) exploiting side-effect similarities (*46*), (iii) applying text-mining literature(*23*), or (iv) matching drug and disease gene expression profiles (*15, 17, 40, 47, 48*). Most of the approaches can only be applied to well characterized drugs whose targets or structures are known. Expression profile based approaches are, on the other hand, more general and do not require prior knowledge of the drugs.

Lamb *et al.(15)* developed a public available database called The Connectivity Map (CMAP) containing a collection of transcriptional expression data from cell lines treated with small molecules. The reference database can be queried with gene signatures of interest, with a compound being identified if the genes in the signature are significantly modulated by that compound. Iorio *et al.(40)* constructed drug-drug similarity networks based on the gene expression profiles in the CMAP and proposed drug repositioning based on drug pairwise similarity. Hu and Agarwal(*47*) and Sirota *et al.(17)* extended the idea by paring drugs and diseases whose gene expression patterns are negatively correlated. They further showed that the anti-correlation relationships between the drugs and diseases can suggest novel therapeutics for existing drugs.

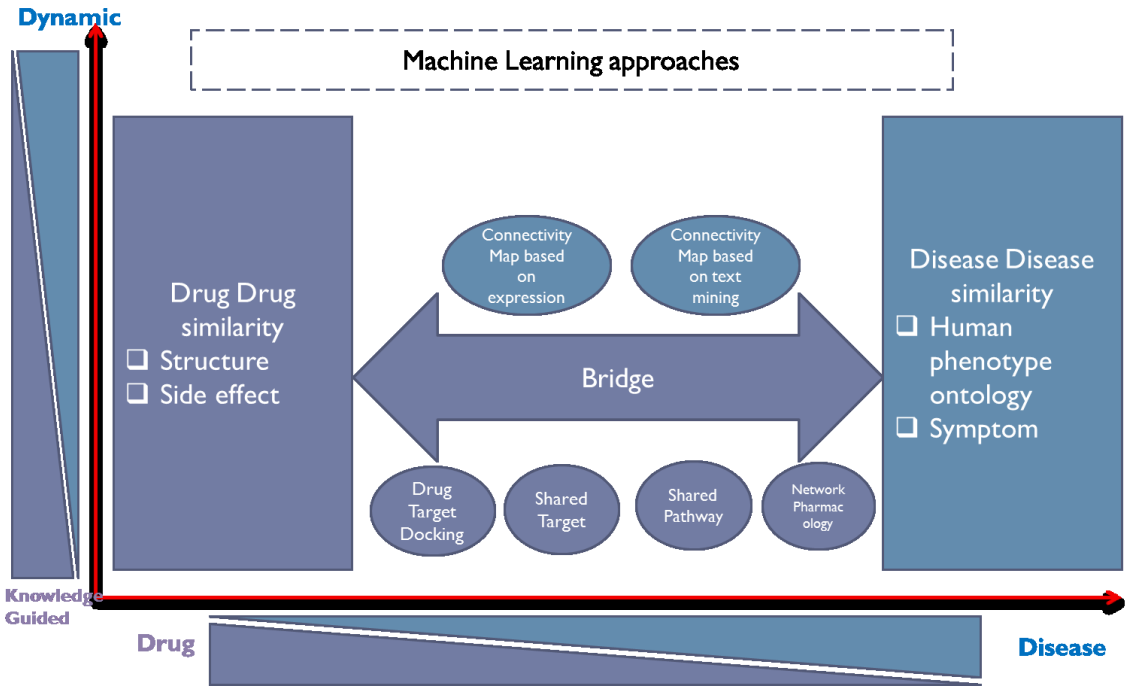

Figure 1-1. Existing approaches for drug repositioning.
All approaches are judged by two dimensions: 1) whether the approach is dependent on the dynamic expression profiles or the approach is based on existing knowledge; 2) whether the approach is based on drug chemical similarities or the approach is based on disease models. In general effective medicine is based on specific disease model and gene expressions from the patients.

1.4 Organization of the Thesis

The thesis is organized as follows:

In Chapter 1, I provide an introduction of biomarker discovery and drug discovery, emphasizing the need for system level approaches. Coupling biomarker discovery and drug discovery is essential for identifying an effective treatment for patients.

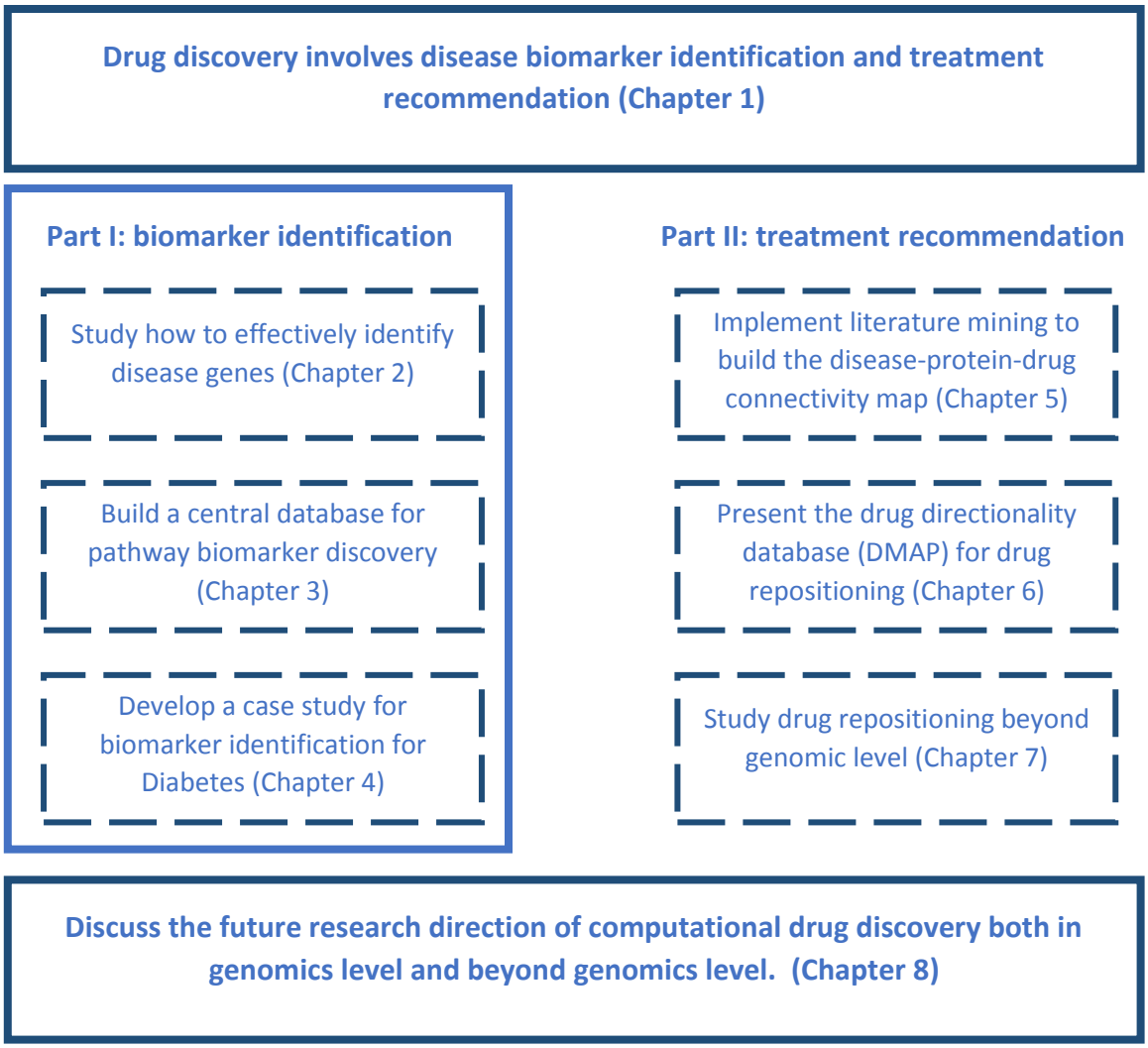

Figure 1-2. A research road map for the thesis.

Chapter 2 is based on my work at (*49*). We describe a simple yet generic computational framework based on protein interaction networks to perform and evaluate disease gene-hunting, using colorectal cancer as a case study. We apply statistical measurements including specificity, sensitivity and Positive Predictive Value (PPV) to evaluate the performance of disease gene ranking methods, which we break down into seed gene

selection, protein interaction data quality and coverage, and network-based gene-ranking strategies. We discover that best results may be obtained by using curated gene sets as seeds, applying protein interaction data set with high data coverage and decent quality, and adopting variants of local degree methods.

Chapter 3 is based on my work at (*50*). We develop an integrated online database, the Pathway And Gene Enrichment Database (PAGED), to enable comprehensive searches for disease-specific pathways, gene signatures, microRNA targets, and network modules by integrating gene-set-based prior knowledge as molecular patterns from multiple levels: the genome, transcriptome, post-transcriptome, and proteome. The online database we developed, PAGED (http://bio.informatics.iupui.edu/PAGED) is by far the most comprehensive public compilation of gene sets. In its current release, PAGED contains a total of 25,242 gene sets, 61,413 genes, 20 organisms, and 1,275,560 records from five major categories. Beyond its size, the advantage of PAGED lies in the explorations of relationships between gene sets as gene-set association networks (GSANs). Using colorectal cancer expression data analysis as a case study, we demonstrate how to query this database resource to discover crucial pathways, gene signatures, and gene network modules specific to colorectal cancer functional genomics.

Chapter 4 is based on my work at (*51*). We present an innovative approach - network expansion and pathway enrichment analysis (NEPEA) for integrative microarray analysis. We assume that organized knowledge will help microarray data analysis in significant ways, and the organized knowledge could be represented as molecular interaction networks or biological pathways. Based on this hypothesis, we develop the NEPEA framework based on network expansion from the human annotated and predicted protein interaction (HAPPI) database, and pathway enrichment from PAGED. We use a recently-published microarray dataset (GSE24215) related to insulin resistance and type 2 diabetes (T2D) as case study, since this study provided a thorough experimental validation for both genes and pathways identified computationally from classical microarray analysis and pathway analysis. We perform the NEPEA analysis for this dataset based on the results from the classical microarray analysis to identify biologically significant genes

and pathways. Our findings are not only consistent with the original findings mostly, but also obtained more supports from other literatures. Chapter 1-4 conclude the first part in this thesis about disease biomarker discovery.

Chapter 5 is based on my work at (*52*). We assess drug pharmacological effect by assuming that "ideal" drugs for a patient can treat or prevent the disease by modulating gene expression profiles of this patient to the similar level with those in healthy people. Starting from this hypothesis, we build comprehensive disease-gene-drug connectivity relationships with drug-protein directionality (inhibit/activate) information based on a computational connectivity maps (C2Maps) platform. An interactive interface for directionality annotation of drug-protein pairs with literature evidences from PubMed has been added to the new version of C2Maps. We also upload the curated directionality information of drug-protein pairs specific for three complex diseases - breast cancer, colorectal cancer and Alzheimer disease. For relevant drug-protein pairs with directionality information, we use breast cancer as a case study to demonstrate the functionality of disease-specific searching. Based on the results obtained from searching, we perform pharmacological effect evaluation for two important breast cancer drugs on treating patients diagnosed with different breast cancer subtypes. The evaluation is performed on a well-studied breast cancer gene expression microarray dataset to portray how useful the updated C2Maps is in assessing drug efficacy and toxicity information.

Chapter 6 is based on my work at (*53*). Critical to drug repositioning involves the reliable measurements of how drug affect disease proteins. We present a computational framework to address those challenges. First, we introduce the Drug directionality Map (DMAP) which consists of directed drug protein relationships for 328,676 drugs. We scale up the coverage of the database by 200 fold compared to experimental based Connectivity Map (CMAP) which suffers from limited drug coverage due to the experimental cost. DMAP enables systematic repositioning for 982 drugs and 622 diseases. With two well-established drug-repositioning methods: drug similarity networks method and K-S scoring method, we demonstrate the feasibility of applying this valuable dataset for systematic drug repositioning. The results demonstrate that the DMAP

database is essential for computational drug repositioning research. Chapter 5 and 6 conclude the drug repositioning work in the genomic level.

In Chapter 7, I explore the possibility of drug repositioning beyond one dimension and beyond genomics level. We hypothesize that clinical side effects (SEs) provide a human phenotypic profile and can be translated into the development of in silico models for predicting novel drug combinations likely to be safe and efficacious. We build a prediction model based on the SE features and test it on a large independent drug combination set. The prediction achieves an accuracy of 0.94 and an AUC of 0.87. We demonstrate that such prediction power is not due to the confounding factors such as biased disease indications or drug targets. To explore the possibility of applying the prediction in practice, we train a rule-based model, namely the decision tree model, and successfully reduce the features to only three 'black box' SEs: *pneumonia, hemorrhage rectum, and retinal bleeding*, whilst maintaining an AUC of 0.80. Based on these results, we propose that avoiding combining drugs with any of these three serious SEs would have better chance of reducing the risks. Finally, we propose a "Two Steps Rule" so that it can help to identify potential safe co-prescriptions or novel fix-dose combinations while maintaining the efficacy.(*54*)

In Chapter 8, I summarize all the research and contributions. I also discuss about the future research direction of computational drug discovery both in genomics level and beyond genomics level. I highlight the significance of incorporating phenotypic features when building the prediction models.

Chapter 2. Analysis of Protein Interaction Network

This section is based on the published work at (*49*). JYC guided the research team by providing ideas and feedback along the way, and revised the manuscript. HH integrated disease genes, generated the protein interaction network, ranked the disease proteins, and wrote the manuscript. JL helped with the specificity, sensitivity and PPV calculation and revised the manuscript.

2.1 Introduction

Disease gene finding is a central topic in biomedical research. If the causal genes are found for a disease, health care solutions may be developed to prevent disease occurrence, diagnose disease early, and make tailored treatment plans, e.g., in (*55, 56*). For nearly a century, there have been two approaches to discover genes related to a specific disease experimentally: biochemical analysis approach and genetic analysis approach (*57*). The first approach attempts to first separate and purify proteins characteristic of disease conditions in model organisms or tissues, and then study the disease-related proteins' biochemical or biophysical altered properties that can be mapped to gene mutations. The second approach normally relies on first studying genetic markers identified in families of diseased populations, and then applying positional cloning techniques and linkage analysis to identify microsatellite markers, chromosomal aberrations, or DNA polymorphisms. However, experimental characterization of proteins or genes involved in diseases is a slow meticulous process. Today, even with advances of genomics technology, one third of all the genes and most of the disease related genes remain functionally uncharacterized (*58*). A promising new experimental technique is genome-wide association studies (GWAS), which may help identify candidate single-nucleotide polymorphism (SNP) genetic markers associated with disease risks.

While most computational approaches to disease gene finding rest on statistical association studies or computational sequence analysis, there are surging interests in taking advantage of molecular interaction networks. The concept is to put candidate genes and proteins in specific disease biology contexts defined by molecular interaction networks or biomolecular pathways, with which a researcher can infer functions of uncharacterized genes or proteins. Such disease biomolecular network context may be

particularly useful for the study of polygenic diseases such as cancer, in which conventional reductionist approaches are ineffective (*55*). In this new approach, disease networks are developed to rank disease relevance of genes/proteins based on properties such as node *degree*s (count of direct PPI connections to a node), *closeness* (path distance of a given node to all other nodes), or *betweenness* (count of geodesic paths that pass through a node). For example, Morrison *et al.* used gene expression network and gene ontology information to rank genes similar to Google's PageRank method (*59*). *Chen et al.* were the first to propose a method that applied disease-specific protein-protein interaction (PPI) networks and modified local node degree measures to prioritize Alzheimer's disease genes (*60*).

While many network-based disease-gene ranking methods have been developed recently, there has not been a consensus how to evaluate their performances. In this work, we describe a simple yet generic computational framework to perform and evaluate network-based disease gene-hunting methods. Using colorectal cancer gene finding as a case study, we report how various seed gene selection, PPI data quality, and ranking strategy could affect final gene-finding results. We also defined how specificity, sensitivity, and positive predictive values (PPV) could be used for performance evaluation criteria. We choose colorectal cancer because it is the third leading cause of cancer death in the US and our current knowledge of colorectal cancer genes is limited, making our results to carry special significance. Next, we will describe our methods and report our findings.

2.2 Methods

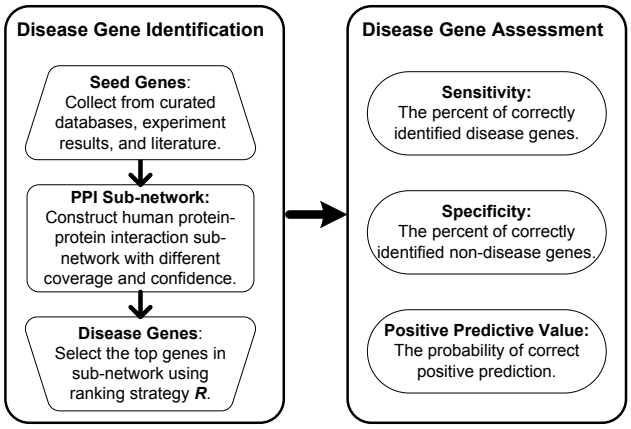

Figure 2-1. Computational Framework for Disease Gene Identification and Assessment

In Figure 2-1, we show an overview of the computational framework used in this study. It consists of two components: (1) *Disease Gene Identification,* in which we expand seed genes to disease-specific protein sub-network and subsequently generate a ranked list of disease-relevant genes; (2) *Disease Gene Assessment*, in which we quantitatively assess disease genes using statistical measurements including sensitivity, specificity and PPV. The relationships between the two components is the following: First, disease gene identification will be performed using a fixed set of gene-seeding, PPI sub-network construction, and disease gene ranking strategies; then, we evaluate how sensitivity, specificity, and PPV are affected by varying choices of seed genes, PPI networks, and ranking strategy.

i. Seed Gene Selection

We consider three sets of colorectal cancer-related genes collected from different resources as *seeds*, which are: (1) the **CORE1** set (i.e. the curated genes), derived from human curated databases by querying the OMIM (*61*) and KEGG (*62*) database for "colorectal cancer" and manually curating the set of genes/proteins; (2) the **CORE2** set(i.e. the expressed genes), derived from high-throughput microarray data in the ONCOMINE (*63*) database by keeping only differentially expressed genes with p-value<0.05 performed for colorectal cancer samples against controls; (3) the **CORE3** set(i.e. text mining genes), derived from the Comparative Toxicogenmics Database (CTD

(*64*)) by searching for colorectal cancer genes associated with >2 chemicals in the database.

ii. Protein Interaction Sub-network Construction

We expand seeds, using PPIs recorded in the Human Annotated and Predicted Protein Interactions (HAPPI) database (*65*) to construct colorectal cancer-specific PPI sub-network. A unique feature of the HAPPI database is that the quality of PPIs comes with estimated confidence scores (a real value between 0 and 1) and star grades (an integer between 1 and 5). The higher the confidence score or the star grade number, the more likely the PPI is attributable to physical PPI events. In this study, we use PPI star grade to control disease-specific sub-network quality and coverage. We refer to the disease-specific PPI sub-network constructed from HAPPI quality star grade *n* and above as PPI-*n*. For example, PP1-3 includes all PPIs from HAPPI with quality grade of 3, 4, and 5.

iii. Disease Gene Ranking Strategy

We treat the disease gene ranking problem as a problem to calculate a weight for each protein in the disease-specific PPI sub-network. There are three ranking strategies being considered in this study: (1) *Global degree* strategy, in which we use the protein's node degree in the global PPI-n network as the weight; (2) *Local degree* strategy, in which we use the protein's node degree in the local (colorectal-specific) PPI-n network as the weight; and (3) *Edge-weighted Promiscuous Hub subtraction* (EPHS) strategy developed in Dr. Chen's lab (*60*), which is a variant of *local degree* strategy adapted by penalize the impact of low-quality promiscuous protein hubs on ranks defined by the following formula:

$$r_p = k * \ln(\sum_{q \in NET} conf(p,q)) - \ln(\sum_{q \in NET} N(p,q)) \quad (1)$$

Here, *p* and *q* are indices for proteins in the constructed network *NET*. *k* is an empirical constant. *conf(p, q)* refers to confidence score in HAPPI Database. *N(p, q)* holds the value of 1 if the protein *p* interacts with *q*. The r_p score is the weight calculated to rank each protein in the network.

In addition, we use *TOP_M* to refer to the *M* highest ranked disease-relevant proteins/genes given by a specific disease gene ranking strategy.

iv. Disease Gene Assessment

To evaluate the disease-related gene list, the sets of Gold Standard Positive (GSP) and Gold Standard Negative (GSN) are constructed as illustrated in Figure 2-2.

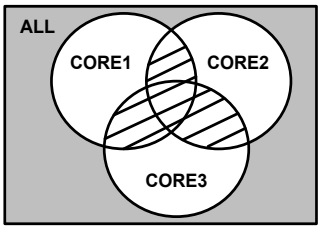

Figure 2-2. Gold standard construction for disease gene assessment.
As shown in the striped area, GSP= (CORE1∩CORE2)∪(CORE1∩CORE3)∪(CORE2∩CORE3). As shown in the gray area, GSN=ALL-(CORE1∪ CORE2∪ CORE3). Note that ALL refers to all HAPPI human proteins.

The following measurements are calculated to evaluate the performance of each disease gene identification method: (1) **Sensitivity**, calculated as the percent of correctly identified disease genes $|TOP\_M \cap GSP|/|GSP|$; (2) **Specificity**, calculated as the percent of correctly identified non disease genes $|GSN-(TOP\_M-GSP)|/|GSN|$; (3) **Positive Predictive Value (PPV)**, calculated as the probability of correct positive predictions $|TOP\_M \cap GSP|/|\ TOP\_M\ |$.

2.3 Results

We developed three colorectal cancer seeds: CORE1, consisting of 148 proteins; CORE2, consisting of 42 proteins; and CORE3, consisting of 721 proteins. With three choices of seeds gene selections (CORE1, CORE2, and CORE3), four PPI qualities (PPI-3, PPI-4, PPI-5, PPI-1), three ranking strategies (EPHS, Local Degree, Global Degree), we tested different combinations to conduct the disease gene findings and assessment for colorectal cancer.

i. Effect of Various Seed Gene Selection Methods

In Figure 2-3, we show how seed selections affect the ranking results. In this experiment, we used PPI-3 as PPI network data source and the EPHS disease protein ranking method. The *ranking index* on the x-axis refers to a number, *TOP_M*, used to indicate the number of all rank-ordered proteins in a given expanded protein set consisting of both seed proteins and PPI-expanded disease sub-network. PPV for the initial top-10 or top-20 proteins for both core-1 and core-2 seeded strategies were at 0.7-0.8 range, suggesting high predictive power of top-ranked proteins for disease-relevance. As ranking index increases, PPV decrease for all core seeded strategies. However, the performance for core-1 is superior to both core-2 and core-3. This is perhaps due to the highly curated nature of core-1 seeds as compared with possible noises introduced by Omics data for core-2 and text mining data for core-3. Core-3 shows an overall poorer PPV performance, particular within top-20 compared with core-1 and core-2. Beyond ranking index of 250, all core seeded strategies converged to low PPV within 0.15. Therefore, the relatively high predictive powers of all disease gene rankings seem to be restricted to the top 50.

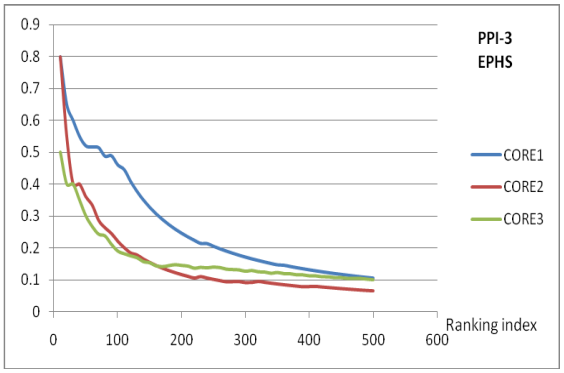

Figure 2-3. PPV performance using different seed choices.

ii. Effect of Various PPI Data Quality and Coverage

In Figure 2-4, we show how PPI data used for network expansion affect the ranking results. In this experiment, we compared results using PPI-1, PPI-3, PPI-4, and PPI-5, using core1 as seeds and the EPHS ranking method. All PPI-n except for PPI-5 showed a similar trend of decreasing PPV. Again, the relatively high predictive powers (PPV>0.5) seem to be achieved at the top 50, except for PPI-5, then continue to decrease to very low levels (PPV<0.15) beyond a ranking index>400. It's counter-intuitive that PPI-5's performance, being the poorest, has a rising phase from ranking index between 10 and 50 before decreasingly significantly. This may be primarily due to the poor coverage of true colorectal cancer proteins in current physical PPI data sets representative of PPI-5 until enough proteins are covered in the top 40 or 50 set. Therefore, data coverage seems quite important in gene ranking performance overall. Also, at least in the top 10 case, the fact that PPI-3 has the best PPV of 0.8 over PPI-1 that has much higher data coverage suggest that PPI data quality is also important to discover disease genes in the most highly ranked protein set. Therefore, balanced data coverage and quality are essential for disease gene finding from such networks.

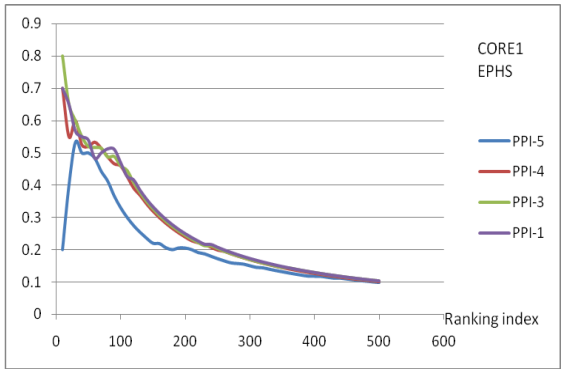

Figure 2-4. PPV performance using different PPI-n networks.

iii. Effect of Various Disease Gene Ranking Methods

In Figure 2-5, we show how the choices of different ranking methods affect the ranking results. The results are performed by fixing seed protein to core1 and using PPI-5 for the expansion network. EPHS and Local Degree methods performed equally, while global degree performed extremely poor—although by sharing similar performance trend of the top-performing methods. The trend for all methods shows two phases: a PPV rising phase

from top 10 to top 60-80; and a PPV decreasing phase from top 80 onwards. The separations of two phases are likely due to balanced PPI data coverage and quality as explained earlier.

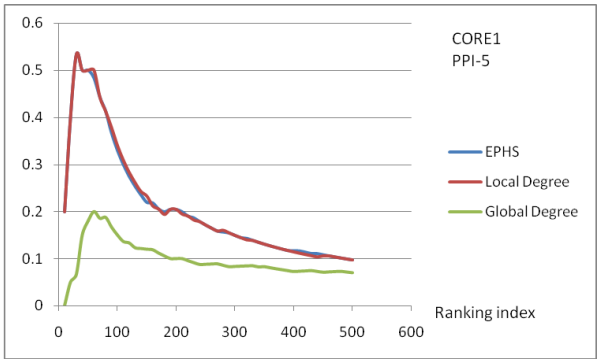

Figure 2-5. PPV performance using different disease gene ranking methods.

iv. Sensitivity and Specificity Comparisons of Top Disease Gene Ranking Methods. We further compared the sensitivity and specificity performances for the best two disease gene ranking methods, EPHS and Local Degree.

Figure 2-6 shows a comparison of their specificity (on the y-axis) performance distributed over different ranking index ranges (on the x-axis). The specificity performances of both methods are quite good overall, even at top 100 range (specificity>0.9). The EPHS ranking method is slightly better (more specific) than Local Degree ranking method. This is primarily because local degree method cannot distinguish nodes with the same number of node degrees, particularly when the node degree drops to small numbers such as 2 or 3 in the high ranking index region.

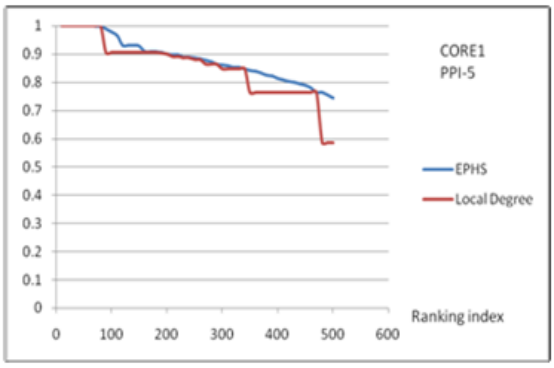

Figure 2-6. A comparison of specificity performance between the EPHS and Local Degree ranking methods.

Figure 2-7 shows a comparison of their sensitivity (on the y-axis) performance distributed over different ranking index ranges (on the x-axis). The sensitivity performances of both methods are decent overall after ranking index range of top 80 (sensitivity>0.75). The local degree ranking method is slightly better (more sensitive) than EPHS ranking method. The reason that local degree method performed better than EPHS ranking method is that there are many tied genes in local degree method due to their sharing the same node degrees. However, since most rankings should be performed in the low ranking index region, this slight loss of sensitivity for EPHS method can be ignored.

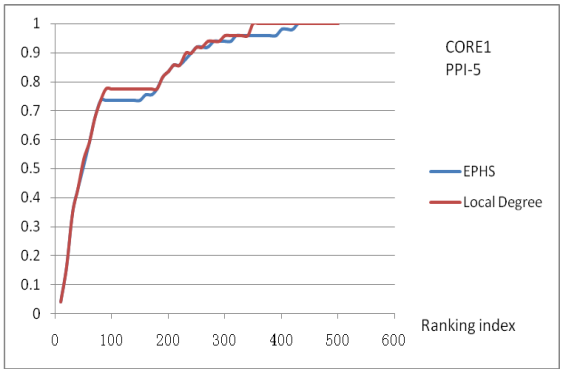

Figure 2-7. A comparison of sensitivity performance between the EPHS and Local Degree ranking methods.

2.4 Conclusions

In this work, we performed disease gene finding from protein-protein interaction networks specific to colorectal cancer. We examined the effects of different seeds, different PPI data quality, and different disease gene ranking methods on the final

performance of the task. While all of these parameters may impact the final performance, our results show that (1) the initial quality of seeds should be based on prior curated knowledge as much possible, with Omics results being the next choice and text mining results being the last resort; (2) disease gene ranking should be performed using PPI data with reasonable quality but as high data coverage as possible; (3) the ranking algorithm that takes advantage of local network parameters should be chosen over those using global network parameters. There are several limitations to our current research approach. For example, the gold standard positive set of genes used for evaluation had to be built by considering seed gene sets used for research studies due to convenience of computation. The observations made for this framework should be carefully validated in other disease contexts before they be generalized.

Chapter 3. Pathway and Gene-set Enrichment Database

This section is based on the published work at (*50*). JYC conceived of this work, guided the research team, and revised the manuscript. HH integrated disease-gene association data, developed the website, designed the case studies and wrote the manuscript. XW participated in the idea initiation, framework development, data quality control, case studies, and manuscript writing. MS integrated various pathways, microRNA, and gene signature data. SNM reviewed the evolvement on pathway analysis and gene-set enrichment analysis. RP helped with the database management and maintenance. KFM tested the website, provided valuable suggestions for substantial improvements, and revised the manuscript. PW assisted with website maintenance.

3.1 Introduction

Pathway analysis and gene-set enrichment analysis are both widely-used methods to identify significant molecular expression patterns from high-throughput data (*27*). Over the last decade, biological pathways have provided natural sources of molecular mechanisms to develop diagnosis, treatment, and prevention strategies for complex diseases (*66-68*). The various and massive functional genomics data are effectively analyzed by gene-set enrichment methods instead of individual gene analysis (*69-72*). Pathway analysis and molecular signature discovery continue to reveal the association between genotypes and phenotypes, which are simply called molecular profiling or molecular phenotypes. At present, researchers intend to combine pathway and gene-set enrichment approaches and network module-based approaches to identify crucial relationships among different molecular mechanisms (*27*).

As sources of prior knowledge for molecular mechanisms, biological pathway databases are heterogeneous, cross multiple levels, and lack annotations (*67*). Different pathway databases may yield divergent results from the same input data. When different databases yield similar results, applying multiple pathway data sources in a single analysis can generate a measure of validation. Unlike candidate pathway analysis, genome-wide pathway analysis does not require prior biological knowledge. In addition, genome-wide pathway analysis can reveal gene interactions across different diseases (*67, 73*) and multiple pathways (*67, 74, 75*). Other studies based on an online integrated human

pathway database (HPD) also provided associations between different pathways with diverse types, sizes, and sources (*28, 76*) on specific phenotypes. Although these efforts have greatly improved the efficiency of pathway analysis, our knowledge of biological pathways is still far from complete.

Gene signature data from the transcriptome level offers a complementary source of information to complete pathway knowledge. In a recent review, Khatri et al. (*27*) categorized pathway analysis into three generations of approaches: the first-generation "over-representation analysis" (ORA) approaches, the second-generation "functional class scoring" (FCS) approaches, and the third-generation "pathway topology" (PT) approaches. To overcome the limitations of ORA approaches (gene-level statistics), FCS approaches, such as gene-set enrichment analysis (GSEA) (*70*), were devised to include overall changes of gene expressions in each pathway/gene set (pathway-level statistics). Third generation approaches also include overall changes of gene expressions based on pathway topology—that is, their upstream/downstream positions within each pathway. Although these third generation approaches were meant to change our understanding of the underlying mechanisms of pathways, they lack information necessary to achieve this: the interdependence between pathways. Annotated knowledge from genome, transcriptome, post-transcriptome, and proteome levels can assist pathway and gene-set enrichment analysis.

Multi-level, multi-scale, knowledge-guided enrichment analysis can enable molecular phenotype discovery for specific human diseases. Currently, the acquisition of prior knowledge and systems modeling poses a challenge for developing tools that go beyond third-generation pathway analysis for disease-specific molecular profiling. Prior knowledge acquisition requires attention to updates and improves the available annotations with descriptive knowledge from multiple levels, especially for information on pathway microenvironment ("condition-, tissue-, and cell-specific functions of each gene") (*27, 67*). Systems biology modeling must incorporate data from the view of systems biology to build systems with multiple scales, which can be used to generate hypotheses that will give detailed and accurate predictions of changes in systems. Both

aspects of this challenge will be addressed by building a database not only containing disease-associated genes, transcript factors, proteins, and microRNAs, but also by organizing their relationships within and between pathways, gene signatures, and any gene sets from existing experiments or papers.

To meet the new challenges of molecular phenotype discovery, we developed in this work an integrated online database, the Pathway And Gene Enrichment Database (PAGED), to enable comprehensive searches for disease-specific pathways, gene signatures, microRNA targets, and network modules, by integrating gene-set-based prior knowledge as molecular patterns from multiple levels—the genome, transcriptome, post-transcriptome, and proteome. The new database can provide the following benefits to biological researchers. First, the new database consists of disease–gene association data, curated and integrated from Online Mendelian Inheritance in Man (OMIM)(*77*) database and the Genetic Association Database (GAD)(*78*); therefore, it has the potential to assist human disease studies. Second, as of March 2012 it also contains all current compiled gene signatures in Molecular Signatures Database (MSigDB)(*72*) and Gene Signatures Database (GeneSigDB)(*71*). Third, it further integrates with microRNA-targets from miRecords(*79*) database, signaling pathways, protein interaction networks, and transcription factor/gene regulatory networks, partially based on data integrated from the Human Pathway Database (HPD) (*28*)and the Human Annotated and Predicted Protein Interaction (HAPPI)(*80*) database. All gene sets or pathways are annotated with molecular interaction details whenever available. We integrated the following version of the database OMIM(*77*) (Feb. 2012), GAD(*78*) (Aug. 2011), GeneSigDB(*71*) (v. 4.0, Sept. 2011), MSigDB(*72*) (v. 3.0. Sept. 2010), HPD(*28*) (2009), HAPPI (*80*)(v. 1.4) and miRecords(*79*) (Nov. 2010), which are the latest versions available. An advantage of our work lies in its representation of relationships between pathways, gene signatures, microRNA targets, and/or network modules. These gene-set-based relationships can be visualized as a gene-set association network (GSAN), which provides a "roadmap" for molecular phenotype discovery for specific human diseases. Using colorectal cancer expression data analysis as a case study, we demonstrate how to query PAGED to

discover crucial pathways, gene signatures, and gene network modules specific to colorectal cancer functional genomics.

3.2 Methods

i. Data sources

We show an overview of the data integration process in Figure 1. Gene-set data were collected, extracted, and integrated from five major categories. The pathway data sources were from HPD (*28*), which has integrated 999 human biological pathway data from five curated sources: KEGG, PID, BioCarta, Reactome, and Protein Lounge. The genome-level disease gene relationships were from OMIM (*77*) and GAD (*78*); the transcriptome-level gene signatures were from MSigDB (*72*) and GeneSigDB (*71*); the post-transcriptome-level microRNA data were from miRecords (*79*); and the proteome level data was from an integrated protein interaction database HAPPI (*80*), which has integrated HPRD, BIND, MINT, STRING, and OPHID databases.

ii. Gene-set data integration

We treat as gene sets all groups of genes, including disease-associated genes, pathway genes, gene signatures, microRNA-targeted genes, and PPI sub-network modules. The raw files from those data sources have various formats including plaintext, XML, and table. We have written Perl/Java parsers to convert them into a common tab-delimited textual format to ensure syntactic-level data compatibility. To integrate across different databases, we mapped the gene/protein IDs in all databases to official gene symbols. The gene-set gene data is stored in the backend ORACLE11g relational database. As of the current release, PAGED contained a total of 25,242 gene sets, 61,413 genes, 20 organisms, and 1,275,560 records. All gene set members are represented by the official gene symbols. All PAGED gene sets were assigned unique PAGED-specific identifiers.

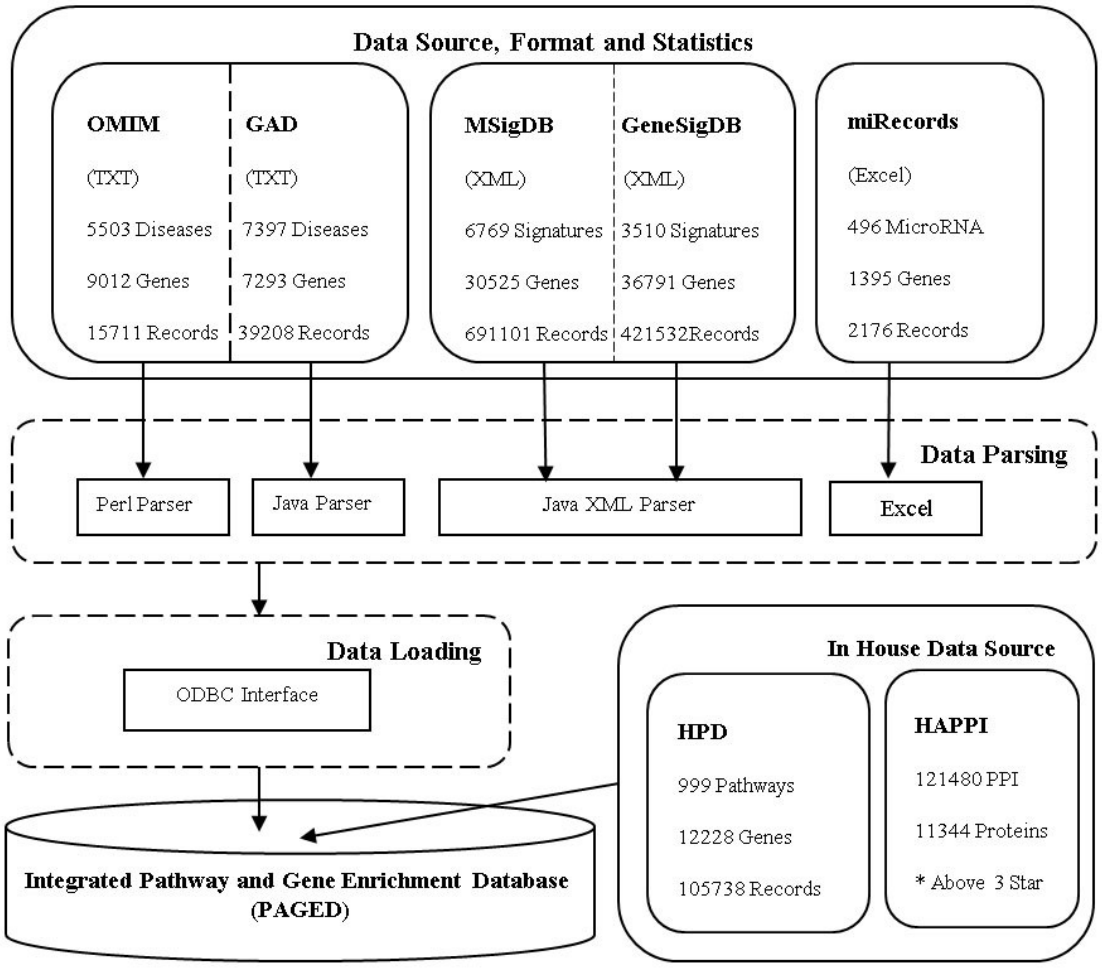

Figure 3-1. The workflow of gene-set data integration and the basic statistics of gene-set data sources.

iii. Online software designing

The PAGED platform follows a multi-tiered design architecture. The backend was implemented as PL/SQL packages on an Oracle 11g database server and the PAGED application middleware was implemented on the Oracle Application Express (APEX) server, which bridged between the Apache webserver and the Oracle database server.

iv. Gene-set similarity measurement

Referring to the pathway similarity definition introduced in (*28*), the similarity score $S_{i,j}$ of two different gene sets is defined by the following formula:

$$S_{i,j} = \alpha \times S_L + (1-\alpha) \times S_R = \alpha \times \frac{|P_i \bigcap P_j|}{|P_i \bigcup P_j|} + (1-\alpha) \times \frac{|P_i \bigcap P_j|}{\min\{|P_i|, |P_j|\}}, (i \neq j) \quad (1)$$

Here, P_i and P_j denote two different gene sets, while $|P_i|$ and $|P_j|$ are the number of genes in each of these two gene sets. Their intersection $P_i \cap P_j$ denotes a common set of genes, while their union $P_i \cup P_j$ is calculated as $|P_i| + |P_j| - |P_i \cap P_j|$. Here, α is a weight coefficient among [0, 1], which is used to count varying degree of contributions from calculations based both on the *overlap* (left item S_L) and the *cover* (right item S_R). S_L is well-known as the Jaccard coefficient (*81*), which is often used to evaluate the similarity between two sets (*82*). When a larger gene set covers a smaller one, we expect their similarity score to be high enough to identify them. In this situation, although the left item S_L is a small number, the right item S_R will be counted as 1.0 to make the final similarity score higher according to our definition in Equation (1), when taking an appropriate α value. We found that when α fell in the interval of [0.7, 0.9], the score distribution would be close to a Poisson distribution. As we know, a Poisson distribution expresses the probability of a number of events occurring during a fixed period of time if these events occur with a known average rate and are time-independent since the last event. Therefore, we chose the middle value, $\alpha = 0.8$, for the rest of the analysis. Our previous HPD paper (*28*) also validates the choice of 0.8 as the pathway similarity measurement.

v. Microarray data

Here we use colorectal cancer (CRC) expression data analysis as a case study to show how to discover crucial pathways, gene signatures, and gene network modules specific to colorectal cancer functional genomics. We downloaded a colorectal cancer microarray dataset GSE8671 from Gene Expression Omnibus, GEO (http://www.ncbi.nlm.nih.gov/geo/) (*83*). This microarray dataset compared the transcriptome data of 32 prospectively collected adenomas with those of the normal mucosa from the same individuals. Hence, we have 32 CRC samples and 32 normal samples. We used maximal expression values for the same proteins mapped from different Probe IDs, the Affy package in BioConductor for quantile normalization, the built-in MicroArray Suite (MAS5) for background correction, and Limma in BioConductor for differential analysis, the result of which is represented as fold changes (FC) of CRC samples vs. normal samples.

vi. Differential gene-set expressions and gene-set association network

We use ABS_FC to denote the absolute value of fold change for each gene. We then define differential gene-set expressions here as

NORM_ABS_FC: The *p**-norm of ABS_FC of all the available differential gene expressions in a gene set.

Usually, *p*-norm = $(\sum_{i=1}^{n}(x_i)^p)^{\frac{1}{p}}$

For unification, we modify it as

$$p\text{*-norm} = ((\frac{1}{n}\sum_{i=1}^{n}(x_i)^p))^{\frac{1}{p}} \quad (2)$$

In the implementation, $p = 6$ performs the best at accentuating highly differential expressions in a gene set.

To visualize the relationships between gene sets, we define a gene-set association network (GSAN) as a network of associations between different gene sets, in which the network element representation is as follows:

- Node: Gene set
- Edge: Association between two gene sets
- Node size: Gene-set scale (Counting genes in each gene set)
- Node color: Differential gene-set expression (NORM_ABS_FC)
- Node line color: Gene-set data source
- Edge width: Similarity score

3.3 Results

i. Database content statistics

Table 3-1 lists the detailed statistics for each data source and the overlap between each pair. For example, MSigDB contains 30,525 genes and GeneSigDB contains 36,791 genes. The number of overlapping genes between these two databases is 17,209. We found a synergistic effect from integrating these two signature databases, resulting in greatly increased gene-set coverage. The same effect was observed for all the remaining pair comparisons. These data sources proved to be complementary.

Table 3-1. Number of overlapping genes between different data sources

	OMIM	GAD	MSigDB	GeneSigDB	miRecords	HPD	HAPPI*
OMIM	9012	1862	3489	2792	231	2559	3849
GAD		7293	6821	6450	432	3202	4922
MSigDB			30525	17209	759	6229	10677
GeneSigDB				36791	900	5904	10395
miRecords					1395	443	725
HPD						12228	10512
HAPPI							21955

* Only PPIs of over 3-star quality are considered here; to calculate the overlap, protein IDs from HAPPI have been first converted to gene symbols.

ii. Gene-set scale distributions

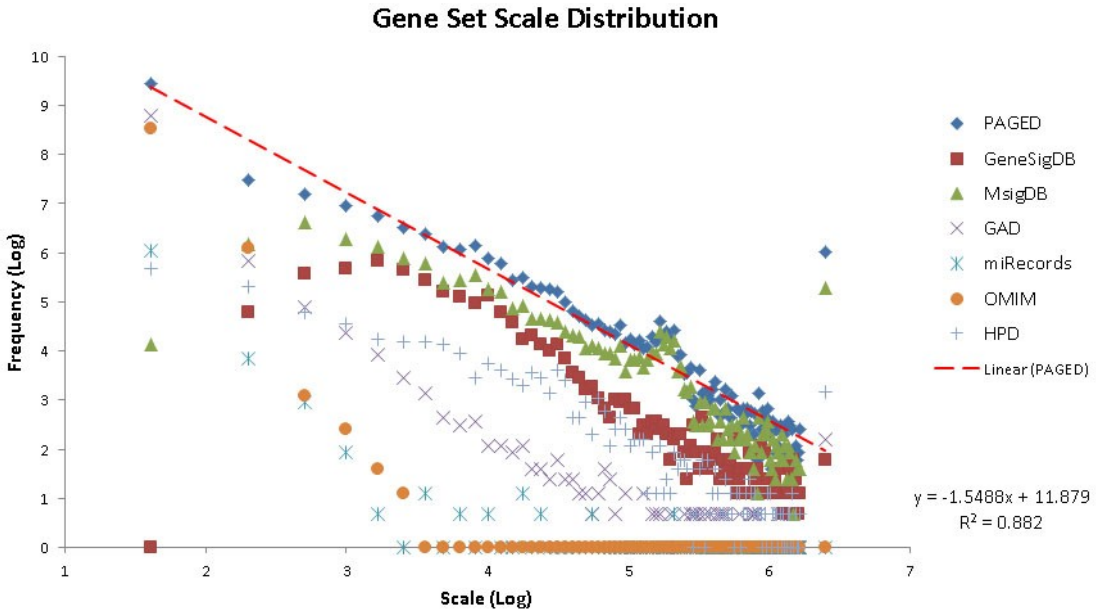

Figure 3-2. Gene-set scale distributions for PAGED molecule data
A gene-set scale refers to the number of molecules (i.e., genes) involved in a given gene set. The frequency on the y-axis refers to the count of all gene sets falling into the category of a particular gene-set scale size on the x-axis. The distributions are plotted under log scale for both the x-axis and y-axis. The linear trend line in red represents linear regression of PAGED distribution and the linear equation and its R-Square are listed.

The gene-set scale can reflect the integrality of information content of a biological topic. In this study, we define gene-set scale as the number of molecules (i.e., gene symbols) in a gene set. We performed a statistical analysis of the gene-set scale distributions of both PAGED and of its individual data sources. Figure 3-2 shows that each data source taken by itself is not very scale-free, especially for OMIM, GAD, and miRecords for higher

scales and HPD, GeneSigDB, and MSigDB for lower scales. The scale distribution of PAGED is relatively scale-free on both the low end and the high end with a linear regression *R*-squared of 0.88. Additionally, the distribution of PAGED always lies above those of its data sources, indicating that the integration has enriched the number of gene sets in all scales instead of exhibiting a bias towards one particular scale. These observations indicate that the integration process of PAGED has resulted in a database that can take account of different gene set scales.

iii. Online functionalities

In Figure 3-3, we show the user interfaces of the PAGED website. It supports both disease-based search and user-defined gene-list search. If users search the disease term in the home page (Figure 3-3A), PAGED will retrieve a list of related gene sets by directly matching the disease term with all the gene-set names; if users instead search a disease term in the advanced search page (Figure 3-3D), PAGED will first retrieve disease-relevant genes from OMIM and GAD and then use those genes to query the whole database, which will retrieve a gene-set list based on disease gene profiles that is more comprehensive than that of either OMIM or GAD individually. Users can also search PAGED using multiple genes in the home page (by delimiting them with a comma) to retrieve a list of related gene sets with the hits number and similarity scores (Figure 3A). In addition, users can upload a file of their genes with one gene per line on the advanced search page (Figure 3-3D) to perform the gene-based search. In the advanced gene-based search (Figure 3-3D), user can also perform an organism specific search though the majority of the gene-sets are human related. All the gene sets are hyperlinked to the original database, where user can further examine the detailed annotations of that specific gene set.

Upon executing the queries, PAGED can retrieve a list of related gene sets in an HTML table (Figure 3-3B, C) with their specific organism information included, which are downloadable as a comma-separated value (CSV) file. On the same page, there are links for downloading all the genes in those gene sets and the association between each gene set. In the gene set association downloading page, a simple heat map is provided for the visualization of gene set similarities.

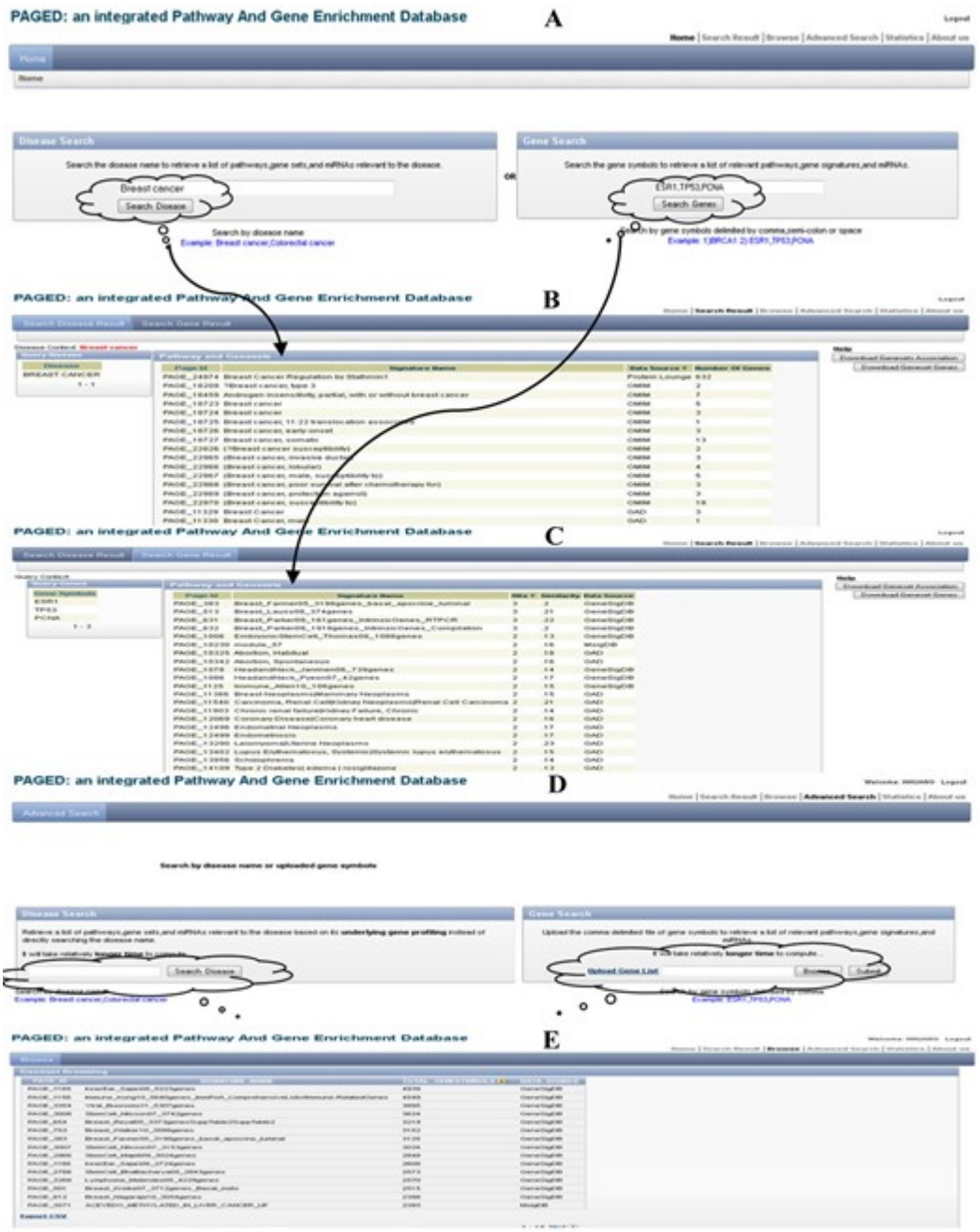

Figure 3-3. An overview for the core functionality of the online PAGED website
(A) The PAGED home page providing search by either disease name or gene list; (B) a webpage containing the list of gene sets retrieved as a result of a disease query; (C) a webpage containing the list of gene sets retrieved as a result of a gene list query; (D) an advanced search page in which the user can either search disease name or upload a gene-list to search; (E) a browse page listing the gene sets, their data source and number of genes.

iv. Case studies

The following case studies use colorectal cancer expression data analysis as a case study to demonstrate how to discover crucial pathways, gene signatures, and gene network modules specific to colorectal cancer functional genomics.

Case study I: Searching disease-associated gene sets based on gene-set names

Using the standard query box provided at the PAGED home page, one can search for *colorectal cancer* in all biological gene sets. PAGED returns a list of gene sets, which can be ordered by decreasing number of genes contained by each gene set. In total, 45 gene sets from three data sources (i.e., OMIM, GAD and KEGG) have been retrieved. Not surprisingly, most of them are disease-related gene sets from either OMIM or GAD. Only 1 (i.e., "Colorectal cancer pathway") out 45 is from KEGG. The top 10 search results are listed in Table 3-2.

Table 3-2. Top 10 search results by querying colorectal cancer at the home page

Gene-set Name	# of Genes	Data Source
Colorectal cancer	433	GAD
Colorectal cancer	134	KEGG
Colorectal cancer	14	OMIM
Colorectal cancer, somatic	12	OMIM
Colorectal cancer, hereditary non-polyposis, type 8	7	OMIM
Colorectal cancer, susceptibility to	7	OMIM
Colorectal cancer, hereditary non-polyposis, type 6	6	OMIM
Breast and colorectal cancer, susceptibility to	5	OMIM
Colorectal Cancer	5	GAD

Case study II: Searching disease-associated gene sets based on gene-set components

Next, a user can search with the same term *colorectal cancer* on the advanced search page, which uses the disease's gene profile to search for gene sets. PAGED first obtained 203 colorectal cancer related genes from OMIM and GAD. Then, it used those genes to retrieve a total of 4,932 gene sets with at least 2 hits. Since we were more interested in gene sets other than disease terms, we excluded those gene sets from OMIM and GAD for further analysis. To rule out the possibility that those gene sets were hit randomly, we did a Fisher's exact test to calculate the *p*-value between those 203 genes and every retrieved gene set. Finally, we obtained 3,879 gene sets with a *p*-value <0.05 and hits ≥ 2.

These gene sets are from all data sources, including MSigDB, GeneSigDB, miRecords, and all pathway data sources from HPD. Both the number of gene sets and their variety support the conclusion that advanced disease search based on gene profiles are more comprehensive than a simple disease search. For other disease query, a similar procedure will be followed to calculate the *p-value* and the number of hits on the fly.

Table 3-3 shows the top results ranked by decreasing number of hits from each data source. Protein Lounge suggests "Molecular Mechanisms of Cancer," "Akt Signaling," and other important pathways in colorectal cancer; BioCarta suggests "wnt signaling pathway"; and NCI Nature curated suggests "Canonical Wnt signaling pathway." These are all very important pathways in colorectal cancer development (*84*). Similarly, "Colorectal cancer" and "p53 signaling pathway" from KEGG, "SIGNAL_TRANSDUCTION" and "KEGG_PATHWAYS_IN_CANCER" from MSigDB, and cancer-related signatures/microRNA from GeneSigDB/miRecords reveal a comprehensive picture of the important gene sets involved in colorectal cancer. Thus, the results of the advanced search yield more insights about colorectal cancer mechanisms than those of the simple search.

Table 3-3. Top search results of colorectal cancer advanced search

Gene-set Name	Hits	P value	FDR	Data Source
Molecular Mechanisms of Cancer	38	2.48E-17	7.04E-10	Protein Lounge
PI3K Signaling	33	2.01E-13	7.04E-10	Protein Lounge
Akt Signaling	27	9.6E-13	7.04E-10	Protein Lounge
ERK Signaling	24	1.53E-10	7.04E-10	Protein Lounge
GSK3 Signaling	23	1.32E-13	7.04E-10	Protein Lounge
inactivation of gsk3 by akt causes accumulation of b-catenin in alveolar macrophages	9	3.7E-11	7.04E-10	BioCarta
atm signaling pathway	8	6.28E-11	7.04E-10	BioCarta
wnt signaling pathway	7	7.7E-09	7.04E-10	BioCarta
cell cycle: g2/m checkpoint	7	2.14E-08	7.04E-10	BioCarta
cell cycle: g1/s check point	7	2.14E-08	7.04E-10	BioCarta
Canonical Wnt signaling pathway	8	9.24E-10	7.04E-10	NCI-Nature
Presenilin action in Notch and Wnt signaling	8	3.16E-08	7.04E-10	NCI-Nature
Plasma membrane estrogen receptor signaling	7	1.41E-08	7.04E-10	NCI-Nature
FOXM1 transcription factor network	7	2.48E-07	7.04E-10	NCI-Nature
LPA receptor mediated events	7	1.45E-06	7.04E-10	NCI-Nature
Metabolism of xenobiotics by cytochrome P450	20	3.3E-25	7.04E-10	KEGG

Drug metabolism - cytochrome P450	17	4.96E-21	7.04E-10	KEGG
Bladder cancer	15	3.29E-18	7.04E-10	KEGG
Cytokine-cytokine receptor interaction	15	1.39E-06	7.04E-10	KEGG
Colorectal cancer	14	4.43E-14	7.04E-10	KEGG
p53 signaling pathway	14	4.92E-14	7.04E-10	KEGG
Prostate cancer	14	1.66E-12	7.04E-10	KEGG
Xenobiotics	5	3.32E-08	7.04E-10	Reactome
Formation of incision complex in GG-NER	5	5.75E-06	7.04E-10	Reactome
Global Genomic NER (GG-NER)	5	5.75E-06	7.04E-10	Reactome
Dual incision reaction in GG-NER	5	5.75E-06	7.04E-10	Reactome
Exocytosis of Alpha granule	5	0.000217	1.95E-08	Reactome
SIGNAL_TRANSDUCTION	55	8.36E-28	7.04E-10	MsigDB
BIOPOLYMER_METABOLIC_PROCESS	49	4.16E-22	7.04E-10	MsigDB
KEGG_PATHWAYS_IN_CANCER	43	9.9E-46	7.04E-10	MsigDB
NUCLEOBASENUCLEOSIDENUCLEOTIDE_AND_NUCLEIC_ACID_METABOLIC_PROCESS	41	2.16E-20	7.04E-10	MsigDB
NUCLEUS	41	1.8E-18	7.04E-10	MsigDB
Immune_Kong10_5640genes_ImmPort_ComprehensiveListofImmune-RelatedGenes	114	3.61E-49	7.04E-10	GeneSigDB
Lymphoma_Melendez05_4229genes	81	1.57E-39	7.04E-10	GeneSigDB
Breast_Farmer05_3198genes_basal_apocrine_luminal	66	1.08E-21	7.04E-10	GeneSigDB
Ovarian_Crijns09_2394Genes_17PathwayPredictor	57	7.94E-30	7.04E-10	GeneSigDB
StemCell_Nilsson07_3742genes	45	4.86E-07	7.04E-10	GeneSigDB
hsa-miR-19a	3	1.49E-05	8.43E-09	miRecords
[hsa-miR-21]	3	0.000116	8.43E-09	miRecords
hsa-miR-204	3	0.000164	1.95E-08	miRecords
hsa-miR-21	3	0.000953	2.72E-07	miRecords
hsa-miR-125b	3	0.003089	2.72E-07	miRecords

Case study III: Searching gene sets similar to user-defined query gene sets

To use the gene-based search from PAGED, we first analyzed a colorectal cancer microarray dataset GSE8671 with BioConductor to identify the differential genes. We selected the top 100 genes ranked by the absolute fold change with *p*-values less than 0.05. After querying PAGED with those 100 genes, we obtained 1,707 gene sets, out of which 1,152 also satisfied Fisher's exact test of a *p*-value less than 0.05. Those gene sets span from all the data sources except BioCarta and miRecords. Table 4 lists the top results ranked by the number of hits. Most of them are cancer-related gene sets. Specifically, "SABATES_COLORECTAL_ADENOMA_DN" and "SABATES_COLORECTAL_ADENOMA_UP" from MSigDB and

"Intestine_Vecchi07_1024genes" and "Colon_Kim04_235genes" from GeneSigDB supports the importance of those 100 query genes to colorectal cancer. This case study also shows the complementary nature of MSigDB and GeneSigDB and thus the benefit of integrating them, which has also been proved by (*85*)

Case study IV: Building disease-specific gene-set association networks (GSANs) based on gene-set similarities

With the unique top 50 gene sets related to colorectal cancer from disease search and

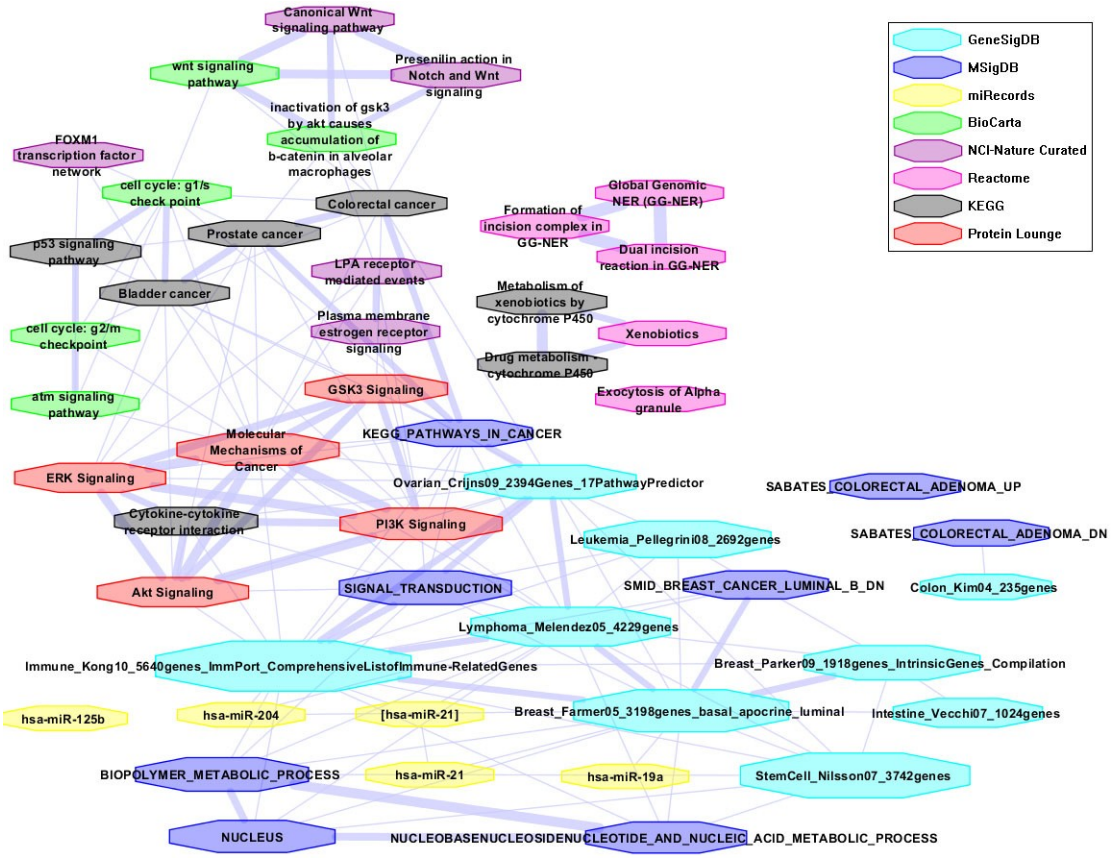

Figure 3-4. CRC-specific gene-set association network (GSAN) on the top gene sets from colorectal cancer study.

Node size: Gene-set scale (Counting genes in each gene set); Node color: Gene-set data source; Edge width: Similarity score (≥ 0.1). All gene sets are highly connected to each other, suggesting their collaborative functions in colorectal cancer.

gene search (Table 3-3 and Table 3-4), we next investigated the gene-set associations between them; 863 associations were found by overlapping the gene symbols between each pair of gene sets, out of which 642 also satisfied Fisher's exact test of a *p*-value and

FDR less than 0.05. A network visualization using Cytoscape (*86*) is shown in Figure 3-4. Most of those gene sets are connected to one another, and a few share a large number of genes, suggesting that they form a collaborative unit in colorectal cancer.

Table 3-4. Top search results of gene-based search from colorectal microarray datasets

Gene-set Name	Hits	P value	FDR	Data Source
SABATES_COLORECTAL_ADENOMA_DN	58	4.57E-96	2.76E-10	MsigDB
Breast_Farmer05_3198genes_basal_apocrine_luminal	35	2.91E-13	2.76E-10	GeneSigDB
SABATES_COLORECTAL_ADENOMA_UP	34	1.62E-57	2.76E-10	MsigDB
Immune_Kong10_5640genes_ImmPort_ComprehensiveListofImmune-RelatedGenes	34	3.56E-08	2.76E-10	GeneSigDB
Leukemia_Pellegrini08_2692genes	32	1.28E-15	2.76E-10	GeneSigDB
Intestine_Vecchi07_1024genes	28	3.91E-23	2.76E-10	GeneSigDB
Viral_Buonomo11_5307genes	25	6.45E-05	0.000109	GeneSigDB
SMID_BREAST_CANCER_LUMINAL_B_DN	23	4.16E-19	2.76E-10	MsigDB
Lymphoma_Melendez05_4229genes	22	2.03E-06	2.76E-10	GeneSigDB
Colon_Kim04_235genes	21	5.18E-30	2.76E-10	GeneSigDB
Breast_Parker09_1918genes_IntrinsicGenes_Compilation	21	1.18E-08	2.76E-10	GeneSigDB

Case study V: Prioritizing disease-associated gene sets by using differential gene-set expressions

First, the differential gene expression value (ABS_FC) for each gene in a gene set is calculated from the differential analysis based on the microarray data GSE8671. Second, the differential gene-set expression value (NORM_ABS_FC) for each gene set in the CRC-specific GSAN is calculated by using Equation (2). Third, a CRC-specific GSAN with differential gene-set expressions is shown in Figure 3-5, in which node size represents gene-set scale (Counting genes in each gene set); node color represents differential gene-set expression (NORM_ABS_FC); node line color represents the gene-set data source; and edge width represents the similarity score. By considering differential gene-set expressions for each gene set, we prioritize top-selected gene sets as shown in Table 3-5. Most of top-ranked gene sets are closely related to colon tissue, colorectal cancer, or other cancers, which implies that the database can not only support comprehensive disease-associated gene-set searching and browsing, but also accurate,

disease-specific gene-set prioritizing by using the concept of differential expressions at the gene-set level.

Table 3-5. Top 20 gene sets ranked by differential gene-set expressions in the CRC-specific gene-set association network (GSAN)

Gene-set name	Scale	Data Source	NORM_ABS_FC
Colon_Kim04_235genes	151	GeneSigDB	48.58225017
SABATES_COLORECTAL_ADENOMA_DN	292	MsigDB	43.9233159
SIGNAL_TRANSDUCTION	1598	MsigDB	32.5957784
Leukemia_Pellegrini08_2692genes	2122	GeneSigDB	31.65148925
SABATES_COLORECTAL_ADENOMA_UP	142	MsigDB	31.65000681
Breast_Parker09_1918genes_IntrinsicGenes _Compilation	1734	GeneSigDB	20.85621131
Lymphoma_Melendez05_4229genes	2570	GeneSigDB	19.38449282
Breast_Farmer05_3198genes_basal _apocrine _luminal	3125	GeneSigDB	18.93820407
SMID_BREAST_CANCER_LUMINAL_B_DN	648	MsigDB	18.13762096
Intestine_Vecchi07_1024genes	796	GeneSigDB	16.68882931
Ovarian_Crijns09_2394Genes _17PathwayPredictor	1586	GeneSigDB	15.29529767
hsa-miR-204	19	miRecords	14.37015815
StemCell_Nilsson07_3742genes	3624	GeneSigDB	12.47045771
Immune_Kong10_5640genes_ImmPort _ComprehensiveListofImmune-RelatedGenes	4549	GeneSigDB	11.91186233
cell cycle: g1/s check point	53	BioCarta	9.84279867
Bladder cancer	89	KEGG	7.885181064
Drug metabolism - cytochrome P450	94	KEGG	7.837851592
Metabolism of xenobiotics by cytochrome P450	103	KEGG	7.837805455
hsa-miR-21	34	miRecords	7.001844224
KEGG_PATHWAYS_IN_CANCER	328	MsigDB	6.792625895

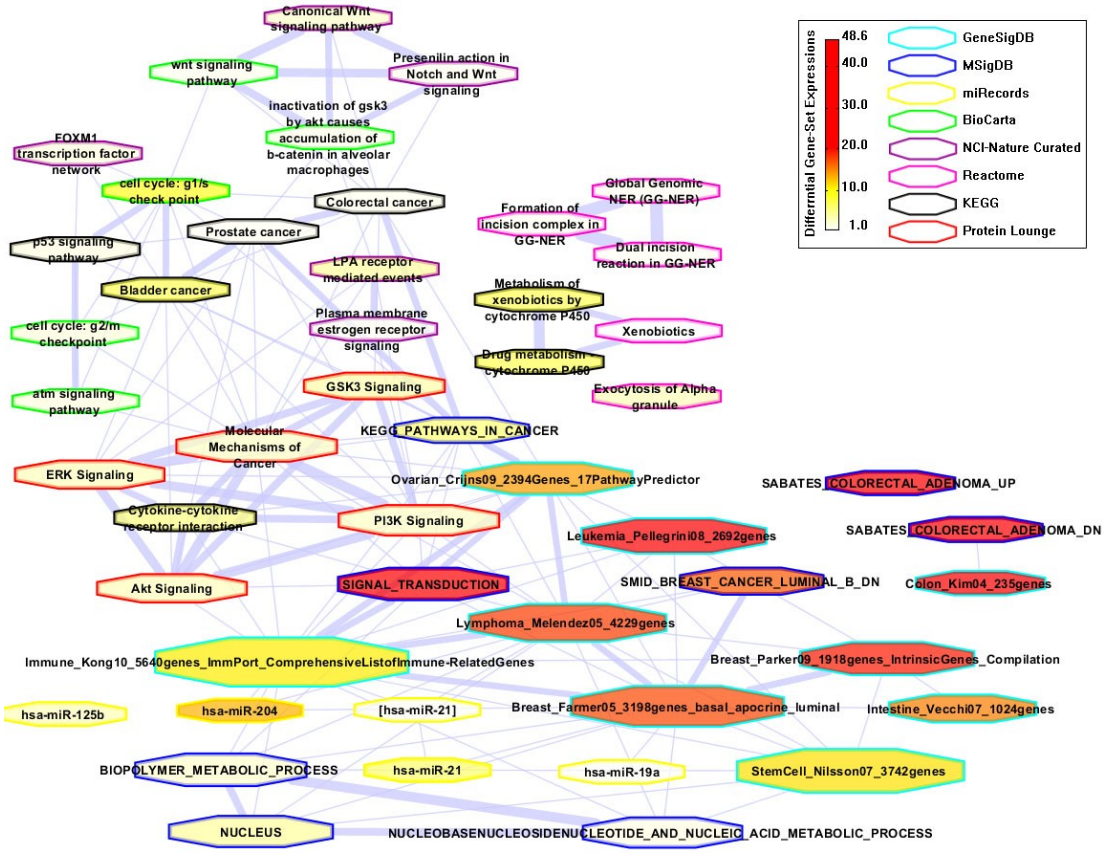

Figure 3-5. CRC-specific gene-set association network (GSAN) with differential gene-set expressions.
The differential gene expressions are from the differential analysis based on the microarray data, GSE8671. Node size: Gene-set scale (Counting genes in each gene set); Node color: Differential gene-set expression (NORM_ABS_FC); Node line color: Gene-set data source; and Edge width: Similarity score (≥ 0.1).

3.4 Conclusions

We developed PAGED, an online database that provides the most comprehensive public compilation of gene sets. In the current release, PAGED contains a total of 25,242 gene sets, 61,413 genes, 20 organisms, and 1,275,560 records from five major categories: the pathway data from HPD, genome-level disease data from OMIM and GAD, transcriptome-level gene signatures from MSigDB and GeneSigDB, the post-transcriptome microRNA data from miRecords, and proteome-level data from HAPPI. The number of overlapping genes between each data source, gene-set scale distribution, and case study in colorectal cancer shows the synergistic effect of integrating data

sources, which greatly facilitate access to gene-set-based prior knowledge. The current PAGED software can help users address a wide range of gene-set-related questions in human disease biology studies.

In the future, one could improve gene-set similarity algorithms by using a global PPI network to calculate their distance. This would provide a more robust measurement for web interface development. A disease browsing function based on disease ontology and a network visualization function to show the gene-set association dynamically could also be added. The final goal is to perform multi-scale network modeling for molecular phenotype discoveries by integrating differential expressions with pathway and network topologies. The current release of PAGED provides a solid foundation for us to develop third-generation pathway analysis tools (*27*).

Chapter 4. Biomarker discovery with Network Expansion and Pathway Enrichment Analysis

This section is based on the published work at (*51*). JYC guided the research team by providing ideas and feedback along the way, and revised the manuscript. HH analyzed the microarray dataset, generated the protein interaction network, ranked the disease proteins, and wrote the manuscript. XW conducted the pathway analysis and wrote the manuscript. TW and SDL helped analyze the result. RP helped maintain the database. CR helped curate the disease genes.

4.1 Introduction

Microarrays make possible the discovery of new functions and pathways of known genes, as they measure all the transcriptional activity in a biological sample (*87*). This high-throughput procedure can be used in medical diagnostics, in biomarker discovery, and in investigating the ways a drug, disease, polymorphism or environmental condition affects gene expression and function (*88, 89*). However, one challenge has arisen because microarray technology generates a large amount of transcriptional data, which is hard to interpret for the results to gain insights into biological mechanisms (*90*). As a result, researchers have sought to analyze microarray data through the use of modern computational tools and statistical methods.

In many cases, crucial genes show relatively slight changes, and many genes selected from differential analysis between groups of samples (e.g. normal vs. disease) by measuring the expression level statistically are also poorly annotated (*88*). From a biological perspective, functionally related genes often display a coordinated expression to accomplish their roles in the cell (*91*). Hence, to translate such lists of differentially expressed genes into a functional profile able to understand the underlying biological phenomena, one approach to aid interpretation is to look for changes in a group of genes with a common function (*88*).

Gene set enrichment analysis (GSEA) is one of the most widely used methods for identifying both statistically and biologically significant genes from high-throughput data such as gene-expression assays (*90*). GSEA relies on pre-defined gene sets, while neglect

gene/protein interaction, pathway upstream or downstream information. Furthermore, GSEA still assumes that more differentially expressed genes are more crucial to the biology, which is not always true (*65*). Currently, gene expression signature analysis and pathway analysis remain two separate processes.

From a view of network biology (*92*), cancer genes and proteins do not function in isolation; instead, they work in interconnected pathways and molecular networks at multiple levels (*93*), one study re-characterized them in a molecular interaction network for BRCA, and identified HMMR as a new susceptibility locus (*94*). Another study integrated protein interaction network and gene expression data to improve the prediction of BRCA metastasis (*95*). These works suggest that protein interaction networks and pathways, although noisy, incomplete and static, can serve as a molecular-level conceptual roadmap to guide future microarray analysis (*7*).

In this work, we present an innovative approach - network expansion and pathway enrichment analysis (NEPEA) for integrative microarray analysis. We assume that organized knowledge will help microarray data analysis in significant ways, and the organized knowledge could be represented as molecular interaction networks or biological pathways. Based on this hypothesis, we develop the NEPEA framework based on network expansion (*96*) from the human annotated and predicted protein interaction (HAPPI) database (*80*), and pathway enrichment from the human pathway database (HPD) (*28*).

We use a recently-published microarray dataset (GSE24215) related to insulin resistance and type 2 diabetes (T2D) as case study, since this study provided a thorough experimental validation for both genes and pathways identified computationally from classical microarray analysis and pathway analysis (*97*). In this study, skeletal muscle samples were collected in all participants (n = 20) in both the basal and insulin-stimulated state before and after bed rest. We perform the NEPEA analysis for this dataset based on the results from the classical microarray analysis to identify biologically significant genes

and pathways. Our findings are not only consistent with the original findings mostly, but also obtained more supports from other literatures.

4.2 Methods

The NEPEA method has three main components: 1) classical microarray analysis for data preprocessing consisting of quality control, normalization and differential analysis, 2) network expansion analysis for significant gene identification consisting of disease gene curation, network construction and significance score calculation, and 3) pathway enrichment analysis consisting of pathway search, pathway differential analysis and ranking. Using the microarray dataset - GSE24215 as an example, we introduce the detailed steps below:

i. Microarray data preprocessing

Quality Control

We use AffyQCReport (applicable for Affymetrix platform) and ArrayQualityMetrics (applicable for Agilent platform) packages in Bioconductor to generate three plots to detect bad chips for each microarray dataset as: 1) examine a heat map that shows array-array Spearman rank correlation coefficients. The map enabled us to plot outliers, failed hybridizations, and mis-tracked samples; 2) make a box plot of all perfect match intensities. The plot enabled us to detect outliers in terms of average intensity; and 3) make a distribution plot of kernel density estimates for perfect match intensities, which enables us to detect outliers in terms of shaped density. After applying ArrayQualityMetrics packages into quality control for microarray dataset - GSE24215, total 3 suspects out of 48 samples are flagged, which are kicked off as bad chips.

Normalization

We use Quantile normalization to normalize all the four qualified microarray datasets; MAS5 for Affymetric platform and normexp for Agilent platform on background correction. We also perform the steps background correction, normalization, probe specific correction, and summary value computation as following: 1) bgcorrect.method: mas; 2) normalize.method:quantiles; 3) pmcorrect.method:pmonly; and 4) summary.method:mas.

Differential analysis

We use Limma (Linear Models for Microarray Data) package (*98*) in Bioconductor to identify differentially-expressed genes for each clinical group comparison from the qualified and normalized microarray datasets as 1) The package Limma uses an approach called linear models to analyze designed microarray experiments; 2) For statistical analysis and assessing deferential expression, Limma uses an empirical Bayes method for more stable inference and improved power, especially for experiments with small numbers of arrays; and 3) Differential genes are obtained by using the filters with p-Value <= 0.05, Fold Change (FC) >= 1.3, and Average Expression Level (AEL) >= 40% after applying Limma package in Bioconductor. Average expression levels (AEL>=40%) have been checked to ensure the presences of the differential genes in the tissue - muscle. Duplicated genes with lower fold changes are eliminated, which implies that only the highest fold change for one gene will be kept. For microarray dataset - GSE24215, we get 495 differential genes from insulin before-bed (IBB) group, and 930 differential genes from insulin after-bed (IAB) group

ii. Network expansion analysis

Disease gene curation

The network expansion analysis is knowledge-guided approach, which relies on the disease-associated genes. Here we use T2D as an example to demonstrate how to curate disease-associated genes, but our method can be applied to any other disease phenotypes. We curate T2D-associated genes from OMIM (http://www.ncbi.nlm.nih.gov/omim) manually, evaluates them semi-automatically through searching in PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) as following: 1) Query: ("Type II Diabetes"[All Fields] OR "Type 2 Diabetes"[All Fields]) AND (prefix star[prop] OR prefix plus[prop]); 2) Results: Records (Entries) -> Genes (Gene Symbol) -> Proteins (Uniprot ID); 3) GENE: Gene name, linked to GeneCards.org; 4) UNIPROT: Uniprot ID, linked to UniProt.org; 5) PUBMED: Count number of references where both term ("Type II Diabetes" OR "Type 2 Diabetes") AND "GENE" appeared in PubMed, linked to PubMed; and 6) Obtain interactions from HAPPI 1.31 for these T2D-associated genes (seed genes) curated from OMIM.

Network reconstruction

We construct a T2D-specific protein-protein interaction (PPI) network by using Oracle SQL Developer with high-quality interaction data in HAPPI version 1.31 and map differentially-expressed genes onto the T2D-specific PPI network by using Cytoscape as following: 1) Expand 39 seed genes (PUBMED >=50) in HAPPI 1.31 (4-Star, h-Score >=0.75), and obtain 702 genes (including 32 seed genes); 2) The left 7 seed genes are also added into the network in order to show their expressions; and 3) Construct a T2D-specific PPI network with 709 nodes and 944 edges, by using Nearest Neighbor Expansion (NNE) approach (*96*).

Significant gene identification

We measure and rank all the differential genes in a T2D-specific protein-protein interaction (PPI) network by considering both differential expressions and network properties. Differential genes are obtained by applying filters with p-Value <= 0.05, Average Expression Level (AEL) >= 40%, and Absolute Fold Change (ABS_FC) >= 1.3. Duplicated genes with lower fold changes have been eliminated, which implies that only the highest fold change for one gene will be kept. The T2D-specific PPI network is reconstructed by expanding all the seed genes curated from OMIM (PUBMED >=0), in HAPPI_1.31 (3-Star) (Confidence: h-Score >=0.45)

We define Gene Significance Score (integrating both gene expression fold change - FC and network connectivity - NC) here as:

Sig_Score = $\left(\alpha_1 + log_2^{|FC|}\right) \times log_2(\alpha_2 + NC)$, $|FC|$ = ABS_FC, absolute fold change.

Constant parameters α_1 and α_2 here are for the balance between differential expressions and network properties. α_1=3 and α_2=1 are chosen since this pair can rank known disease genes on the top. Network connectivity (for un-weighted networks) NC: Weight_1 = Number of direct neighbors for each node. Network connectivity (for weighted networks) NC: Weight_2 = Sum of connection strength values on all neighbored edges. In the implementation, we use Weight_2 here. Connection strength here is the confidence for an interaction: h-Score.

iii. Pathway enrichment analysis

Pathway search

We search curated T2D-associated genes by using Oracle SQL Developer with comprehensive integrated pathway data in HPD version 2.1 (including pathway data from NCI-Nature curated, KEGG, BioCarta, and Protein Lounge), and map differentially-expressed genes onto the pathways obtained. We obtain 92 pathways with (HITS/Pathway Scale) >= 3.5% AND HITS >=2 by querying 39 seed genes (PUBMED >=50) in HPD 2.1.

Pathway differential analysis

We provide average differential gene expressions in a pathway as:

AVG_ABS_FC: The average of ABS_FC of all the available differential gene expressions in a pathway.

We define pathway differential expressions here as:

NORM_ABS_FC: The p*-norm of ABS_FC of all the available differential gene expressions in a pathway

Usually, p-norm = $(\sum_{i=1}^{n}(x_i)^p)^{\frac{1}{p}} = (SUM((x_i)^p)^{1/p}$

For unification, we modify it as p*-norm = $(\frac{1}{n}\sum_{i=1}^{n}(x_i)^p)^{\frac{1}{p}} = (AVG((x_i)^p)^{1/p}$

In the implementation, p = 6 have best performance to emphasize highly differential expressions in a pathway.

We also provide maximal differential gene expressions in a pathway as:

MAX_ABS_FC: The maximum value of ABS_FC of all the available differential gene expressions in a pathway;

and count number of differentially expressed genes as:

CNT_DIFF: Count number of differentially expressed genes (FC >= 1.3 AND p-Value <= 0.05) in a pathway.

We rank all the pathways by their pathway differential expressions - NORM_ABS_FC defined above.

4.3 Results

i. Findings on insulin before-bed (IBB) group

Top-20 differential genes

Totally 495 differential genes are obtained, which are differentially-expressed in Insulin Before Bed (IBB) Group from the microarray dataset – GSE24215. Differential genes are obtained by using filters with p-Value <= 0.05, Fold Change (FC) >= 1.3, and Average Expression Level (AEL) >= 40% after applying Limma package in Bioconductor. Average expression levels (AEL>=40%) have been checked to ensure the presences of the differential genes in the tissue - muscle. Duplicated genes with lower fold changes are eliminated, which implies that only the highest fold change for one gene will be kept. Top-20 differential genes in IBB from GSE24215, ordered by absolute fold change (ABS_FC), are listed in Table 4-1.

Table 4-1. Top-20 differential genes in IBB from GSE24215, ordered by FC (FC>=1.3, p-value <=0.05 and AEL >=40% after applying Limma package in Bioconductor). Note: Gene names are linked to GeneCards.org, UniProt IDs are linked to UniProt.org, and Evidences are linked to PubMed.

Gene Symbol	*p*-Value	FDR	Log2_FC	ABS_FC	Evidences
SOCS3	0.00193	0.08858	2.54455	**5.83426**	28
PDK4	0.00000	0.00074	-2.34193	**5.06980**	16
THBD	0.00001	0.00243	2.25714	**4.78043**	0
CISH	0.00013	0.01380	2.19425	**4.57651**	0
G0S2	0.00000	0.00003	2.05403	**4.15264**	0
MYC	0.00064	0.04234	1.97513	**3.93164**	23
PDE4B	0.00000	0.00042	1.82895	**3.55280**	0
ADAMTS4	0.00061	0.04111	1.76371	**3.39569**	1
GADD45A	0.00002	0.00373	1.76132	**3.39008**	0
RGS16	0.00217	0.09630	1.72508	**3.30598**	1
EGR1	0.01342	0.29638	1.71863	**3.29125**	3
HES1	0.00000	0.00012	1.71837	**3.29065**	1
CCL2	0.00019	0.01796	1.71466	**3.28219**	111
KLF15	0.00000	0.00000	-1.66849	**3.17882**	3
PYCR1	0.00000	0.00000	1.66356	**3.16797**	0
CITED2	0.00000	0.00000	-1.65106	**3.14064**	0
OTUD1	0.00006	0.00778	-1.56650	**2.96186**	0
ARRDC4	0.00000	0.00000	1.51143	**2.85092**	0
NR1D1	0.00000	0.00003	-1.50730	**2.84277**	1
PIK3R1	0.00000	0.00000	1.50274	**2.83379**	9

Top-20 significant genes

Totally 130 significant genes in IBB from GSE24215 are obtained from all the differential genes in a T2D-specific protein-protein interaction (PPI) network, measured by using significant score (considering both differential expressions and network properties). The T2D-specific PPI network is reconstructed by expanding all the seed genes curated from OMIM (PubMed >=0), in HAPPI_1.31 (3-Star) (Confidence: h-Score >=0.45). Top-20 significant genes in IBB from GSE24215, ordered by significant score (Sig_Score), are listed in Table 4-2.

Table 4-2. Top-20 significant genes in IBB from GSE24215, ordered by Sig_Score, which is measured in the T2D-specific PPI network
(PubMed >=0, h-Score >=0.45) for all the differential genes (FC>=1.3, p-value <=0.05 and AEL >=40% after applying Limma package in Bioconductor) in IBB from GSE24215.

Gene	*p*-Value	FDR	FC	ABS_FC	Weight_1	Weight_2	Score	Evidence
CCL2	0.00019	0.01796	1.71466	3.28219	98	75.2645	**29.48048**	111
IL6	0.00164	0.08069	0.96338	1.94987	140	112.808	**27.07169**	52
AKT2	0.00133	0.06955	-0.83609	1.7852	104	66.6378	**23.32247**	21
IRS2	0.00011	0.01276	-0.78851	1.72729	60	49.3156	**21.4162**	124
VEGFA	0.01432	0.30888	0.52022	1.43417	57	44.6818	**19.40888**	28
PIK3R1	0	0	1.50274	2.83379	13	11.8228	**16.57294**	9
MYC	0.00064	0.04234	1.97513	3.93164	10	8.8236	**16.39929**	23
UCP3	0.00001	0.00147	-1.04039	2.05679	22	15.2626	**16.25647**	45
SOCS3	0.00193	0.08858	2.54455	5.83426	7	6.3574	**15.96385**	28
UCP2	0	0.00093	-0.80665	1.74915	22	15.7096	**15.46493**	78
SCARB1	0.00115	0.06287	-0.38842	1.30896	30	19.9446	**14.87011**	11
HSD11B1	0.01673	0.33975	-0.46192	1.37737	24	17.478	**14.56683**	20
SORBS1	0	0	-1.13404	2.19472	13	10.08	**14.34464**	2
KLF11	0	0.00001	-0.5958	1.51131	20	14.1964	**14.11588**	8
AQP7	0.00023	0.02066	-0.71786	1.64474	9	7.305	**11.35427**	8
RRAD	0.00018	0.01748	1.45449	2.7406	7	4.8134	**11.31166**	8
LPIN1	0.00898	0.23354	-0.49106	1.40548	13	7.849	**10.98119**	9
SMAD3	0.01644	0.33728	0.4402	1.35679	9	8.1113	**10.96617**	7
ICAM1	0.00219	0.09689	0.74517	1.67617	8	6.539	**10.91482**	4
TNFRSF1A	0.00039	0.02999	0.5936	1.509	8	6.6686	**10.56144**	0

A T2D-significant protein-protein interaction (PPI) network (See Figure 4-1) is reconstructed by connecting Top-20 significant genes in IBB from GSE24215, with and

within the T2D-associated genes (seed genes) curated from OMIM (PubMed >= 50), in HAPPI_1.31 (3-Star) (Confidence: h-Score >=0.75)

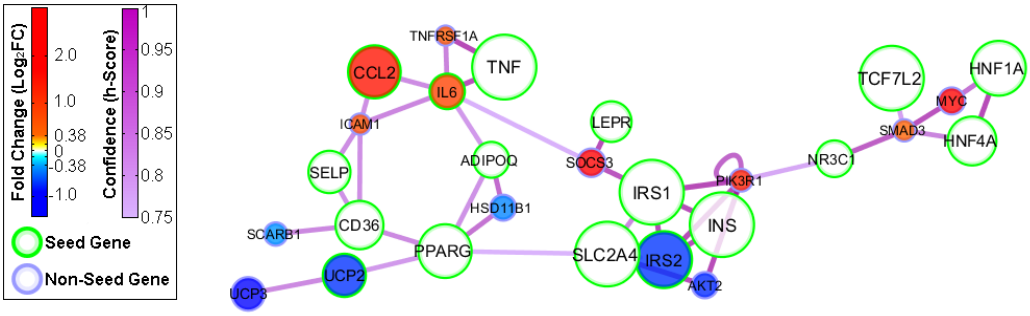

Figure 4-1. Top-20 significant genes in IBB from GSE24215, interacted with T2D-associated genes.

Node size represents Evidence for each gene, node color represents Log2_FC, red color implies over-expressed and blue color implies under-expressed. Green circled nodes are seed genes (T2D-associated genes curated from OMIM). Edge color represents Confidence (h-Score) for each interaction. Note: this figure was generated by Dr. Wu and used here with his permission. For details, please refer to (*51*).

Top-20 significant pathways

Totally 51 significant pathways (*p**-norm >= 1.2) in IBB from GSE24215 are obtained from all the differential pathways, measured by using pathway differential expressions (*p**-norm). Top-20 significant pathways in IBB from GSE24215, ordered by pathway differential expressions (*p**-norm), are listed in Table 4-3.

Table 4-3. Top-20 significant pathways in IBB from GSE24215

They are ordered by pathway differential expressions (p*-norm), which is measured with all the available differential gene expressions in IBB from GSE24215.

PATHWAY_NAME	DB_SOURCE_ID	NORM_ABS_FC	MAX_ABS_FC
IL-9 Pathway	KEGG	**2.56782**	3.97011
IL-10 Pathway	KEGG	**2.35238**	3.97011
IL23-mediated signaling events	NCI-Nature Curated	**2.29305**	3.97011
EPO signaling pathway	NCI-Nature Curated	**2.24041**	3.97011
Murine MSP-STK Signaling	KEGG	**2.20505**	3.28219
IL6-mediated signaling events	NCI-Nature Curated	**2.19750**	3.97011
Type II diabetes mellitus	KEGG	**2.16107**	3.97011
Signaling events mediated by PTP1B	NCI-Nature Curated	**2.13761**	3.97011
growth hormone signaling pathway	BioCarta	**2.09285**	3.31174
IL-4 Pathway	KEGG	**2.08410**	3.97011
Growth Hormone Signaling	KEGG	**2.06889**	3.97011
LDL Oxidation in Atherogenesis	KEGG	**2.05923**	3.28219
IL4-mediated signaling events	NCI-Nature Curated	**2.02714**	3.97011

Adipocytokine signaling pathway	KEGG	**1.97685**	3.97011
FoxO family signaling	NCI-Nature Curated	**1.97503**	3.39008
C. pneumoniae Infection in Atherosclerosis	KEGG	**1.89211**	3.28219
Calcineurin-regulated NFAT-dependent transcription in lymphocytes	NCI-Nature Curated	**1.87618**	3.29125
Jak-STAT signaling pathway	KEGG	**1.85834**	3.97011
MSP-RON Signaling	KEGG	**1.84756**	3.28219
Insulin signaling pathway	KEGG	**1.80963**	3.97011

ii. Findings on insulin after-bed (IAB) group

Top-20 differential genes

Totally 930 differential genes are obtained, which are differentially-expressed After Bed (IAB) Group from the microarray dataset – GSE24215. Differential genes are obtained by using filters with *p*-Value <= 0.05, Fold Change (FC) >= 1.3, and Average Expression Level (AEL) >= 40% after applying Limma package in Bioconductor. Average expression levels (AEL>=40%) have been checked to ensure the presences of the differential genes in the tissue - muscle. Duplicated genes with lower fold changes are eliminated, which implies that only the highest fold change for one gene will be kept.

Top-20 differential genes in IAB from GSE24215, ordered by absolute fold change (ABS_FC), are listed in Table 4-4.

Table 4-4. Top-20 differential genes in IAB from GSE24215

They are ordered by FC, (FC>=1.3, p-value <=0.05 and AEL >=40% after applying Limma package in Bioconductor). Note: Gene names are linked to GeneCards.org, UniProt IDs are linked to UniProt.org, and Evidences are linked to PubMed.

Gene Symbol	*p*-Value	FDR	Log2_FC	ABS_FC	Evidences
NR4A3	0.00000	0.00000	4.18431	**18.18032**	2
SOCS3	0.00005	0.00780	4.12982	**17.50651**	28
GADD45B	0.00054	0.03440	3.56927	**11.87016**	1
THBD	0.00000	0.00100	3.48248	**11.17714**	0
ADAMTS4	0.00019	0.01838	3.40269	**10.57580**	1
PDE4B	0.00000	0.00005	3.33522	**10.09258**	0
FOS	0.00031	0.02436	3.31416	**9.94630**	18
EGR1	0.00002	0.00362	3.11271	**8.65008**	3
JUNB	0.00004	0.00743	3.08829	**8.50488**	2
RGS16	0.00044	0.03038	2.96393	**7.80246**	1
ZFP36	0.00012	0.01425	2.92543	**7.59700**	2
MYC	0.00026	0.02179	2.86543	**7.28754**	23

CISH	0.00000	0.00049	2.78449	**6.88992**	0
CCL2	0.00013	0.01462	2.59339	**6.03513**	111
CXCL2	0.00006	0.00858	2.35828	**5.12758**	2
ATF3	0.00202	0.07254	2.31257	**4.96766**	9
SERPINA3	0.00732	0.14867	2.16589	**4.48742**	0
NFIL3	0.00002	0.00434	2.15060	**4.44013**	0
GADD45A	0.00382	0.10196	2.14953	**4.43682**	0
IL6	2.0786	0.00044	0.03051	**4.22398**	52

Top-20 significant genes

Totally 237 significant genes in IAB from GSE24215 are obtained from all the differential genes in a T2D-specific protein-protein interaction (PPI) network, measured by using significant score (considering both differential expressions and network properties). The T2D-specific PPI network is reconstructed by expanding all the seed genes curated from OMIM (PubMed >=0), in HAPPI_1.31 (3-Star) (Confidence: h-Score >=0.45). Top-20 significant genes in IAB from GSE24215, ordered by significant score (Sig_Score), are listed in Table 4-5.

Table 4-5. Top-20 significant genes in IAB from GSE24215.
They are ordered by Sig_Score, which is measured in the T2D-specific PPI network (PubMed >=0, h-Score >=0.45) for all the differential genes (FC>=1.3, p-value <=0.05 and AEL >=40% after applying Limma package in Bioconductor) in IAB from GSE24215. Note: Gene names are linked to GeneCards.org, UniProt IDs are linked to UniProt.org, and Evidences are linked to PubMed.

Gene	*p*-Value	FDR	Log2_FC	ABS_FC	Weight1	Weight2	Score	Evidence
CCL2	0.00013	0.01462	2.59339	6.03513	98	75.2645	**34.9751**	111
IL6	0.00044	0.03051	2.0786	4.22398	140	112.808	**34.68919**	52
IRS1	0.00902	0.16494	-0.68912	1.6123	103	83.1045	**23.58864**	280
IL6R	0.00002	0.00404	0.79678	1.73722	70	55.9728	**22.14359**	7
VEGFA	0.00017	0.01724	0.92831	1.90304	57	44.6818	**21.6589**	28
APP	0.03329	0.31458	-0.43503	1.35194	80	68.3971	**21.01141**	15
SOCS3	0.00005	0.0078	4.12982	17.50651	7	6.3574	**20.52815**	28
ADRB2	0.00016	0.0164	1.0105	2.01461	37	31.2254	**20.09311**	10
FOXO1	0.00008	0.01056	0.52622	1.44015	65	44.529	**19.42493**	59
MYC	0.00026	0.02179	2.86543	7.28754	10	8.8236	**19.33394**	23
SOD2	0.00164	0.06478	0.82246	1.76841	50	29.5486	**18.85629**	4

DGKD	0.02661	0.28238	0.5992	1.51487	42	34.932	**18.59775**	2
FOS	0.00031	0.02436	3.31416	9.9463	7	6.3552	**18.17698**	18
PIK3R1	0	0.00008	1.67218	3.18695	13	11.8228	**17.1966**	9
XBP1	0.00338	0.09606	0.40352	1.32273	42	26.9454	**16.35234**	9
AGT	0.00295	0.08852	0.59516	1.51064	24	18.032	**15.28071**	42
UCP3	0.00163	0.06449	-0.75312	1.68543	22	15.2626	**15.10061**	45
UCP2	0.00186	0.06897	-0.60181	1.51762	22	15.7096	**14.63273**	78
PPARGC1A	0.00479	0.117	0.53018	1.44411	17	13.6	**13.65437**	111
GFPT2	0.04473	0.36274	0.54879	1.46286	14	12.2852	**13.2432**	2

A T2D-significant protein-protein interaction (PPI) network (See Figure 4-2) is reconstructed a by connecting Top-20 significant genes in IAB from GSE24215, with and within the T2D-associated genes (seed genes) curated from OMIM (PubMed >= 50), in HAPPI_1.31 (3-Star) (Confidence: h-Score >=0.75)

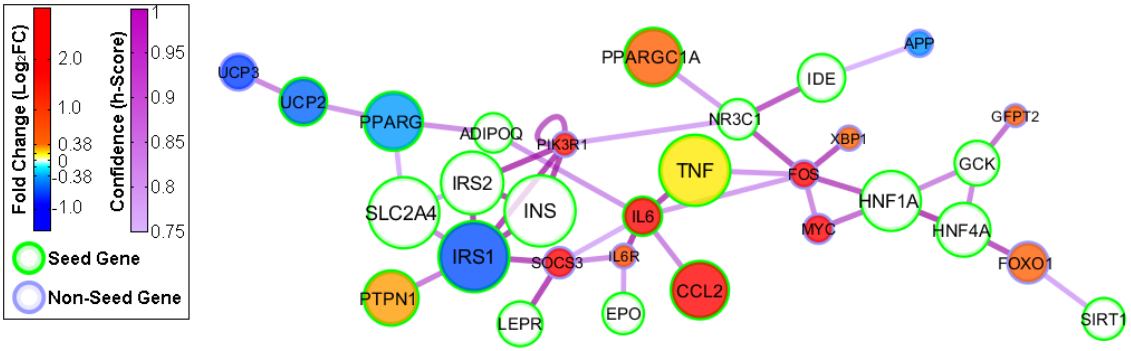

Figure 4-2. Top-20 significant genes in IAB from GSE24215, interacted with T2D-associated genes.
Node size represents Evidence for each gene, node color represents Log2_FC, red color implies over-expressed and blue color implies under-expressed. Green circled nodes are seed genes (T2D-associated genes curated from OMIM). Edge color represents Confidence (h-Score) for each interaction.

Top-20 significant pathways

Totally 64 significant pathways (p*-norm >= 1.2) in IAB from GSE24215 are obtained from all the differential pathways, measured by using pathway differential expressions (p*-norm). Top-20 significant pathways in IAB from GSE24215, ordered by pathway differential expressions (p*-norm), are listed in Table 4-6.

Table 4-6. Top-20 significant pathways in IAB from GSE24215.
They are ordered by pathway differential expressions (p*-norm), which is measured with all the available differential gene expressions in IAB from GSE24215.

PATHWAY_NAME	DB_SOURCE_ID	NORM_ABS_FC	MAX_ABS_FC
IL-9 Pathway	KEGG	6.70137	10.92515
IL-10 Pathway	KEGG	6.43693	10.92515
IL6-mediated signaling events	NCI-Nature Curated	6.34297	10.92515
EPO signaling pathway	NCI-Nature Curated	6.13163	10.92515
IL23-mediated signaling events	NCI-Nature Curated	6.01717	10.92515
igf-1 signaling pathway	BioCarta	5.98831	9.94630
Type II diabetes mellitus	KEGG	5.90782	10.92515
Signaling events mediated by PTP1B	NCI-Nature Curated	5.83706	10.92515
IL-4 Pathway	KEGG	5.71129	10.92515
signal transduction through il1r	BioCarta	5.68010	9.94630
Growth Hormone Signaling	KEGG	5.67342	10.92515
Calcineurin-regulated NFAT-dependent transcription in lymphocytes	NCI-Nature Curated	5.66004	9.94630
IL4-mediated signaling events	NCI-Nature Curated	5.53802	10.92515
FOXM1 transcription factor network	NCI-Nature Curated	5.44978	9.94630
Adipocytokine signaling pathway	KEGG	5.40776	10.92515
GDNF-Family Ligands and Receptor Interactions	KEGG	5.18210	9.94630
HIF-1-alpha transcription factor network	NCI-Nature Curated	4.99991	9.94630
Regulation of nuclear SMAD2/3 signaling	NCI-Nature Curated	4.91493	9.94630
Insulin signaling pathway	KEGG	4.89924	10.92515
Jak-STAT signaling pathway	KEGG	4.80603	10.92515

4.4 Conclusions

In this work, we apply both classical microarray analysis (such as differential analysis in Bioconductor) and the knowledge-guided analysis (network expansion analysis and pathway enrichment analysis). From the evidence from literature (PubMed), Top 20 significant genes from our analysis have more supports than Top 20 differential genes from simple differential analysis, in the case study on the microarray dataset - GSE24215. This implies the vitality of our hypothesis on which organized knowledge will help microarray data analysis in significant ways.

For GSE24215 dataset, both of the two networks (shown in Figure 1 and Figure 2) consist of two subnetworks. The bigger one includes genes that are highly related to

diabetic type 2. Some of genes are shared between before bed network and after bed network, like insulin receptor, peroxisome proliferator-activated receptor gamma and so on. For those shared genes, their expressions are different between these two conditions. We can see from the figure that IRS1, PPARG are under-expressed in after bed condition while IRS2 are under-expressed in before bed condition. Beside those shared molecules, some only show in after bed condition, like PPARGC1A, IDE, IL6R, APP, and PTPN1 while others only show in before bed condition, like CD36, SCARB1, SELP, ICAM1, TNFRSF1A, HSD11B1, and AKT2. The smaller one is relatively small sub-network. Commonly shared gene between before bed and after bed are HNF1A, HNF4A and MYC with similar expression level. Some genes like TCF7L2 only show up in before bed network while GCK, GFPT2, FOXO1, and SIRT1 only show in after bed network.

Another interesting finding is that the molecules which connect the red sub-network and blue-subnetwork are different. In before bed network, SMAD3 play this important role while in after bed network it is FOS that connect these two subnetworks. In fact, FOS and SMAD3 are physically interacting with each other and together Smad3 cooperates with c-Jun/c-Fos to mediate TGF-beta-induced transcription. Finally though IGF1 doesn't show up in the network, yet the IGF1 pathway is highly ranked (refer to pathway analysis part) in the after bed condition.

The key finding on GSE24215 was that bed rest was associated with a paradoxically increased response to insulin of genes involved in acute-phase response and inflammation, including IL-6 signaling, IL-10 signaling, and the ER stress pathway, contrasting the development of severe peripheral insulin resistance of glucose metabolism in young healthy men. The present study demonstrated that 9 days of bed rest induces severe transcriptional changes of genes potentially involved in the pathogenesis of insulin resistance and T2D in skeletal muscle, which might to some extent explain the harmful effect of a sedentary lifestyle on human metabolism. Impaired expression of HK2, VEGFA, NDUFB6, PPARGC1A, and OXPHOS genes in general, as well as a markedly increased expression of RRAD, are among the prime candidates contributing to the development of insulin resistance during bed rest.

Our analysis on this microarray dataset also shows that Insulin-stimulation After Bed-rest (IAB) is associated with the same significant genes: VEGFA (Rank: 5), PPARGC1A (Rank: 19), HK2 (Rank: 23), and RRAD (Rank: 29). We also found IAB is associated the same/similar pathways: IL-10 Pathway (Rank: 2) from KEGG database, IL6-mediated signaling events (Rank: 3) from NCI-Nature Curated pathway database, igf-1 signaling pathway (Rank: 6)from BioCarta database, Type II diabetes mellitus (Rank: 7) from KEGG database, Growth Hormone Signaling (Rank: 11) from KEGG database, Insulin signaling pathway (Rank: 19) from KEGG database, Jak-STAT signaling pathway (Rank: 20) from KEGG database, il 6 signaling pathway (Rank: 27) from BioCarta database, and role of erbb2 in signal transduction and oncology (Rank 31) from BioCarta database.

Chapter 5. Drug Repositioning using Literature Mining: Computational Connectivity Map

This section is based on the published work at (*52*). JYC conceived this work, guided the research team by providing ideas and feedback along the way, and revised the manuscript. HH involved the drug efficacy evaluation hypothesis development, carried out annotation web page development, designed the case studies and wrote the manuscript. XW participated in the drug efficacy hypothesis initiation, directionality concept development, case studies and manuscript writing. RP updated the PubMed database and tables in the Apex application. JL developed the previous version of C2Maps and performed case studies about the web server in her paper. GZ helped with annotation web page development.SI performed the directionality curations for breast cancer and colorectal cancer.

5.1 Introduction

Screening millions of chemical compounds to identify "hit" compounds for specific disease gene/protein targets has been a mainstream paradigm for modern drug discovery(*99*). While the conventional "One disease, One gene, and One drug" paradigm (*9*)works effectively for simple genetic disorders, it fails to produce effective drugs for complex diseases such as cancer (*10*). In complex diseases, many genes may be contributing to the disease's phenotype; therefore, identifying a "magic bullet" drug compound can be quite elusive.

Polypharmacology, which focuses on multi-target drugs, has become a new paradigm in drug discovery. Polypharmacology drugs have conventionally been viewed to have undesirable 'promiscuity'. However, recent research studies show, in the case of both older psychiatric drugs and modern anticancer therapies, that this promiscuity is intrinsic to the drug's therapeutic efficacy(*11*). Although there are over 40 drug-target (protein-compound interaction) databases according to Pathguide(*100*), (e.g. DrugBank(*101*), STITCH(*102*), CTT (*103*), CTD (*104*)and BindingDB(*105*), etal), a disease-specific searching platform is still needed to fully understand drug effects on the human body.

A new cancer systems biology approach to drug discovery has emerged in recent years. The primary focus of this paradigm is to understand the actions of drugs by considering targets in the context of the biological networks. By focusing on a systems level, it provides a better way to examine complicated diseases that can be caused by several gene mutations, such as cancer (*106*). However, most methods published so far focus on modeling the structure of the drug target network qualitatively (*107*). To examine a drug's effect on a molecular network representative of the disease, more quantitative and accurate modeling techniques need to be developed by utilizing the concept of network pharmacology (*106*) or network medicine (*13*).

In post-genome biology, molecular connectivity maps have been proposed to establish comprehensive knowledge links between molecules of interest in a given biological context (*108*).Molecular connectivity maps between drugs and genes/proteins in a disease-specific context can be particularly valuable because they allow researchers to evaluate drugs against each other using their unique gene/protein-drug association profiles. The functional approach to drug comparisons helps researchers gain global perspectives on both the toxicological profiles and therapeutic profiles of candidate drugs. Furthermore, the time it takes to develop high quality drugs in new therapeutic areas can also be reduced by using this method.

One approach for developing molecular connectivity map data is to generate disease-specific protein-drug association profiles computationally by mining bio molecular interaction networks and PubMed literature (*24*) . The Computational Connectivity Maps (C^2Maps) web server (*109*) is an online bioinformatics resource that provides biologists with potential relationships between drugs and genes/proteins in specific disease contexts based on network mining and literature mining. It's based on the concept of network pharmacology by examining many drugs at the same time and studying the drug disease relationship based on the underlying protein interaction network instead of drugs' direct target. C^2Maps provides quantitative measurements of protein's and drug's relevance to a specific disease by applying networking mining and the statistical testing methods in text mining and thus offers new insight to assess overall drug efficacy and toxicity.

Occurrences between proteins and drugs from literature mining of C^2Maps don't necessarily tell research what type of relationships they have, therapeutic or toxic. To overcome these limitations, we further standardize the classifications between proteins and drugs and then perform literature curations to determine drugs' effect on proteins on higher resolutions. Such valuable information is not readily available from the existing drug-target (protein-compound interaction) databases (e.g. DrugBank, STITCH, CTT, *et al.*) they may be scattered within a description or referenced text.

To assess drug pharmacological effect, such as drug efficacy and toxicity, we assume that "ideal" drugs for a patient diagnosed with a certain disease should modulate the gene expression profiles of this patient to the similar level with those in normal healthy people. Therefore, for those statistically over-expressed genes, drugs should be able to inhibit their expression level to the normal range. Similarly, for those statistically under-expressed genes, drugs should be able to activate their expression level to the normal range. In this way, drugs can treat or prevent the disease through reversing the gene expression level from disease status to the normal range, thus modulating cellular function as in normal cells.

By assuming that if the gene expression profiles of disease and drug are opposing, then the drug might be a potential treatment option of the disease, (*108*) identified novel drug indications in diet-induced obesity or Alzheimer's disease. Another work by Atul(*17*) utilized the same gene expression data and algorithms with large scale gene expression data from GEO to study associations between 100 diseases and 164 drug molecules. They found candidate therapeutics for 53 of the diseases. These studies are proof of principle that how using public genomics database and similar hypothesis can benefit drug discovery. Though gene expression data are publicly available for more than 1000 compounds in the second release of (*108*), yet there are numerous compounds that are not part of the database. Another limitation of this overly simplified hypothesis lies in it doesn't differentiate important genes from unimportant ones. Ideally a biological meaningful scoring methods needs developed.

Drugs effect data from literatures could be complementary here. In this work, we focus on building comprehensive disease-gene-drug connectivity relationships with drug-protein directionality (inhibit/activate) information based on the C^2Maps platform (*109*). To show the feasibility of applying the data for computational drug discovery, we took previous hypothesis a step forward by but assigning different weights to different genes. However this work aims to provide the data for the future network pharmacology research instead of developing a drug efficacy prediction method .This work has the following contributions:

1) The C^2Maps website itself has been not published before though (*109*) only provides the underlying computational methodology and relative low disease coverage such as Alzheimer's Disease.

2) We create an interactive interface for directionality annotation of drug-protein pairs with literature evidences from PubMed.

3) We curate the directionality information of drug protein pairs for three disease phenotypes: breast cancer, colorectal cancer and Alzheimer disease from 5133, 4869 and 3928 PubMed abstracts, respectively. We also upload these curated directionality information into the C^2Maps, and perform a statistical analysis on them. Curation of additional diseases, like pancreatic cancer and autism, is still on-going.

4) We enhance the functionality of disease-specific searching for relevant proteins and drugs with directionality information.

5) We update the comprehensive disease-gene-drug connectivity data in the C^2Maps databases, including 19,569,563 PubMed abstracts in the current version and 142,523 unique 3 star protein interactions in the current version.

6) We also use breast cancer as a case study to demonstrate the functionality of disease-specific searching for relevant drug-protein pairs with directionality information.

7) Based on the searching result, we show the feasibility of performing drug pharmacological effect evaluation for two important breast cancer drugs to show the power of updated C^2Maps in drug efficacy and toxicity assessment.

5.2 Methods

i. Data sources and systems design

As shown in Figure 5-1, the C^2Maps platform incorporates three major components in its systems design:

- *Network mining* component takes a query disease term as the input, and generates a ranked list of disease-relevant proteins as the output, through 1) MeSH term matching, 2) disease-associated gene searching from OMIM (*110*),3) network expanding in HAPPI (*80*), and 4) network-based protein ranking;
- *Text mining* component takes an input list of genes or proteins, and creates a list of enriched disease candidate drugs that are significantly associated with the disease-relevant proteins from the previous component as the output, through 1) gene/protein name mapping using UniProtKB, 2) article abstract retrieving from PubMed, 3) drug/chemical compound identification using MeSH term, and 4) disease-specific drug-protein pair ranking;
- *Drug effect annotating* component can allow users to 1) retrieve disease-specific drug-protein association list, 2) curate drug-protein directionality information from PubMed abstract, 3) annotate these drug-protein directionality information interactively, and 4) browse disease-specific drug-protein directionality information online.

In specific, we apply the network mining method originally developed by Chen et al.(*60*) to fish cancer relevant proteins from the protein interaction network. We expand cancer related genes/proteins using PPIs recorded in the Human Annotated and Predicted Protein Interactions (HAPPI) database to construct cancer-specific PPI sub-network. A protein's cancer relevance score r_p is calculated as the function (1).

$$r_p = k \ln(\sum_{q \in NET} conf(p,g)) - \ln(\sum_{q \in NET} N(p,q)) \quad (1)$$

Here, p and q are indices for proteins in the cancer-related interaction network *PPI*, k is an empirical constant (k=2 in this study), *conf(p, q)* is the confidence score assigned to each interaction between protein p and q, and *N(p, q)* holds the value of 1 if the protein p interacts with q.

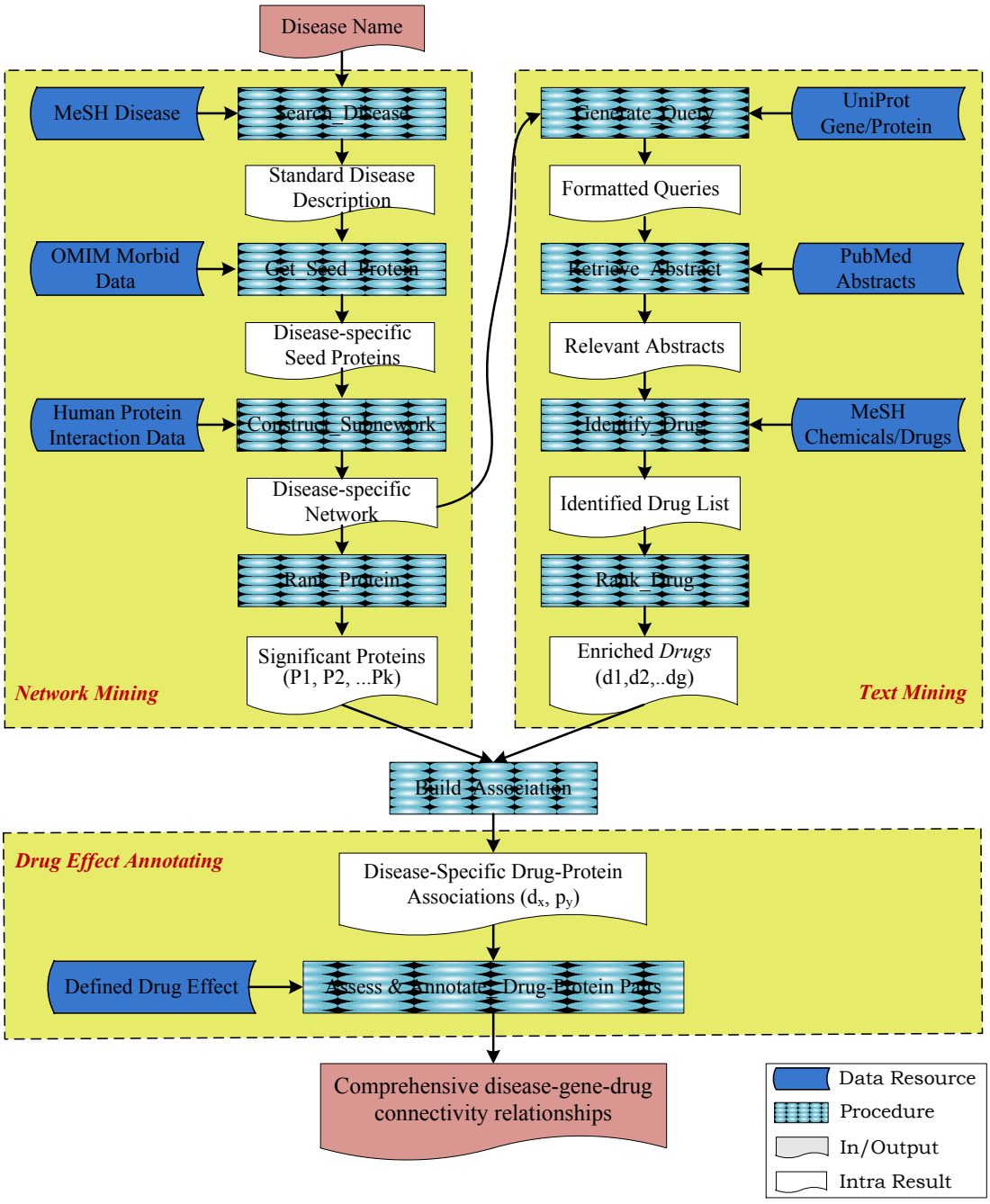

Figure 5-1. C2Map workflow for a given disease-specific study.

To retrieve the important drugs for cancers and to parse out drug terms in the articles, we acquire PubMed abstracts with the list of cancer-related genes/proteins derived earlier from PPI as queries. Drug term frequency is calculated and compared to term statistical distributions from the PubMed abstracts to get the p-value of drugs using function (2).

$$\Delta_j = \frac{(\overline{df(d_j \mid T'_{NET})} - \overline{df(d_j \mid T'_{Random})})}{\sqrt{\frac{Var(d_j \mid T'_{NET})}{N_{NET}} + \frac{Var(d_j \mid T'_{Random})}{N_{Random}}}} \quad (2)$$

Here, $T'_{NET} = \{T'_{NET1}, T'_{NET2}, \ldots\}$ is generated by sampling the entire collection of retrieved abstracts T'_{NET}. $N_{NET} = |T'_{NET}|$ is the size of each sample. $T'_{Random} = \{T'_{Random1}, T'_{Random2} \ldots\}$ refers to a random sample generated by randomly sampling the entire number of PubMed abstracts; the size of the random sample is N_{Random}. $df(d_j \mid T'_{NET})$ and $df(d_j \mid T'_{Random})$ refer to average document frequencies of d_j in T'_{NET} and T'_{Random}. $Var(d_j \mid T'_{NET})$ and $Var(d_j \mid T'_{Random})$ refer to document frequency variances of d_j in T'_{NET} and in T'_{Random}. A two-sided tails t-test was then performed to calculate the p-value. A thorough description of the computational components and algorithms used, along with data sets and data processing parameters, is described in detail by Li et al. (*109*).

The C^{2}Maps platform follows a multi-tier architecture design. The back end was implemented as PL/SQL packages in the Oracle 11g database server, with the Oracle Text engine enabled, to ensure scalable querying of PubMed text documents. The C^{2}Maps application middleware was implemented in the Oracle Application Express (APEX) server, which bridged between the Apache web server and the Oracle database server.

The current release of C^{2}Maps uses the following data sets: 19,569,563 records in the PubMed/MEDLINE baseline database (*111*), 142,523 human protein-protein interactions above 3-star confidence ratings in the HAPPI database(*80*), 26,142 descriptors in the MeSH database (Category C for diseases and Category D for chemicals and drugs) (*112*), 20,331 entries for the curated human proteins in the UniProtKB database (*113*), 18,344 entities in the OMIM database (*61*), and 4,772 entities in the DrugBank. Current statistics for the included database records is also shown in Table 5-1. The top 500 drug-protein

pairs for 'Alzheimer disease', 'Breast cancer' and 'Colorectal cancer' from C^{2}Maps were manually curated by assigning the effects of drugs on proteins as defined in the next section. As a result, C^{2}Maps platform contains 3928, 5133 and 4869 curated records for Alzheimer disease, Breast cancer and colorectal cancer respectively. All data is warehoused in a local Oracle 11g database.

Table 5-1. Current statistics for the included database records

Dataset	Data Resource	Record count
Biomedical Literature	PubMed	19,569,563
Human Protein-Protein Interaction	Unique HAPPI 3-star interactions	142,523
Disease and Drug Terminology	MeSH descriptors	26,142
Human Protein	UniProtKB	20,331
Disease-Gene relationships	OMIM	18,344
Drug Information	DrugBank	4,772

ii. Drug effect annotation

Since our hypothesis is that ideal drugs for a patient diagnosed with a certain disease should modulate the gene expression profile of this patient to the similar level with those in healthy people, we annotate a drug's pharmacological effect on a protein using one of the following three categories (also illustrated in Figure 5-2):

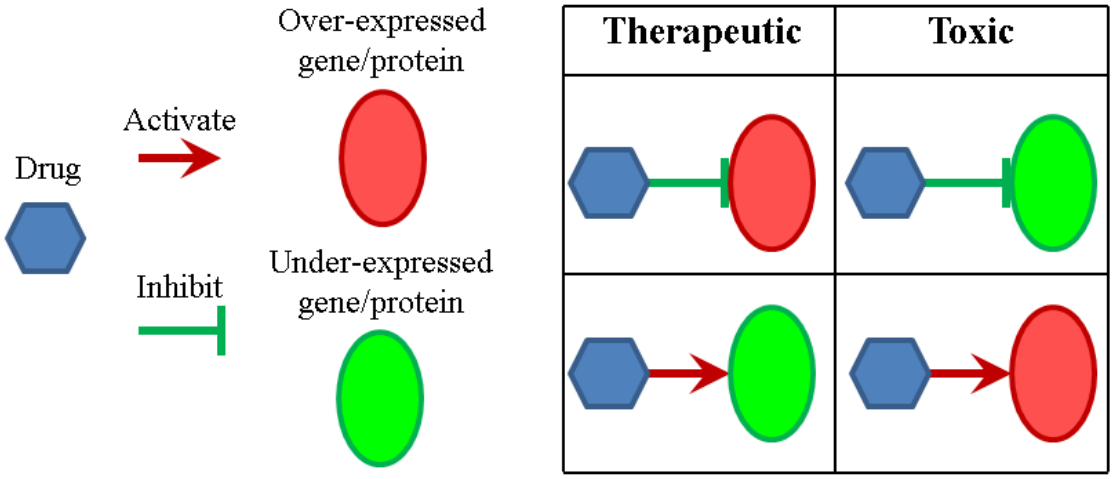

Figure 5-2 Illustration of drug pharmacological effects based on directionality information for drug-protein pairs

- *Therapeutic*: if the drug activates the under-expressed protein or inhibits the over-expressed protein, we define that the drug has a therapeutic effect on that protein
- *Toxic*: if the drug activates the over-expressed protein or inhibits the under-expressed protein, we define that the drug has a toxic effect on that protein
- *Ambiguous*: if there is missing directionality information for either the nodes (i.e. proteins/drugs) or edges.

iii. Perturbation Effects of Drugs on Proteins/Genes

We use breast cancer as an example to illustrate how we curate the directionality of a drug-protein pair retrieved from the C^2Maps platform. The followings are four categories of an annotated drug-protein relationship pair (also, refer to Table 5-2):

- *Activation* - "Subsequent injection of tamoxifen triggers the transient activation of Akt/PKB in mice." (Tamoxifen and AKT1_HUMAN, PMID: 12640620).
- *Inhibition* - "Treatment of cells with Cycloheximide (CHX) prevented the activation of p53 in all phases of the cell cycle and its accumulation in G1/S and S." (P53_HUMAN and Cycloheximide, PMID:9484835).
- *Indirect Yes* - "Hydroxyurea-mediated DNA synthesis arrest of S phase MCF7 cells led to a loss of BRCA1 from these structures." (BRCA1_HUMAN and Hydroxyurea, PMID:9267023).
- *Ambiguous* - "GRalpha and GRbeta transcripts are coordinately upregulated in CEM-C7 cells and coordinately downregulated in IM-9 cells by dexamethasone." (GCR_HUMAN and Dexamtheasone, PMID:12974663).

Table 5-2. Curation of drug-protein relations from Pub-Med abstracts

Relation	Protein	Drug	PMID	Relation Proof
Up-Regulated	BRCA1	Estradiol	7553629	"BRCA1 mRNA and protein levels were significantly decreased in estrogen-depleted MCF-7 and BT20T cells and **increased again** after stimulation with beta-estradiol".
Down-Regulated	P53	Cycloheximide	9484835	"Treatment of cells with cycloheximide (CHX) **prevented the activation** of p53 in all phases of the cell cycle and its accumulation in G1/S and S".
Indirect	BRCA1	Hydroxyurea	9267023	"Hydroxyurea-mediated DNA synthesis **arrest of S phase MCF7 cells led to a loss of BRCA1** from these structures".
Ambiguous	GCR	Dexamethasone	12974633	"GRalpha and GRbeta transcripts are

				coordinately **upregulated**in CEM-C7 cells and coordinately **downregulated** in IM-9 cells by dexamethasone".
Unknown	AKT1	Phosphothreonine	11087733	The drug-protein relation is not mentioned in the text.

We can see that the literature does contain such information and thus provides basis for the directionality information retrieval. We focus on curating drug actions in the disease context rather than only in cell lines like in (*108*). Different research works under specific contexts may produce conflicting conclusions regarding drug protein relationship. Take Tamoxifen and estrogen receptor as examples. As shown in Table 5-3, we successfully extracted 7 article abstracts which support the inhibitory effects of Tamoxifen on estrogen receptor and 2 PubMed abstracts which support the stimulatory effects of Tamoxifen on estrogen receptor. The pre-dominant evidence showing Tamoxifen's inhibition on estrogen receptor (*114*) matches well with the fact that Tamoxifen acts as an antagonist for estrogen receptor. Beside checking the majority vote of all the related papers, the original references was also checked. For Tamoxifen, it inhibits estrogen receptor in the mammary tissue while activating estrogen receptor in bone density. In the breast cancer case study, we decided Tamoxifen's inhibiting effect on estrogen receptor since the gene expression experiment was based on breast tissue. In the future, one could add additional contexts such as experimental conditions, disease subtypes and so on. In the current version, they are not added due to limited availability of those data in abstracts.

Table 5-3. PubMed evidence for Tamoxifen's effect on ESR1

Drug	Protein	PMID	Direction
Tamoxifen	ESR1_HUMAN	14507640	-1
Tamoxifen	ESR1_HUMAN	2359140	-1
Tamoxifen	ESR1_HUMAN	14759988	-1
Tamoxifen	ESR1_HUMAN	11774281	-1
Tamoxifen	ESR1_HUMAN	2137212	-1
Tamoxifen	ESR1_HUMAN	9328205	-1
Tamoxifen	ESR1_HUMAN	11261829	-1
Tamoxifen	ESR1_HUMAN	12767276	1
Tamoxifen	ESR1_HUMAN	11812086	1

1 represents that the drug up regulates the protein while -1 represents down regulation

iv. Data access and website usage

The C^2Maps online platform (http://bio.informatics.iupui.edu/cmaps) provides researchers a web-based bioinformatics user interface, following principles described in (*115*). As shown in Figure 5-3, users can begin with a single disease term as a query and navigate to extract significant subsets of the disease-specific C^2Maps.

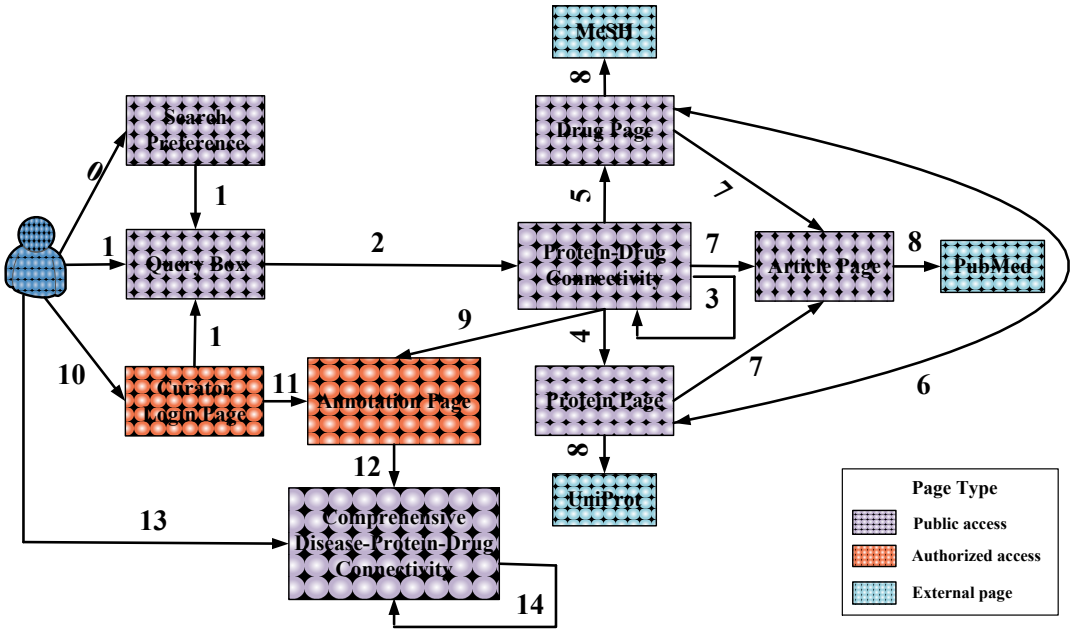

Figure 5-3. The navigational site map of the C2Map platform.
The numbers refer to: 0. Configure search preference; 1. Input interest disease; 2. Generate Protein-Drug connectivity; 3. Refine initially constructed connectivity map; 4. Link to protein page; 5. Link to drug page; 6. Link between protein and drug pages; 7. Link to evidence article pages; 8. Link to external data resources (MeSH, UniProt and PubMed); 9. Import enriched disease specific protein-drug associations for further annotation; 10; Authorized users (curators) set up profiles and login in; 11. Annotate effects of drugs on protein/genes; 12. Release annotation results; 13. Browse annotated disease-protein-drug connectivity relationships; 14. Filter or search for interested subset of connectivity relationships.

We also show snapshots of C^{2}Maps output web pages in Figure 5-4, using "Breast Cancer" as a query example. The statistically significant relationships between proteins and drugs are extracted from the literature are displayed in tabular format (Figure 5-4).

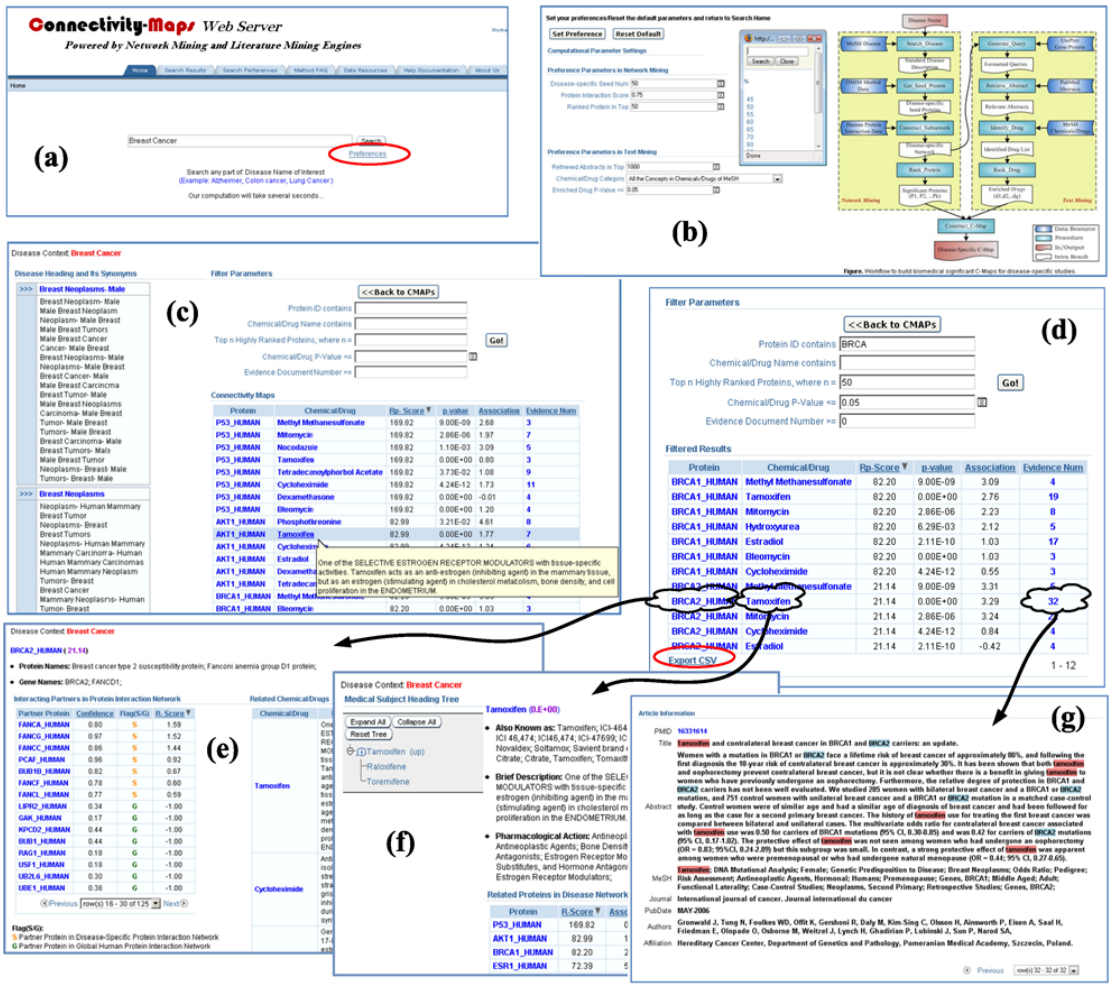

Figure 5-4. Web Interface for C2Maps basic query function.
(a) C2Maps default home page with a query box; (b) Preference configuration page with editable parameters in network mining component and text mining component; (c) Disease-specific protein-drug connectivity map page shows disease-relevant proteins, their ranking score, and drugs, their p-value, their relative frequency, as well as evidences supporting the protein-drug associations; (d) Filter result page shows the subset of the connectivity map; (e) Protein detail page shows the disease-related protein, its partner proteins in interaction network, and related drugs in retrieved abstracts; (f) Drug detail page shows the disease-related drugs, its neighbor drugs in Medical Subject Heading Tree, and related proteins in disease-specific protein interaction network; (g) Article detail page shows the literature references in PubMed with protein entity and drug entity highlighted. The url is http://bio.informatics.iupui.edu/cmaps

The proteins, drugs, and evidence numbers are further linked to protein, drug, and the article detail page, respectively. The search results can also be sorted by the protein ranking score (R-Score) and Chemical/Drug significance (*P-Value*). In addition, the page also lists Disease Context (disease name of user interest) and Disease Terminology (disease name containing query term in the controlled vocabulary of Medical Subject Headings). Users can also specify advanced search criteria for further biological/pharmacological analysis. Annotations for the extracted relations could be performed from the 'Annotate Data' tab.

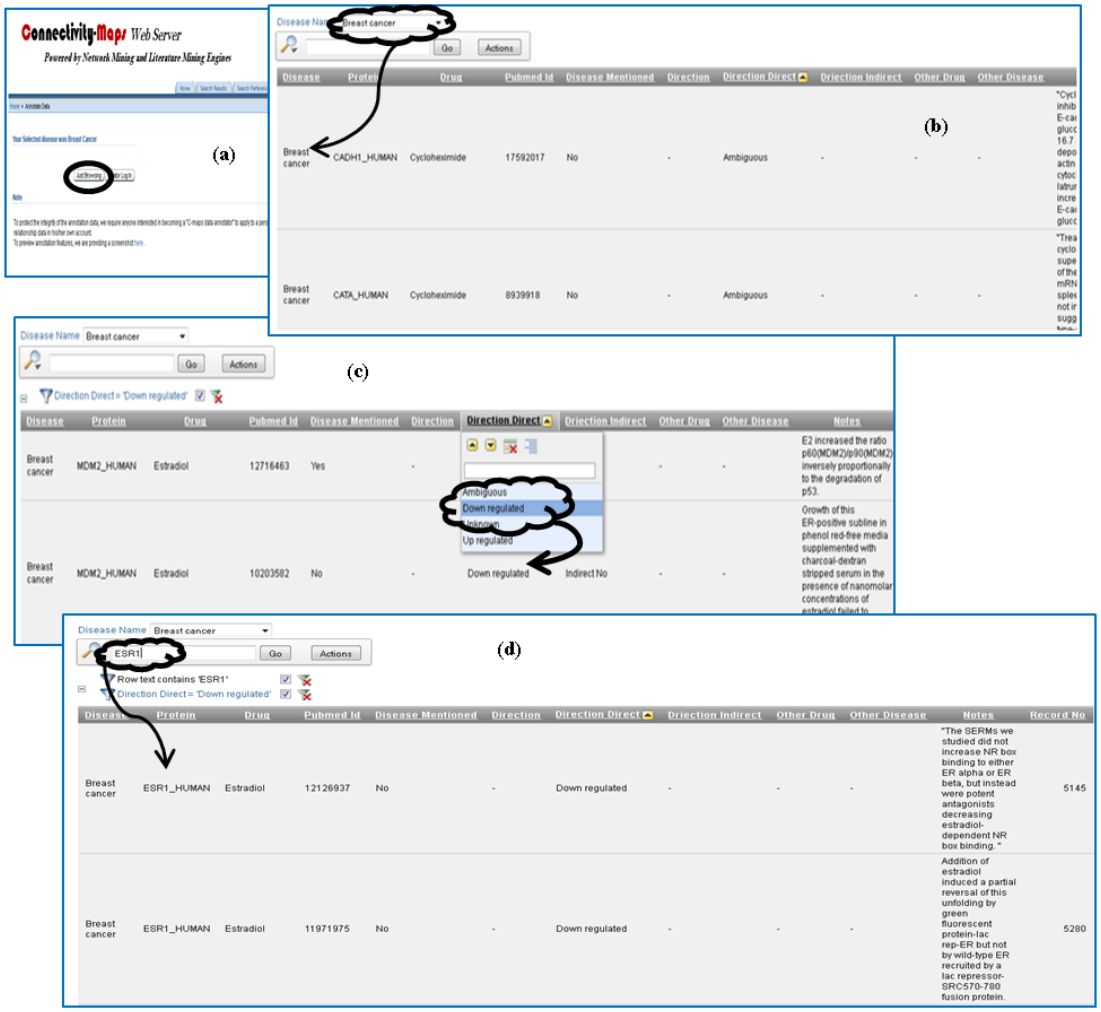

Figure 5-5. Web Interface for C2Map Annotation data browse.
(a) Annotation Data: main search page allowing either public browse or authorized curation; (b) Annotation Data Browse: displaying curated directionality between drug and proteins for certain disease; (c) Filtering: Each column support filtering (e.g., only show 'Down regulated' directionality); (d). Search function: search by either drug or protein. For example, search all records for ESR1 protein. The url is http://bio.informatics.iupui.edu/cmaps

v. Browsing Disease-specific drug-protein relationship information

Any public C^{2}Maps database user can access the well curated drug-protein directionality data. The database will display disease, protein, drug directionality, and PudMed evidence for each record. Each column can be sorted or filtered. One can also display only drug-protein directionality belonging to certain disease by selecting it from the drop down list (Figure 5-5b). Furthermore, the user can also search the keywords, such as

protein name or drug name, to retrieve only specific records. Currently, the C^2Maps platform contains 3928 curated records for Alzheimer disease, 5133 curated records for breast cancer, and 4869 curated records for colorectal cancer. More curation information will be updated regularly.

vi. Interactive interface for directionality annotation

An authorized C^2Maps database user can also annotate selected C^2Maps contents by performing manual curation from the 'Annotate Data' tab. The user may apply for an annotator's account to edit protein-drug interactions suggested by the C^2Maps automated recommendation system. This editing is provided through a separate user interface that enables the annotator to categorize protein-drug relationships as direct (including activation, inhibitory, ambiguous), indirect, or unknown (Figure 5-6f). The user will first select the assigned disease and the C^2Maps webserver will populate disease relevant protein and drug pairs. All the PudMed abstracts mentioning both the relevant protein and drug will be pulled out and the curator can read the abstract to annotate the directionality between the drug and the protein. The user can also edit (Figure 5-6d) each record or delete it.

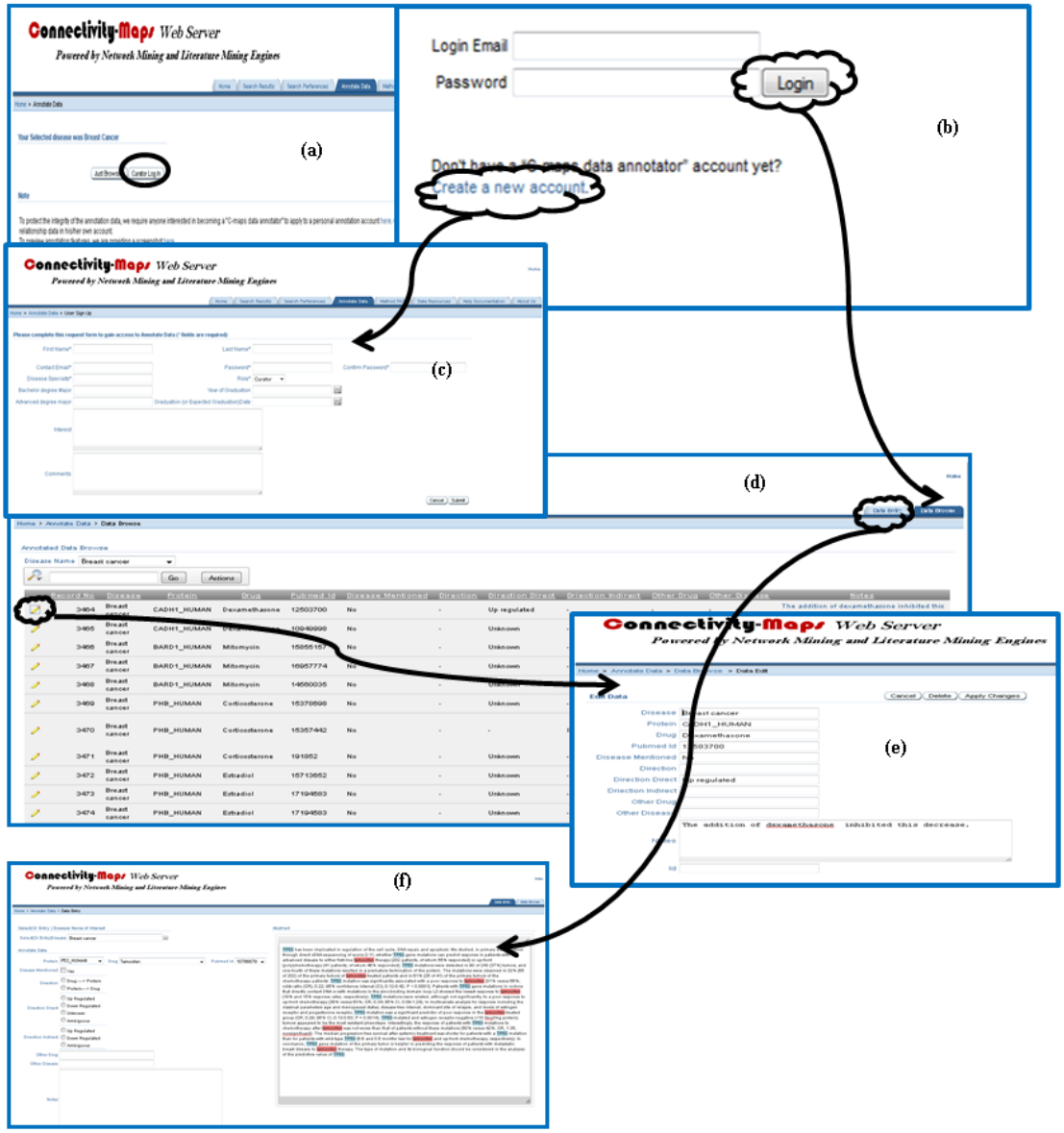

Figure 5-6. Web Interface for C2Map Annotation data curation.
(a) Annotation Data: main search page allowing either public browse or authorized curation; (b) Curator login: require login to curate directionality for certain disease; (c) User sign up: application form to create an account as a curator; (d) Data Browse: the browse page after login; (e) Data Edit: update or delete previously curated records; (f) Data Entry: 1. select disease; 2. relevant protein based on 1; 3. relevant drug based on 2; 4. relevant PubMed ID based on both 2 and 3; 5. relevant abstract based on PubMed ID from 4; 6. Curate directionality based on 5.

5.3 Results

i. Statistical analysis for reliability

The major computational components of the C^2Maps platform were developed using validated computational techniques. In the network mining component, protein interaction network expansion was able to reduce the initial biases and low data coverage, which may have existed in the seed list of protein. We used the new HAPPI database instead of other protein interaction databases because of its overall better data quality (comparable or better than data in the HPRD database for quality star grades of 3 and above) and coverage (more than 280,000 human protein interactions with star grades of 3 and above), which was thoroughly described in Chen et al.(*80*). In the text mining component, the PubMed abstract retrieval for each protein was shown to improve Information Retrieval (IR) recall performance without sacrificing precision The quality of disease drug identifications was shown to outperform comparable systems with balanced sensitivity, specificity, and positive predictive values (for details, refer to Li et al. (*109*)).

In Table 5-4, we show a summary of C^2Maps platform performance, by comparing its overall sensitivity, specificity, PPV (positive predictive value), F-score, and ACC (accuracy) measures among a number of cancers. The result confirmed that C^2Maps performed well consistently across different disease studies.

Table 5-4. Performance assessment of C2Maps in varying cancers.

	Bladder	Breast	Leukemia	Lung	Lymphoma	Melanoma	Ovary	Pancreas	Prostate
Sensitivity	80.84%	79.80%	83.16%	78.44%	81.20%	77.39%	80.88%	84.99%	82.84%
Specificity	87.11%	84.91%	86.11%	89.37%	87.60%	91.53%	84.34%	86.45%	88.38%
ACC	86.70%	84.27%	85.63%	87.86%	86.78%	90.17%	84.09%	86.34%	87.93%
PPV	30.51%	43.01%	53.39%	54.06%	48.92%	49.24%	28.84%	33.82%	38.88%
F-Score	44.30%	55.89%	65.03%	64.01%	61.05%	60.19%	42.52%	48.38%	52.92%

The experiment were performed using a protein interaction confidence minimum threshold of 3 star and above (i.e., reliability score of >0.75) and retrieved drug p-value at a minimum threshold of 0.05. The detailed evaluation procedures and measurement definitions, they can be found in the "Method FAQ" page of the C^2Maps website and as supplemental materials. The table was developed by Jiao Li and was used here by her permission (*52*).

ii. A case study on breast cancer specific searching for relevant drug-protein pairs with directionality information

We evaluated breast cancer drugs from C^2Maps based on our hypothesis. First, we obtained top 500 drug protein pairs for breast cancer from the C^2Maps web server, 23 drugs and 103 proteins, respectively. The r_p scores for those proteins range from 1.69 to 169.82 and the P-Values for those drugs are all below 0.05. Well known breast cancer related proteins, like BRCA1, or related drugs, like tamoxifen, were included in these 500 pairs. All supporting evidence for each drug protein relations, a total of 5,225 PubMed abstracts, was manually curated to extract the relevant drug effect information. Out of those 500 pairs, 155 pairs contained information of how the drug affects the protein in the literature, totaling 19 drugs and 52 proteins. After performing manual curation, 79 drug protein pairs contained only up-regulation information, 57 only down-regulation information, 11 primarily up-regulation information, and 8 primarily down-regulation information. The distribution of directionality categories for breast cancer is shown in Figure 5-7a. A subnetwork based on the directionality information specific for Tamoxifen can be constructed from C^2Maps directionality data (shown in Figure 5-7b). Another subnetwork based on the directionality information specific for Plicamycin is also shown in Figure 5-7e.

iii. A case study on drug efficacy evaluation with C^2Maps

Drug efficacy can be measured by the ability of a drug to produce the desired phenotypic effect or molecular effect. To evaluate the drug efficacy in the molecular level based on our hypothesis illustrated in Figure 5-2, we need know how drugs can affect the expression of it interacting genes and how those genes expressed in disease conditions. We have got the former from the above case study of C^2Maps. To get the latter, we did the differential analysis of a well-studied microarray dataset-GSE3191(*116*). This experiment contains breast cancer subtype luminal A, basal-like and also normal breast tissues. We obtained the differential genes for both breast cancer subtypes - luminal A and basal-like when compared to normal. We identified 579 differential genes between luminal A and normal, 773 differential genes between basal like and normal. We used these two sets for the following case study.

iv. Tamoxifen efficacy and toxicity assessment for the luminal A subtype

Tamoxifen is a standard drug clinically used for breast cancer and has 15 interacting proteins with directionality annotations from C^2Maps (shown in Table 5-5). We intersected differential genes from luminal A microarray experiment with Tamoxifen's interacting partners from C^2Maps. Four proteins out of 15 are differentially expressed between luminal A and normal, including ESR1. In Figure 5-7c, drugs are represented as hexagons and proteins as circles. The size of a protein node is proportional to the r_p score, an indication of importance of this protein to breast cancer. Red nodes stand for over-expressed proteins in breast cancer while green ones represent under-expressed proteins. For edges between drugs and proteins, red symbolizes that the drug activates the protein while green symbolizes inhibition. From the Figure, Tamoxifen has 3 therapeutic effects: it inhibits over-expressed ESR1, activates under-expressed JUN and activates MYC. Tamoxifen also has one toxic effect, activating over-expressed ERBB2, which might help explain certain side effects when using Tamoxifen. Considering that ESR1 is more significant for breast cancer compared with the other three proteins, overall, Tamoxifen has more of a therapeutic value in Luminal A patients by reversing the gene expression of important disease proteins in the network level (Figure 7c).

Table 5-5.Tamoxifen relevant proteins and their directionality

Drug	Protein	RpScore	Association	Direction
Tamoxifen	AKT1_HUMAN	82.99	1.77	1
Tamoxifen	BRCA2_HUMAN	21.14	3.29	1
Tamoxifen	CADH1_HUMAN	17.57	0.95	1
Tamoxifen	CDK2_HUMAN	2.94	2.12	1
Tamoxifen	E2F1_HUMAN	2.95	1.59	1
Tamoxifen	ERBB2_HUMAN	2.07	3.19	1
Tamoxifen	ESR1_HUMAN	72.39	5.11	-1
Tamoxifen	IRS1_HUMAN	2.51	1.29	-1
Tamoxifen	JUN_HUMAN	2.91	1.51	1
Tamoxifen	MYC_HUMAN	3.49	2.5	1
Tamoxifen	NCOA3_HUMAN	2.61	3.01	-1
Tamoxifen	NCOR1_HUMAN	2.81	4.05	1
Tamoxifen	P53_HUMAN	169.82	0.8	-1
Tamoxifen	P85A_HUMAN	2.92	1.57	1
Tamoxifen	PTEN_HUMAN	3.98	0.98	1
Plicamycin	MYC_HUMAN	3.49	4.16	-1
Plicamycin	SP1_HUMAN	3.32	6.24	-1

1 represents that the drug up regulates the protein while -1 represents down regulation. The association score was calculated based on the co-occurrence between the drug and protein. For details, please refer to (*109*).

v. Tamoxifen efficacy and toxicity assessment for the basal-like subtype

In Figure 5-7d, we portray the drug protein interaction for Tamoxifen in basal patients. Three proteins out of its 15 interacting proteins are differentially expressed between basal patients and normal. Tamoxifen has only 1 therapeutic effect by activating under-expressed JUN, while 2 toxic effects by activating over-expressed E2F1 and inhibiting under-expressed IRS1. However, all these three proteins are relatively not important for breast cancer. This implies a neutral role overall when using Tamoxifen in basal patients since it is not able to reverse its interacting proteins in basal condition (Figure 5-7d). This agrees well with the clinical fact that basal or triple negative breast cancer patients fail to benefit from Tamoxifen treatment.

vi. Plicamycinefficacy and toxicity assessment for the luminal A subtype

Plicamycin was an approved antineoplastic antibiotic for a variety of advanced forms of cancer. It has been withdrawn from market in 2000. In Figure 5-7f, we showed the drug protein interaction for Plicamycin in Luminal A patients. It has 2 interacting proteins with directionality annotations (shown in Table 5-5) and both are not significant in breast cancer with a low r_p score. Only 1 protein out of these 2 is differentially expressed between luminal A and normal. Plicamycin has a toxic effect overall by inhibiting under-expressed MYC. This implied a neutral or toxic effect when using Plicamycin in Luminal A subtype breast cancer patients since it is not able to reverse its interacting proteins in the disease condition (Figure 5-7f). This may help explain why it was withdrawn in 2000.

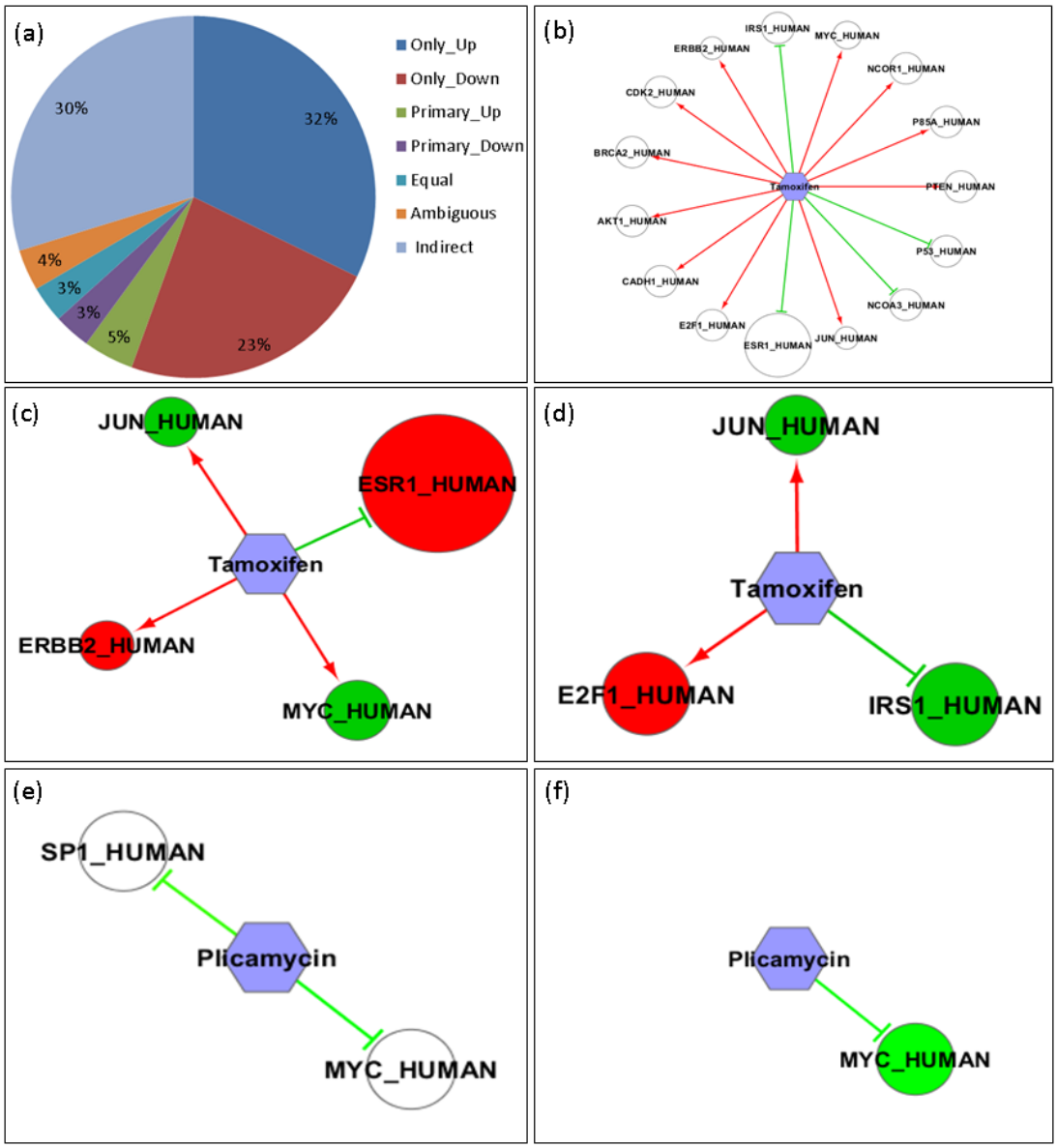

Figure 5-7.Breast cancer case study for drug pharmacological effectevaluation with C2Maps.

(a) Distribution of directionality categories for breast cancer; (b) A drug-target subnetwork with directionality information specific for Tamoxifen; (c) Drug effect evaluation for Tamoxifenon breast cancer subtype – luminal A; (d) Drug effect evaluation for Tamoxifenon breast cancer subtype – basal-like; (e) A drug-target subnetworkwith the directionality information specific for Plicamycin; (f) Drug effect evaluation for Plicamycinon breast cancer subtype – luminal A.

5.4 Conclusions

In this study, we present an upgraded C^2Maps platform to evaluate drug pharmacological effects based on the hypothesis that an ideal drug can reverse the gene expression level in a disease back to those in normal conditions. This online platform will enable users to query high-coverage protein-drug connectivity maps in real time. It enables users to research up-to-date knowledge of connectivity maps for a specific disease, explore therapeutic protein targets, design repurposed drug compounds, and assess toxicological impacts of drug compounds on disease-relevant genes/proteins. Three efficacy case studies prove the feasibility to apply the literature mined drug directionality data from C^2Maps for drug efficacy study. It will be a major resource to biomedical researchers interested in developing disease-specific therapeutic and diagnostic applications based on progresses in network biology and network pharmacology.

From the case study on breast cancer drug effect evaluation, we can see that there is still room for improvement, although the two breast cancer drugs get well-evaluated. The information for judging whether a drug have global therapeutic effects on other diseases is limited due to the manual curation procedure, while these information are very valuable for drug repurposing. In the current version of C^2Maps, drug-protein pairs are mainly come from literature mining based on disease-gene searching and network mining, which can be supplemented by plenty of publicly-available drug target databases. One could continue to update C^2Maps and improve its usability through achieving the following functionalities.

1) One could increase the functionality of drug-orientated searching for relevant disease phenotypes and proteins in the C^2Maps. It will allow users to input drug names, not just disease names. It should be able to retrieve all the disease names and genes/proteins related to this drug. This function will be very useful for drug repurposing.
2) One could increase the functionality of disease-orientated browsing for relevant proteins and drugs in the C^2Maps by using disease phenotype trees. It will allow users to browse the database by clicking the disease name. The current version

only supports the disease-specific searching function without any browsing function.

3) One could also enhance the functionality of interactive directionality information annotation for drug-protein pairs in the C^2Maps by using natural language processing (NLP) techniques. The literature curations for breast cancer, colorectal cancer and Alzheimer disease took three experts nearly one year's effort to complete. While it ensure the data quality, it's time consuming. With those golden standard dataset from curation, one could use NLP techniques to allow users to curate and annotate directionality information from PubMed abstracts more easily and semi-automatically.

Chapter 6. Drug Repositioning using Drug directionality Map (DMAP)

This section is based on my work at (*53*). JYC guided the research team by providing ideas and feedback along the way, and revised the manuscript. HH constructed the DMAP, performed drug repositioning with K-S algorithms and drug similarity network approach, and wrote the manuscript.

6.1 Introduction

Repositioning of drugs (*17, 22, 40*) already approved for human use could alleviate the cost (*41*) associated with early stages and offer a shorter path for new approval(*43*). Current computational methods for drug repositioning include: (i) studying the structural similarity of each drug to their targets' ligand set using chemoinformatic tools (*45*) or drug–drug and disease–disease similarity with machine learning methods(*22*), (ii) exploiting side-effect similarities (*46*), (iii) applying text-mining literature(*23*), or (iv) matching drug and disease gene expression profiles (*15, 17, 40, 47, 48*). Most of the approaches can only be applied to well characterized drugs whose targets or structures are known. Expression profile based approaches are, on the other hand, more general and do not require prior knowledge of the drugs.

Although the Connectivity Map (CMAP) approaches are gaining popularity for expression profile based drug repositioning, the limitation of these approaches is due to the coverage of the dataset. Lamb *et al.(15)* developed a public available database called CMAP containing a collection of transcriptional expression data from cell lines treated with small molecules. Iorio *et al.(40)* proposed drug repositioning by constructing drug-drug similarity networks from CMAP. Hu and Agarwal(*47*) and Sirota *et al.(17)* paired drugs and diseases whose gene expression patterns are negatively correlated. They further showed that the anti-correlation relationships between the drugs and diseases can suggest novel therapeutics for existing drugs. Despite of the success, the main limitation of studies based on CMAP (*15*) lie in the fact that it is simply impossible to screen all the drugs in the database due to experimental cost.

Critical to the success of the expression profile based drug repositioning is the resource for how drug affects the disease proteins. In this work we developed a computational Drug directionality Map (DMAP) database which contains ranked drugs' effects (i.e. activation or inhibition) on their interacting proteins (Figure 6-1). The database offers a better coverage consisting of directed drug-protein relationships for 328,676 drugs. To check its quality for drug repositioning, we applied the following two representative CMAP based drug-repositioning methods in literatures: (i) we calculated pairwise drug similarity (*40*) based on the DMAP for drug repositioning, (ii) we implemented the Kolmogorov–Smirnov algorithms (*15, 17*) based on the dataset from DMAP. We not only successfully recalled known drugs for breast cancer, colorectal cancer, lung cancer, diabetes, etc. but we were also able to propose novel indications for drugs in NCATS(*117*).

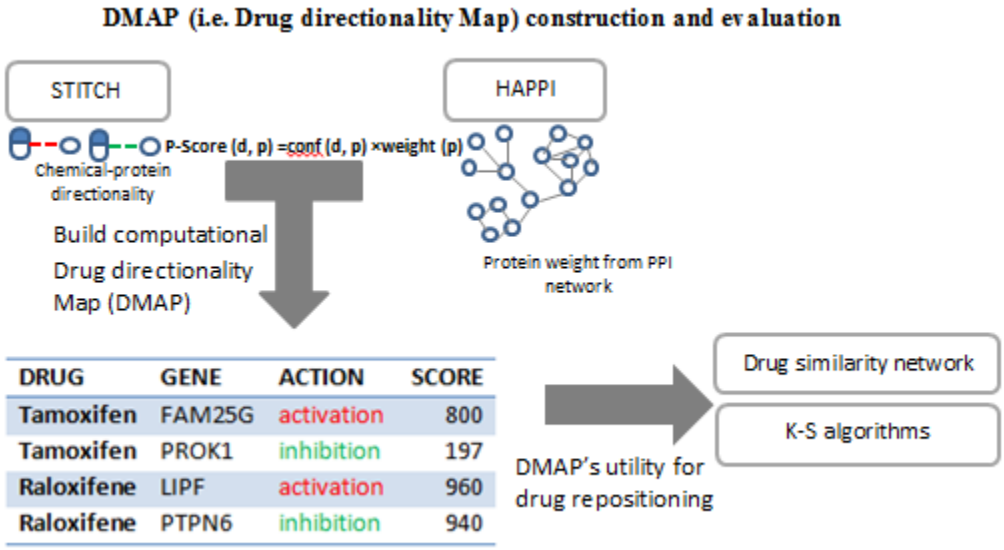

Figure 6-1. Computational framework.

6.2 Methods

i. Construct the DMAP data set

DMAP contains drug protein/gene directionality information in a compatible format with CMAP. The main data sources for DMAP are STITCH(*32*) and HAPPI (*80*). STITCH is an aggregated Cheminformatics database of interactions connecting over 300,000 chemicals and 2.6 million proteins. We first parsed out chemical protein interactions for Homo sapiens with those edge actions being activation or inhibition. Next we did a

probabilistic weighted average summary of all the evidence to come up with the overall action for each specific chemical-protein relationship.

To rank each relationship, we used HAPPI (*80*), an integrated protein interaction database including HPRD, BIND, MINT, STRING, and OPHID to assign a weight for each drug's interacting proteins (*60*). Finally, we developed an intuitive pharmacology score, or P-Score, to combine the probability for each interaction and the weight of the interacting protein:

$$P\text{-}Score(d,p)=conf(d,p)\times weight(p) \quad (1)$$

Here, *d* and *p* are specific drugs and proteins, respectively. *conf(d,p)* measures probability of each drug-protein relationship with a positive sign to indicate activation and a negative sign to indicate inhibition. It is thus within a range of [-1,1]. *weight(p)* is the measurement of the importance of the protein in the pathway as shown in the function (2)

$$Weight(p) = k\ln(\sum_{q\in NET} conf(p,g)) - \ln(\sum_{q\in NET} N(p,q)) \quad (2)$$

Here, p and q are proteins on the pathway, k is an empirical constant (k=2 in this study), conf(p, q) is the confidence score assigned by HAPPI to each interaction between protein p and q, and N(p, q) holds the value of 1 if the protein p interacts with q.

P-Score contains both the information of each drug's action on their interacting proteins and the importance of their proteins in the biological network. This is different than the expression level based ranking in CMAP, which may be more suitable for biomarker discovery instead of drug discovery. With P-Score for each drug-protein relationship, DMAP is thus in a compatible format with CMAP (*15*).

Compared to STITCH alone, DMAP differed in that i) it left out 'spurious' drug protein relationships by requiring the interacting protein to have biological significance as measured by the protein interactions in HAPPI; ii) it assigned P-Score to the drug protein relationship with Equation (1), different from the pure probabilistic score in STITCH database.

ii. Integrate drug therapeutic indication data

For repositioning existing drugs for other uses, we need have the approved indications for each drug. Thus we integrated the Therapeutic Target Database (TTD) (*118*) and the dataset from the PREDICT (*22*) paper to come up with a list of known drug indications. TTD is a database that provides information about drugs' known therapeutic protein targets and their targeted diseases. The PREDICT (*22*) paper provides a compiled list of drug indications. We integrated these two sources to get 2,912 drug indication associations corresponding to 1,180 drugs and 726 indications.

iii. Prepare disease expression signatures and drug expression signatures

The expression based drug repositioning need the disease gene expression as inputs. We retrieved the disease gene expression profiles from Pacini C et al. (*119*)'s paper. In total, 87 disease associated microarray experiments were compiled to represent 45 distinct diseases. According to Pacini C's paper, these datasets were obtained from the GEO microarray repository (*120*). The raw CEL files were normalized with RMA(*121*). For those gene expression profiles representing the same disease, they were combined with the median rank normalization by Warnat et al.(*122*).

The drug gene expression datasets were obtained from Iorio et al.(*40*)'s paper instead of directly from CMAP (*15*) to reduce the batch effect. Iorio et al.(*40*) computed a single synthetic ranked list of genes, called Prototype Ranked List (PRL), by merging all the ranked list of the same compound in CMAP. Only consistently overexpressed/underexpressed genes are placed at the top/bottom of the RPL. This helped capture a consensus transcriptional response for each drug. We thus chose to use the PRL to represent the drug signatures from CMAP in this study.

iv. Design drug similarity measurement

To measure the similarity among each drug pair, we computed $SIM(d_x,d_y)$ based on the Tanimoto Coefficient between their interacting proteins (3)

$$SIM(d_x,d_y)=\frac{|p_{x+}\cap p_{y+}|-|p_{x-}\cap p_{y-}|}{|p_x\cup p_y|} \quad (3)$$

Here, d_x and d_y represent the two specific drugs. p_x represents the set of interacting proteins for d_x. p_y represents the set of interacting proteins for d_y. $|p_x \cup p_y|$ is the number

of total distinct proteins in $p_x$and p_y. $|p_{x+} \cap p_{y+}|$ is the number of overlapped proteins on which both drugs have consistent effects (i.e. both activate or inhibit the shared proteins). $|p_{x-} \cap p_{y-}|$ is the number of overlapped proteins on which the drug pair have in-consistent effects (i.e. one activates while the other inhibits the shared proteins). *SIM(d_x,d_y)* lies in the range of [-1,1] with 1 representing that the two drugs share the same interacting proteins and drugs' action on each protein is the same while -1 representing that the two drugs share the same proteins but drugs' action on each protein is opposite.

v. Implement Kolmogorov–Smirnov strategy

We implemented the nonparametric, rank-based strategy based on the algorithm originally introduced by Lamb et al.(*15*) to generate a ranked list of candidate drugs for each disease. For each disease signature, we computed an enrichment score separately for the up- or down- regulated genes: es_{up} and es_{down}. In specific, we constructed a vector V of the position of each of the up- or down- regulated genes on the basis of the values from the reference drug dataset. The vector was then sorted in ascending order such that V(*j*) is the position of disease gene *j*. The computation of the enrichment score is based on Kolmogorov–Smirnov statistic and the details can be referred to in the supplementary material in Lamb et al. (*15*). The drug score is set to zero, where es_{up} and es_{down} have the same algebraic sign. Otherwise, we set the drug score to es_{up}-es_{down}. We hypothesized that those drugs with a statistically significant negative score might be a possible treatment for the disease of interest.

vi. Perform literature validation

To check whether the predicted drug-disease pairs have clinical literature evidence, we used the esearch API provided by NCBI. The query term we used is '*drug name* AND *disease name* AND (Clinical Trial[ptyp] OR Clinical Trial, Phase I[ptyp] OR Clinical Trial, Phase II[ptyp] OR Clinical Trial, Phase III[ptyp] OR Clinical Trial, Phase IV[ptyp])'. The total number of clinical type PubMed articles for each association was recorded.

6.3 Results

i. Drug directionality Map (DMAP) Construction

We constructed a probabilistic-based Drug directionality Map (DMAP). It records the directionality (i.e. activation/inhibition) between chemicals and their interacting proteins

and the strength of such directed relationships. To generate a ranking for each relationship, we developed an intuitive pharmacology score by combining the probability of a drug's action (i.e. activation/inhibition) and the significance of each interacting protein. This ranking system renders DMAP in a format essentially compatible with gene expression profiles in CMAP. Therefore, DMAP serves as a valuable alternative for researchers interested in CMAP based studies.

DMAP contains 9,486,081 ranked chemical protein interactions for 328,676 chemicals. It significantly increases the chemical coverage by over 200-fold (Table 6-1) compared to the 1,309 chemicals covered in the second release of CMAP (*15*). A Venn diagram shows the number of shared chemicals between DMAP, CMAP and drugs with known indications which we compiled from the TTD database(*118*) and literature (*22*) (Figure 6-2). CMAP contains 394 drugs with known indications. Among these, 380 drugs can also be captured by DMAP and thus only 14 drugs are uniquely covered by CMAP. On the other hand, 982 drugs in DMAP have known indications. Among these, 602 drugs are uniquely covered by DMAP. Thus, we argue that DMAP provides a valuable resource for repositioning existing drug for new uses. To demonstrate this, in the following section we applied two representative drug-repositioning methods with DMAP dataset and proved its utility for computational drug repositioning.

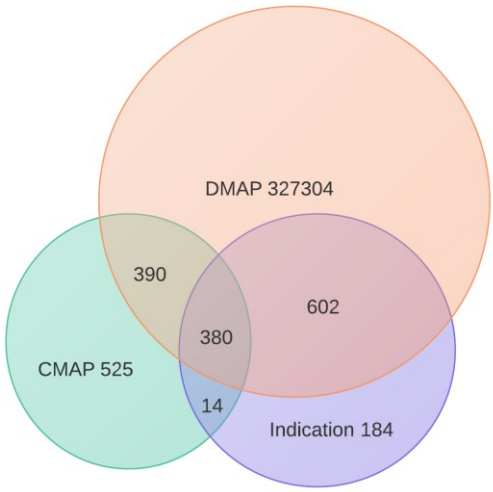

Figure 6-2.The Venn diagram of drugs from DMAP drug signatures, CMAP drug signatures and drugs with Indication

Table 6-1.Statistics summary of DMAP

Performance indicator	CMAP (build 02)	DMAP Version 1.0 (Oct 2013)
Count of drugs	1,309	328,676
Count of drugs with known indications	394	982
Count of drug protein relationship	15,472,380	9,486,081
Count of up regulations	7,710,741	4,458,335
Count of down regulations	7,763,186	5,027,746

ii. DMAP's utility for drug repositioning

To check DMAP's utility for drug repositioning, we applied the following two known drug-repositioning methods in literatures: (i) drug similarity approach, (ii) Kolmogorov–Smirnov algorithms. The former approach was nearest neighbor based approach: if two drugs were similar, the disease indication for one drug could be potentially assigned to the other drug. The latter approach was a hypothesis driven approach. It assumed that the ideal drug could reverse the gene expression in the disease condition back to that in the healthy condition. This approach had more structural assumption imbedded and thus was

different from the similarity based approach. These two approaches were the two mainstream drug repositioning approaches.

a) Drug similarity network approach based on DMAP

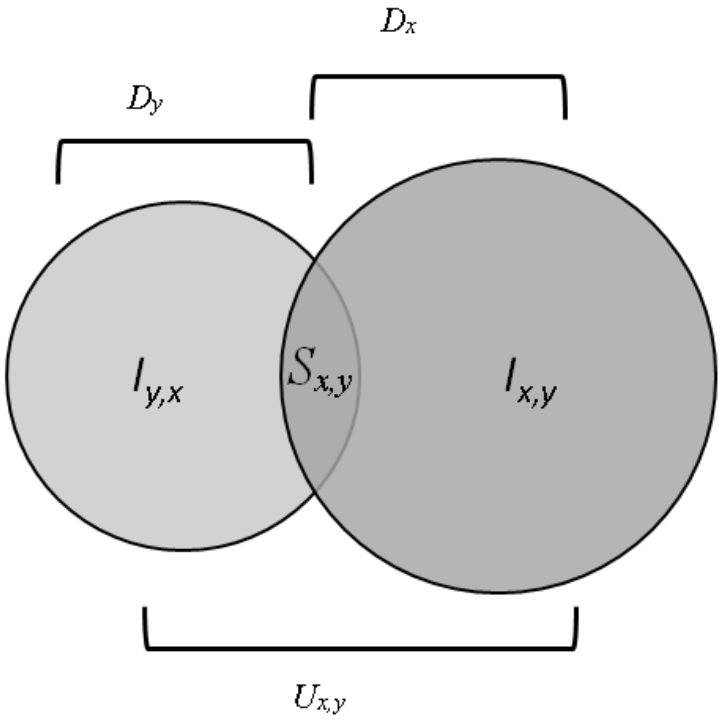

Figure 6-3. A schematic representation of the GBA method.
Given two drugs x and y and their corresponding indication profiles Ix and Iy, respectively, the potential novel uses for drug x is Iy,x. Similarly, potential novel drug uses for drug y is Ix,y.

We first checked the quality of DMAP by applying them for drug repositioning with the drug similarity network approaches developed by Iorio *et al.(40)*. We computed 481,671 pairwise drug similarities for the 982 drugs with known indications by calculating the Tanimoto Coefficient between their interacting proteins profiles.

To assess the prediction performance, we implemented the 'Guilt by Association' (GBA) concepts (Figure 6-3.) presented by Chiang *et al.(126)* and conducted "Leave-One-Out" cross-validation. For each drug, we removed its known indications and attempted to recover them by considering the indications for its top N similar drugs found. We calculated overall sensitivity and specificity by varying N—the number of similar drugs—from 1 to 981. The area under the ROC curve (AUC) score was used to measure the performance.

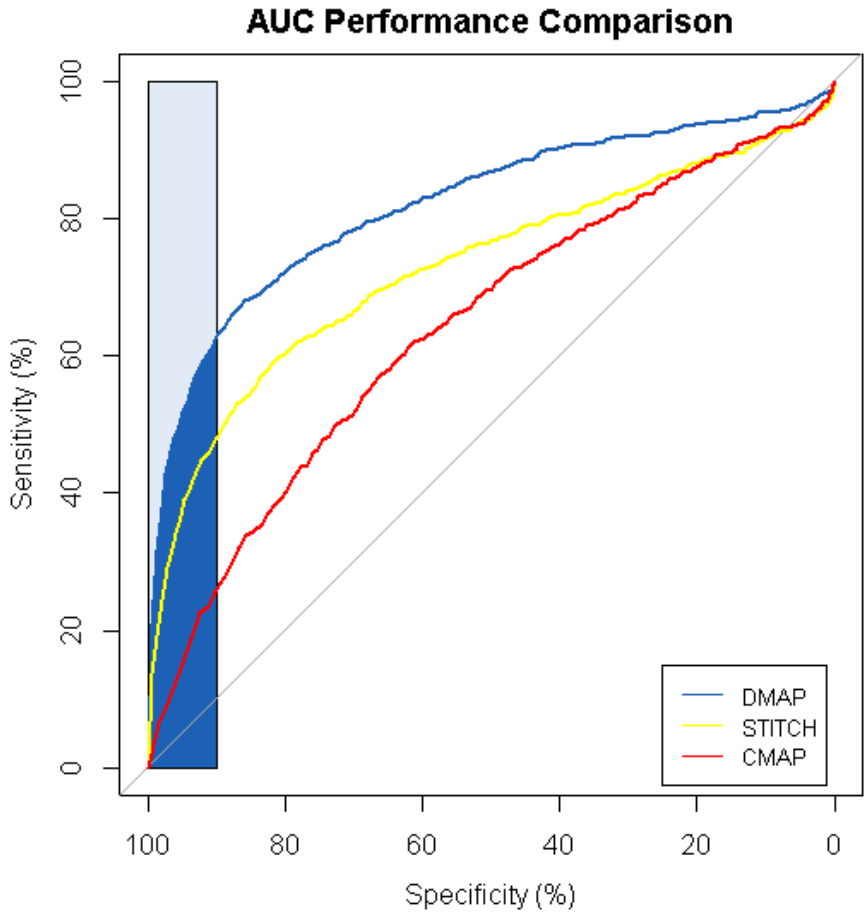

Figure 6-4. ROC curves for the prediction performance based on DMAP (blue line), STITCH (yellow line) and CMAP (red line).
Blue shade area provides a partial ROC area corresponding to specificity 90% above.

The Overall AUC for the prediction based on DMAP achieved 0.82. Most importantly, early retrieval performed well, with a partial AUC of 0.72 for specificity of 90% or above(*127*). Since one could only test the limited number of drugs in experimental setting, the good performace in high specificity region, approximately corresdponing to the top

ten candidates of all the predictions, would make the proposed drug repositioning more mearningful in practice.

In comparison, 1) we performed similar analysis based on CMAP transcriptome data and the overall AUC was 0.64. The early retrival performance was only 0.55; 2) we performed the analysis directly using the STITCH data and we got an overall AUC of 0.70. The early retrival performance was only 0.65. Figure 6-4 showed that the ROC curve based on DMAP was above the curve from STITCH which was in turn above the curve of CMAP.

To rule out the possibility that the performance difference was purely due to the drug coverage difference between DMAP and CMAP, we conducted the ROC analysis with only the shared drugs between DMAP and CMAP. The DMAP achieved an AUC of 0.81 while CMAP only achieved an AUC of 0.64 (Figure 6-5).

Out of all the possible drug pairs, we identified 3,014 significant pairs by requiring the number of overlapped proteins no less than two and the drug similarity score at the top 5% of the distribution. The resulting drug network (Figure 6-6A) showed a scale free property (Figure 6-6B), commonly observed in a biological network. Most of the drugs are well connected and formed communities. In fact, 451 drug pairs out of these 3,014 significant pairs have shared at least one known disease indication. For the remaining 2,563 pairs without overlapping indications, the novel drug disease associations from 1,206 drug pairs were supported by at least one clinical type PubMed article. Table 6-2 list the top 20 drug-disease pairs and could be a good starting point for further experimental validations.

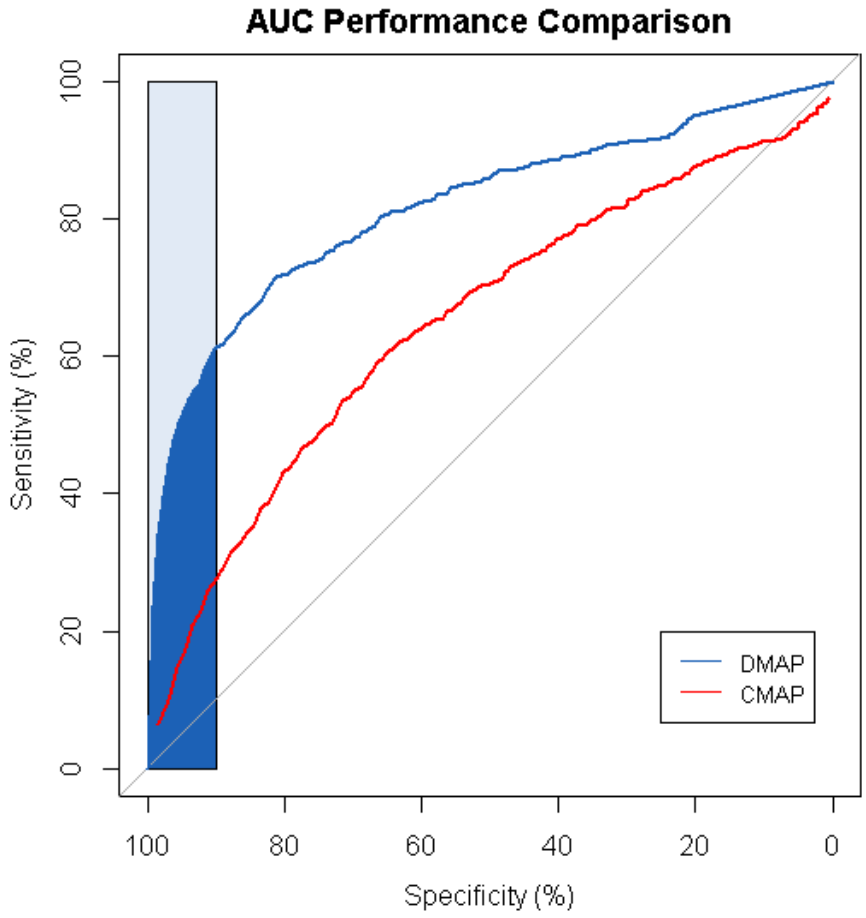

Figure 6-5. The ROC curves for DMAP and CMAP using the overlapped drugs.

Table 6-2.Top 20 novel drug repositionings and the number of clinical type publication support

Drug	Disease	PubMed(Clinical)
Rocuronium	Pain	126
Clemastine	Allergies	80
Mometasone	Asthma	78
Nicotinamide	Alzheimer Disease	45
Sotalol	Hypertension	42
Sertraline	Alzheimer Disease	40
Ifosfamide	Leukemia, Acute Myeloid	40
Gabapentin	Anxiety disorder	33
Vinorelbine	Prostate Cancer	32
Lumiracoxib	Pain	28
Hydrocodone	Anesthetic	25
Zileuton	Inflammatory diseases	20
Irbesartan	Cardiovascular disease	17

Moclobemide	Parkinson Disease	13
Fluvoxamine	Alzheimer Disease	10
Ranolazine	Dysrhythmias	6
Trihexyphenidyl	Depression	5
Nicotinamide	Breast Cancer	5
Methylphenidate	Obesity	5
Pemetrexed	Colon cancers	1

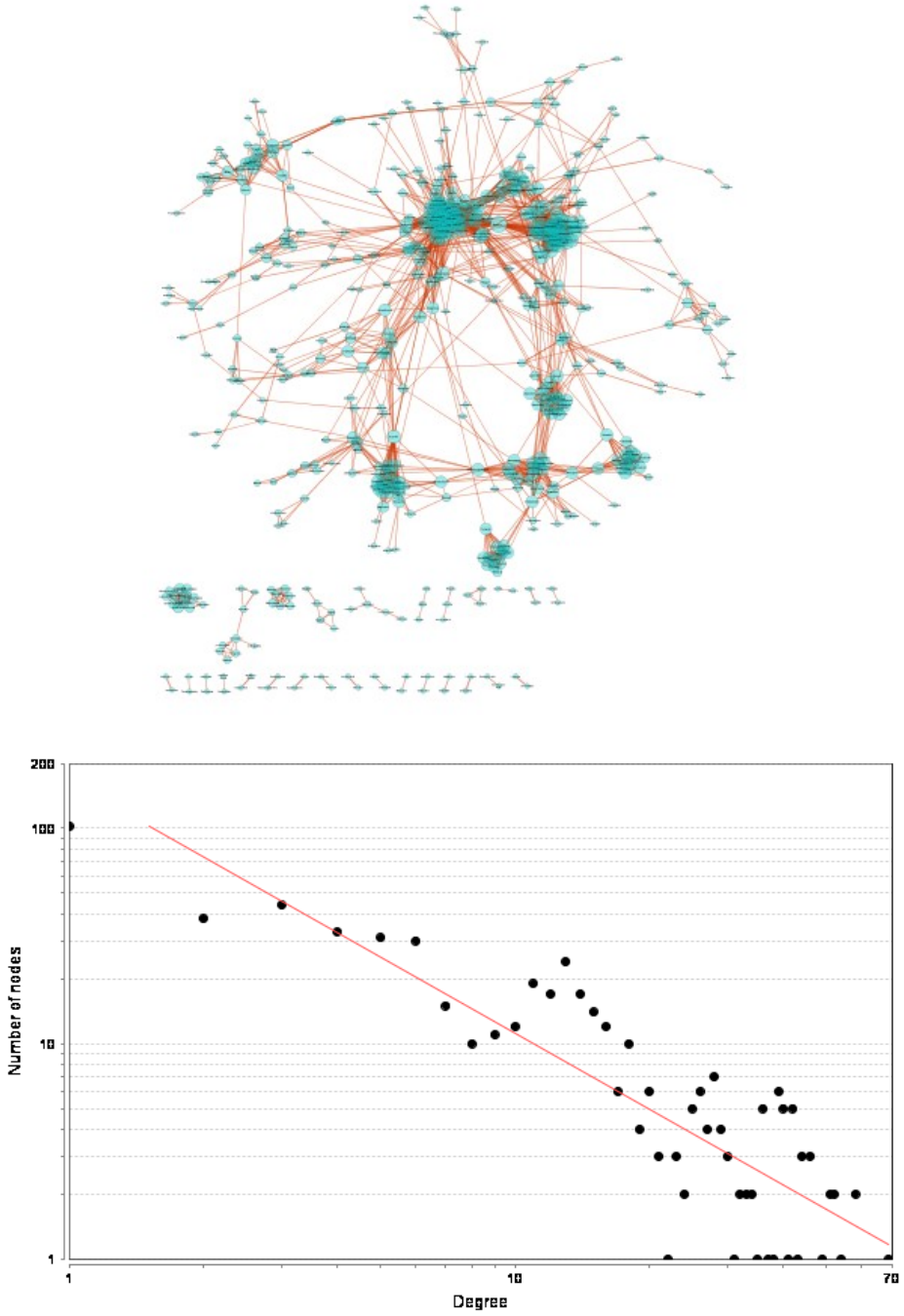

Figure 6-6. (A) Drug similarity network based on DMAP. (B) Power-law degree distribution of the network.

b) Kolmogorov–Smirnov approach based on DMAP

We next applied DMAP with the nonparametric Kolmogorov–Smirnov strategy based on the methodology originally developed by Lamb *et al.(15).* We followed *(15)* and *(17)* to use the recall of known drug indication relationship as the performance measurement for better benchmarking with previous works. We compiled the gene expression profiles for 45 distinct diseases and then queried them against DMAP and CMAP, respectively, to generate a ranked list of potential treatments for each of the diseases of interest. We calculated a similarity score for every drug-disease pair. If the similarity score is negative, the drug overall causes a reverse set of changes in the gene expression compared to that in disease condition. We hypothesized that this drug could potentially have therapeutic effects for that particular disease. To evaluate the statistical significance of the similarity score, we applied a permutation approach by randomly selecting any drug signatures and re-calculated the similarity score accordingly. We did the permutation 200 times for each drug-disease pair and computed the *p-value* by checking the actual similarity score with the score distribution after randomization.

We examined results for diseases that are the leading causes of death in the US (*128*). For breast cancer, we successfully retrieved Anastrozole, Capecitabine, Doxorubicin, Estradiol, Megestrol, Paclitaxel, Testosterone and Testolactone as possible therapeutic drugs for breast cancer. When the CMAP dataset was utilized, only Paclitaxel was retrieved as a potential therapeutic drug. For lung cancer, we retrieved Cisplatin and Etoposide by using the DMAP. However, when CMAP was used, we were not able to retrieve any drugs for lung cancer. Table 6-3 also contains the results for other diseases. To have statistical significance, we required the *p-value* less than 0.05 in Table 6-4. CMAP did relatively better in the case for Alzheimer's disease and Leukemia. For these known relationships covered in CMAP but not DMAP, or vice-versa, some were due to having borderline *p-value* while others were due to violating our hypothesis of negative correlation. Overall, DMAP and CMAP database were complimentary to each other.

Table 6-3.Retrieval of known disease drug relationships from DMAP and CMAP, respectively

The star rating is labeled according to the following criteria:
K-S Score<-0.3: ★★★★★
-0.3≤K-S Score<-0.2: ★★★★
-0.2≤K-S Score<-0.1: ★★★
-0.1≤K-S Score<0: ★★
K-S Score≥0 or *p-value*≥0.05:★

Disease	Drug	DMAP	CMAP
Breast Cancer	Anastrozole	★★★★	
Breast Cancer	Capecitabine	★★★★★	
Breast Cancer	Doxorubicin	★★★★	★
Breast Cancer	Estradiol	★★★	★
Breast Cancer	Megestrol	★★★★★	★
Breast Cancer	Paclitaxel	★★★★★	★★★★
Breast Cancer	Testolactone	★★★★★	
Breast Cancer	Testosterone	★★★	★
Colorectal Cancer	Capecitabine	★★★	
Colorectal Cancer	Leucovorin	★★★★	
Colorectal Cancer	Raltitrexed	★★★★★	
Lung Cancer	Cisplatin	★★★★	
Lung Cancer	Etoposide	★★★★	★
Prostate Cancer	Docetaxel	★★★★★	
Prostate Cancer	Leuprorelin acetate	★★★★★	
Parkinson's disease	Galantamine	★★★★★	★
Parkinson's disease	Trihexyphenidyl	★	★★★★
Alzheimer's disease	Galantamine	★	★★★★
Alzheimer's disease	Memantine	★	★★★★
Alzheimer's disease	Tacrine	★★★★★	★
Diabetes	Liraglutide	★★★★★	
Diabetes	Vildagliptin	★★★★★	
Leukemia	Carmustine	★	★★★★
Leukemia	Celecoxib	★	★★★★
Leukemia	Idarubicin	★★★★	
Leukemia	Irinotecan	★	★★★
Leukemia	Isotretinoin	★	★★★★
Leukemia	Methotrexate	★	★★★
Leukemia	Pentostatin	★★★★★	
Asthma	Bambuterol	★	★★★★
Asthma	Dexamethasone	★★★	★
Asthma	Methylprednisolone	★	★★★★

Asthma	Orciprenaline	★	★★★★
Asthma	Prednisone	★★★★★	★
Asthma	Salbutamol	★★★★★	★
Asthma	Salbutamol sulphate	★★★★★	
Asthma	Theophylline	★	★★★★

Besides recalling the known drug-disease relationships, this method could also propose novel drug-disease associations. National Center for Advancing Translational Sciences (NCATS) (*117*) provides a list of drugs for translational medicine researches. We cross checked the novel predictions with their drug list. Here, we highlight a few drug-disease relations.

A novel relation that DMAP results suggest is between Vincristine, a drug typically used for Leukemia, and Wilm's tumor. A recent study performed by Indolfi *et al.(129)* revealed that there is a potentially higher rate of survival in patients with bilateral Wilm's tumor when patients are given a dosage of vincristine/actinomycin D.

Nifedipine is usually used to treat high blood pressure and angina. The DMAP results suggest that Nifedipine can also be used to treat asthma. Since Nifidipine is a PKC inhibitor and PKC is a potential therapeutic target for asthma (*130*), it is a potential treatment for asthma. Cheng *et al.*(*131*) demonstrated in their study that Nifedipine can help control the constriction involved in sensitized tissue in asthma. Furthermore, another study by Barnes *et al.(132)* suggested that Nifidipine modifies exercise-induced asthma.

Progesterone is a prescription drug used for women taking estrogens after menopause and is also used for treating amenorrhea. The DMAP results suggest that progesterone can be used to treat breast cancer. In the study by Groshong *et al.* (*133*), it was determined that treatment with Progesterone can be used to regulate Breast Cancer cell growth.

Table 6-4 summarized the all the novel drug repositioning predicted by both similarity approach and KS algorithms, which could be starting point for further experimental validation.

Table 6-4. Drug repositioning predicted by both similarity approach and KS algorithms

Drug	Disease Indication
Mebendazole	Alzheimer's disease
Amiodarone	Asthma
Anastrozole	Asthma
Anastrozole	Asthma
Anastrozole	Asthma
Benztropine	Asthma
Chlorzoxazone	Asthma
Drospirenone	Asthma
Econazole	Asthma
Fluvoxamine	Asthma
Itraconazole	Asthma
Methylergonovine	Asthma
Methylergonovine	Asthma
Methylergonovine	Asthma
Oxybutynin	Asthma
Oxybutynin	Asthma
Rivastigmine	Asthma
Sertraline	Asthma
Sitaxsentan	Asthma
Tioconazole	Asthma
Tioconazole	Asthma
Vinorelbine	Asthma
Amodiaquine	Breast Cancer
Atovaquone	Breast Cancer
Flucytosine	Breast Cancer
Fluticasone Propionate	Breast Cancer
Fluvoxamine	Breast Cancer
Methylergonovine	Breast Cancer
Raltitrexed	Breast Cancer
Repaglinide	Breast Cancer
Rivastigmine	Breast Cancer
Spirapril	Breast Cancer
Tioconazole	Breast Cancer
Trifluridine	Breast Cancer

Trimethoprim	Breast Cancer
Voriconazole	Breast Cancer
Zafirlukast	Breast Cancer
Atovaquone	Colorectal Cancer
Cyclizine	Colorectal Cancer
Flucytosine	Colorectal Cancer
Suprofen	Colorectal Cancer
Tolcapone	Colorectal Cancer
Trifluridine	Colorectal Cancer
Valdecoxib	Colorectal Cancer
Acenocoumarol	Parkinson's disease
Anastrozole	Parkinson's disease
Bivalirudin	Parkinson's disease
Capecitabine	Parkinson's disease
Cyclizine	Parkinson's disease
Pyrimethamine	Parkinson's disease
Suprofen	Parkinson's disease
Valdecoxib	Parkinson's disease
Flucytosine	Prostate Cancer
Rivastigmine	Prostate Cancer
Trimethoprim	Prostate Cancer
Trimethoprim	Prostate Cancer
Voriconazole	Prostate Cancer

6.4 Conclusions

Critical to drug repositioning involves the reliable measurements of how drug affect disease proteins. In this work we presented a computational drug directionality resource called DMAP to address the challenges. We demonstrated that the resource can greatly facilitate the drug discovery process for the following reasons: access to disease gene drug relationship data with high coverage and quality; incorporating prior knowledge about biological significance with protein interaction network.

This study differs from previous research in that it provides a comprehensive database of computationally derived drug-protein relationships. Previous efforts (*17, 40, 47, 48*) on paring the expression of drugs and diseases mainly rely on experimental connectivity map. For example, Sirota *et al.(17)* performed a large-scale integration of expression signatures of human diseases from the public data with CMAP drug signatures. This

work provides another alternative resource of directed drug-protein relationships. The drug similarity study proves the validity of the probabilistic based directionality for each drug-protein relationship. The implementation of K-S algorithm proves the compatibility of the pharmacology score based ranking with the expression based ranking in CMAP for the drug repositioning research. With these two major drug repositioning approaches, the knowledge base from DMAP performed better than directly using the microarray data from CMAP. It can thus serve as a valuable resource for drug repositioning studies.

One limitation of DMAP lies in that the number of interacting proteins for each drug is not a constant number. For the gene expression based profiles in the CMAP database, each drug was measured against the same number of proteins in experiments while in DMAP the number of interacting proteins varies from drug to drug In DMAP, 64,034 drugs have at least 10 activated and inhibited proteins. 13,098 drugs have at least 50 activated and inhibited proteins and 3,515 drugs have at least 100 activated and inhibited proteins. Despite of this limitation, the database served its purpose for systematic drug repositioning as demonstrated in this work.

Chapter 7. Drug Repositioning using Side Effect Features: from 1D to 2D

This section is based on my work at (*54*). HH and LY conceived the idea. HH constructed the drug combination database, built the decision tree model, evaluated the prediction performance, validated the out of sample predictions and wrote the manuscript. XAQ helped analyze the case studies.

7.1 Introduction

The use of multiple drugs with different mechanisms or modes of action may treat the disease more effectively (*142-144*). The traditional "one drug – one target – one disease" approach has been used to develop successful drugs. However such "magic bullet" sometimes shows limited efficacy, especially for complex diseases (*145*). It is often due to factors such as network robustness (*146*), redundancy (*147*), compensatory and neutralizing actions (*148*). Polypharmacology, which focuses on multi-target drugs, has the potential (*11*) to address those limitations. High-throughput screening was used to identify possible drug combinations (*149*); however, it is impractical to screen all possible drug combinations for every indication. Therefore, computational methods (*150-153*) have been developed to predict new drug combinations. For example, network biology was introduced to investigate drug combinations by studying the molecular networks or pathways affected by the drugs (*154*) yet the incompleteness of molecular networks limits the practical use of such approaches for prediction of novel drug combinations.

Besides the molecular information-based approaches, clinical phenotypic information has not been adequately investigated for its power in predicting drug combinations. The advantages of leveraging on clinical phenotypic information include better translational power when comparing with animal models (*155*) since it mimics a phenotypic screening of the drug effects, both therapeutic effect (*46, 156*) and toxic effect (*157, 158*), on human. In this work we propose an innovative approach by using observed side effects reported in clinical findings to identify novel safe and efficacious drug co-prescriptions or fix-dose combinations.

In this study, we summarized the prediction of novel drug combination as a two-step effort: 1) to minimize the potential side effect of the new combination; 2) to avoid reduction of the efficacy via pairing the indication for each drug in the new combination. We hypothesized that drugs that can be put together usually do not have serious adverse drug reactions (ADRs) in common. We tested this hypothesis by identifying a set of three FDA blacklisted side effects from marketed drug combinations and evaluated its prediction performance in both the training and the validation set. Our results support that using these features, clinicians could rule out unsafe drug pairs with high confidence. We further demonstrated such classification power is not due to the synthetic confounding factors such as biased disease indications or drug targets. We further proposed both components in the pair to treat the same disease so that therapeutic effects from each component could be added in the combination. This two-step rule provides a novel approach to identify novel drug co-prescriptions or combination from using of clinical side effects, which should be less of a translational issue compared to animal model. We applied this approach to identify 977 candidate drug combinations. 144 pairs (15%) are supported by clinical trials from clinicaltrial.gov for the same indication, leaving 85% potential novel combinations to be evaluated in future clinical studies.

7.2 Methods

i. Preparation of datasets

Side effect datasets. SIDER is a side effect database containing information on marketed medicines and their recorded adverse drug reactions. The information is extracted from public documents and package inserts (*159*). In this study, we downloaded the entire database from http://sideeffects.embl.de/. Besides relying on drug label as sources for drug side effects, we also checked FAERS, a database that contains information on adverse event submitted to FDA and is designed to support the FDA's post-marketing safety surveillance program for drug and therapeutic biologic products. OFFSIDES is such a side effect database by mining FAERS system while controlling those confounding factors such as concomitant medications, patient demographics, and patient medical histories and so on. OFFSIDES contains 1332 drugs and 10097 side effects. 438 drugs and 2322 side effects are shared between SIDER and OFFSIDE. In the final

integrated side effect database, drugs are represented with STITCH ID while side effects represented with MedDRA terms so that they could be integrated across databases.

The TWOSIDES database identifies 59,220 pairs of drugs with 1,301 adverse events by carefully matching groups of patients in the post-marketing surveillance system FAERS. It provides a reliable and comprehensive database of side effects for drug pairs. It is thus used to identify the features enriched in approved DDCs compared to random drug pairs. In contrast, when doing the DDC prediction, we only used the side effect for single drugs from drug label and OFFSIDES since it is logical to only have single drugs' side effect data before such pair has come into being.

Drug combination datasets. The Drug Combination Database (DCDB) is a database collecting and organizing known examples of drug combinations. The current version contains 145 drug combinations. Peer Bork's paper also lists 178 drug combinations, mainly collected from FDA orange book. We also curate 236 FDA approved or registered drugs from literature. After mapping them to STITCH ID, we get a comprehensive list of 349 drug combinations.

Drug target and ATC code. DrugBank (http://www.drugbank.ca) is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information. Current version contains 6711 drugs and 4081 targets. We downloaded the full database in xml format and parsed out the drug target pairs and drug ATC pairs.

ii. Analysis methods

Two Step Rule for making safe drug combination or co-prescriptions. First step is to make sure what drugs can be safely put together. We hypothesize that the drugs that can be put together usually do not have overlap in some serious adverse drug reactions (ADR), but might share some side effects that contribute to the therapeutic effect (*46, 156*). Here we came up with a practical black list consisting of three side effects for clinicians to decide the safe drug pairs with high accuracy; at the second step, we

required that those safe pairs should further have the same disease indications to achieve the similar efficacy (Figure 7-1).

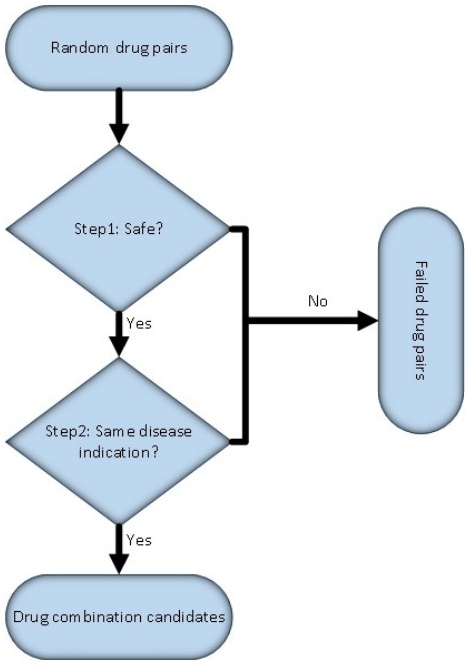

Figure 7-1. Illustration of the Two-Step Rule to predict the drug combinations.

Feature selections For each side effect, we built a two by two table and performed a Fisher's exact test to determine whether that side effect is differentially shown up between positive DDCs and negative DDCs. Then we used *p-value* less than 0.05 as the threshold to pick the significant features. When we developed the black list consisting of three side effects, we first used information gain as the statistical significance measurement to identify the top ten features. To get the biological significance, we then chose three out of the top ten according to the origin of their organs/human systems.

Machine learning models we used decision tree as main machine learning models, J48 decision tree algorithms in WEKA(*160*). We also tested the performance with Naïve Bayes, Logistic regression and random forest. All of them gave even better AUC and accuracy than decision tree. Since we are more interested to develop a simple rule to be

easily applied in clinical than achieving a better AUC, we presented the results based on decision trees model in this work.

PubMed and Clinical Trial Validation To validate whether the predicted drug pairs have clinical literature supports, we used the esearch API provided by NCBI to count the co-occurrence of the drug components for each proposed DDCs. The query term we used are ‘drug name1 AND drug name2 AND (Clinical Trial[ptyp] OR Clinical Trial, Phase I[ptyp] OR Clinical Trial, Phase II[ptyp] OR Clinical Trial, Phase III[ptyp] OR Clinical Trial, Phase IV[ptyp])’. We also checked clinicaltrial.gov to see whether predicted drug pairs are co-mentioned in the same registered clinical trials.

7.3 Results

i. Construction of the data set

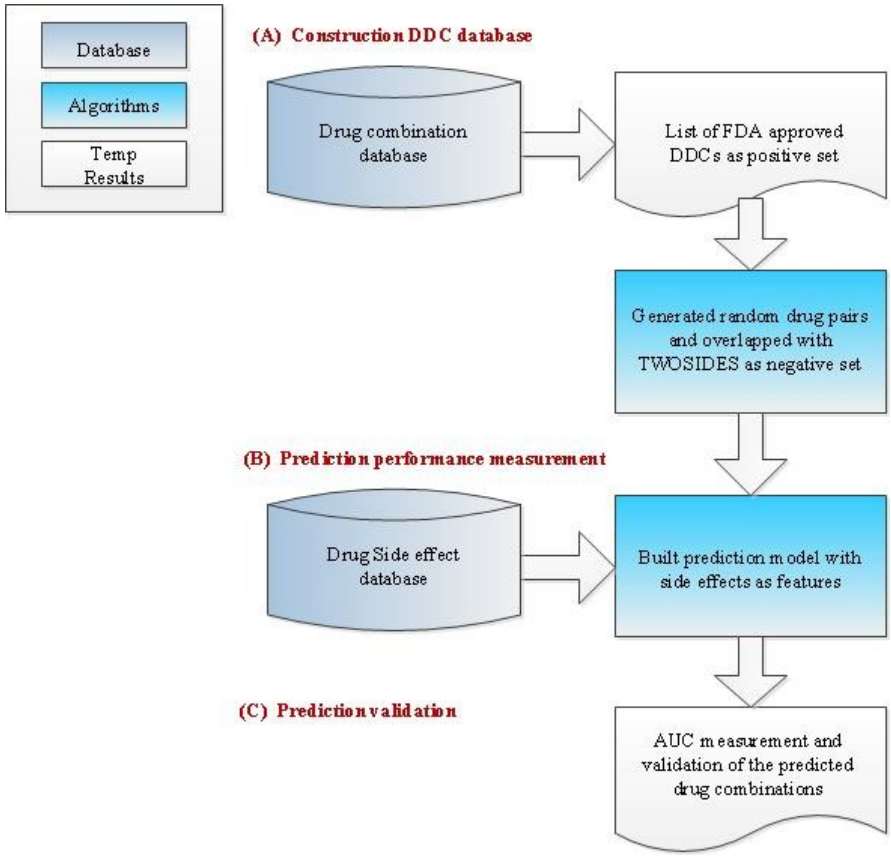

Figure 7-2. Workflow of applying logistic regression and decision tree models to measure the DDC prediction performance with side effects as features

We constructed a comprehensive drug combination database (Figure 7-2A) which contains 349 approved pair-wise drug-drug co-prescriptions/combinations (DDC) from three different sources: drug combination database DCDB (*161*), a recent drug combination paper (*153*) and manual literature curation of the FDA approved or registered DDCs. To resolve different naming issues in different data sources, DDCs were represented by their two components whose names were mapped to STITCH ID (*32*) for comparison. (Venn diagram comparison of these three sources was shown in Figure 7-3)

To annotate drugs with their side effect features, we extracted side effect information from drug labels using SIDER (*159*) and OFFSIDES (*157*) (Figure 7-2B) . SIDER

derives side effects from drug labels and OFFSIDES mines side effects from post-marketing surveillance system FAERS (i.e. FDA Adverse Event Reporting System). Of the 349 approved DDCs, 239 DDCs can be annotated with side effects for both components, which correspond to 245 individual drugs and 7,888 side effects. As a comparison, previous work (*153*) used 181 pair-wise DDCs, out of which only 75 contains both side effects and indication annotation due to the limited data sources for DDCs, side effects and indications. Therefore the coverage of this database is much more comprehensive.

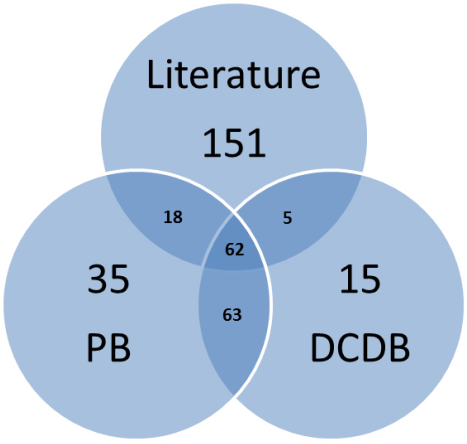

Figure 7-3. The Venn diagram of drug combinations, where the numbers indicate how many drug combinations can be covered by each data source

We also constructed a negative training set consisting of unsafe drug pairs for training the DDC prediction model. We defined the unsafe co-prescriptions as those causing unexpected side effects as tracked in TWOSIDES (*157*), a database of reported side effects only caused by the combination of marketed drugs rather than by any single drugs from FAERS. For those 245 drugs in the positive set, we generated all the possible pairs of combinations while excluding those 239 positive DDCs. Then the left drug pairs were overlapped with those drug pairs in TWOSIDES. A resultant set of 2291 unsafe drug pairs (8% of all the possible drug combinations for the 245 drugs) were identified and used as the negative training set for training the DDC prediction model.

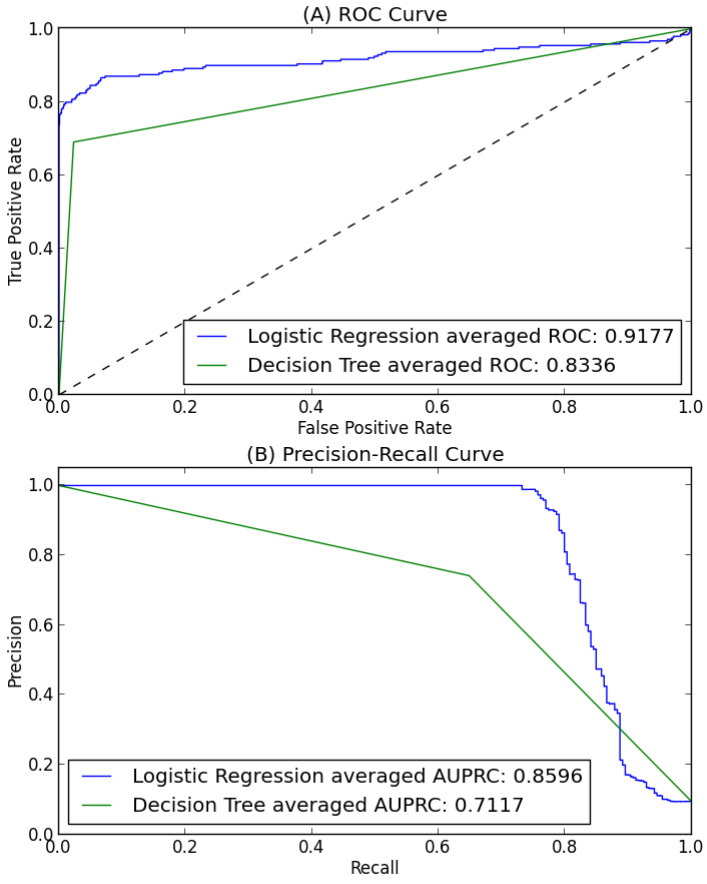

Figure 7-4. Evaluation of logistic regression and decision tree models based on the full dataset (i.e., 239 marketed DDCs and 2291 unsafe drug pairs).
(A) ROC curve. (B) Precision-Recall curve.

ii. Evaluation of the power of predicting DDCs based on the side effects features

We used 239 marketed DDCs as positive set along with 2291 unsafe drug pairs as negative set. Each side effect of a drug is called a feature and a drug pair can be represented as a vector of side effect features with value of 0, 1 and 2 depending whether zero, one or both drugs have such side effect. We applied logistic regression model with 10-fold cross validation to evaluate the performance. We measured the model performance with both AUC (area under the ROC curve) and AUPRC (area under the precision-recall curve). We repeated the cross-validation experiment 100 times with random seeds, and computed the mean and the standard deviation of AUC and AUPRC over the 100 repetitions. In the experiment, logistic regression model achieved an AUC of 0.92±0.01 and AUPRC of 0.86±0.01 (Figure 7-4). The overall AUC is 0.94 and the early retrieval performs as well with a corrected pAUC of 0.92, which enables us to keep false positives low (*162*) while sacrificing some true positives.

To explore how unbalanced positive set and negative set affects AUC, we randomly sampled from the negative set 100 times. Each time we made the negative set with the same number of drug pairs with positive set. The average AUC was 0.95 (Figure 7-5) with standard deviation of 0.02.

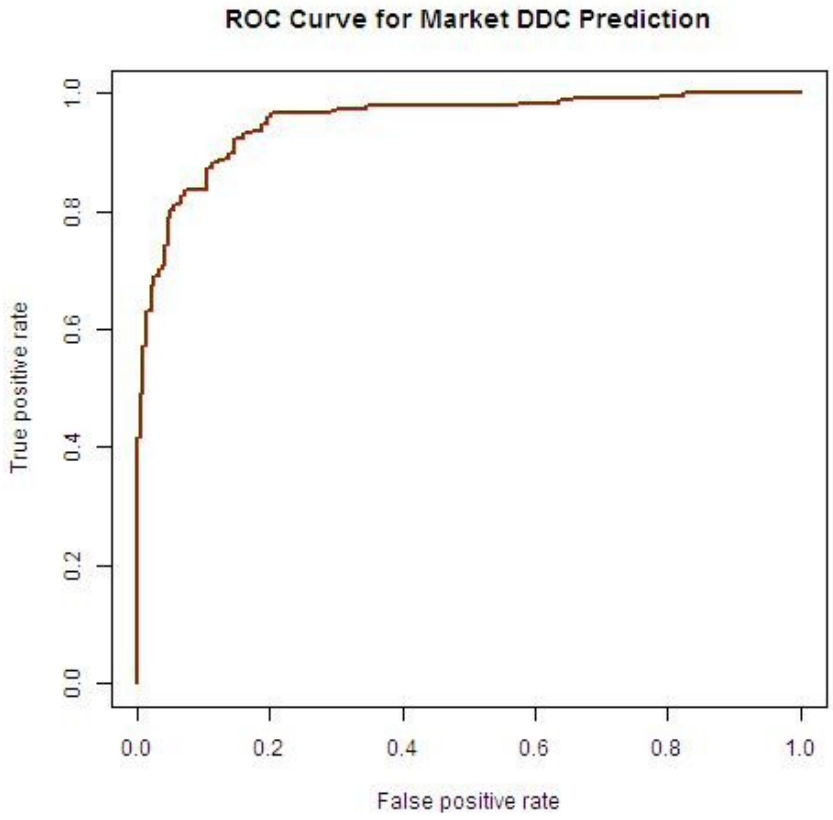

Figure 7-5. Evaluation of logistic regression based on 239 marketed DDCs with balanced positive set and negative set.

To exclude the fact that the good AUC score is just due to homologous relationships between structurally similar drugs, we mimicked the method in Gottlieb's work (*163*) by removing the drug pairs with Tanimoto similarity coefficient larger than 0.50. We re-run the logistic regression 10-fold cross-validation experiment 100 times and still achieved AUC of 0.92±0.01 and AUPRC of 0.86±0.01,which is the same with previous results to two decimal places.

Since the datasets are made of drug pairs, it is possible that some drugs occur in both the training set and testing set although no drug pairs are shared between these two sets. To further characterize the predictive model, hold-drug-out validation had been used to evaluate the performance of the method. The original 2530 drug pairs are made of 245 distinct drugs. From the 245 drugs, we randomly chose 60 drugs as testing drugs (i.e., about 25%) and 185 drugs as training drugs (i.e., about 75%). We held out all drugs pairs

involved with the testing drugs, rather than holding out drug pairs directly. From the 2530 drug pairs, we selected the pairs only involving with training drugs as training data. Then we constructed predictive models with the training data. From the 2530 drug pairs, we selected the pairs only involving with testing drugs as validation data for testing the models. Again, we repeated the hold-drug-out validation experiment 100 times with randomly partitions, and computed the mean and the standard deviation of AUC and AUPRC over the 100 repetitions. In the experiment, logistic regression model achieved an AUC of 0.87±0.03 and AUPRC of 0.76±0.07. The additional results show that the predictive model performed still well even in the situation where none of the pair members in the test set are within the training set.

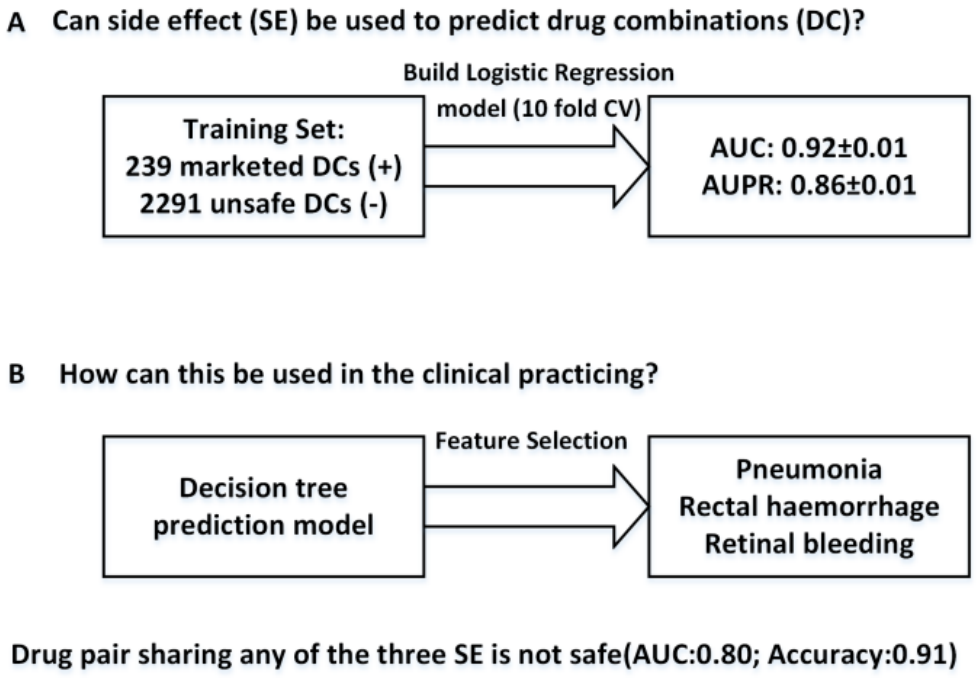

Figure 7-6. The outline of this study
(A) build logistic regression models to measure the DDC prediction performance with side effects as features; (B) build rule based model that can be easily applied in clinical settings

The results of logistic regression showed the strong performance of the DDCs prediction with side effect features. Next we focused on how to develop a simple rule for the

clinicians or the drug developers to use in their daily work in making co-prescriptions or fix-dose drug combination (Figure 7-6). A different model, decision tree model (*164*), was thus tested. The model showed an AUROC of 0.83±0.01 and AUPRC of 0.71±0.01, not as good as the performance in the logistic regression model. However considering that decision tree model is easier for interpretations in practice and such a white-box model is much more accessible to clinicians, we used the decision tree model for the further analysis.

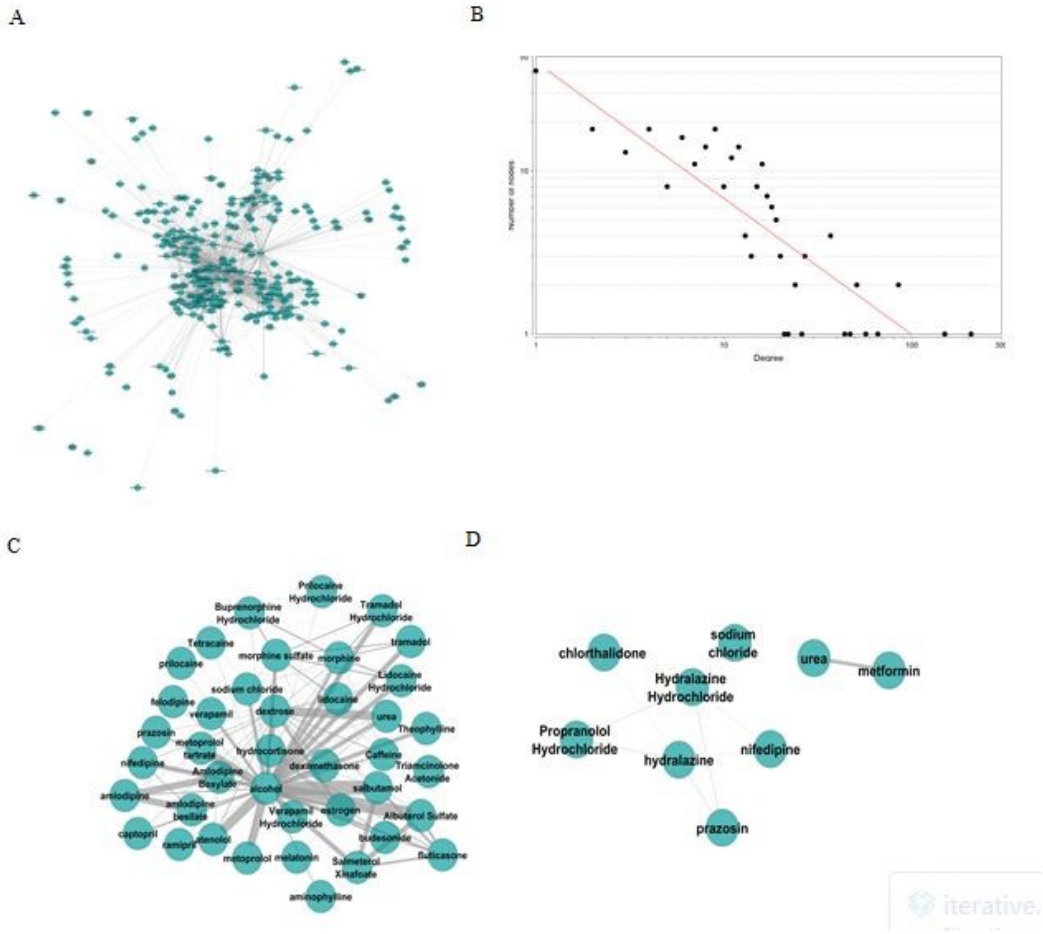

Figure 7-7. Drug combination networks.
(A) A network view of the 1508 drug pairs with prediction probability above 0.99 and support from at least 10 clinical type publications.(B) Degree distribution of the network. (C) The sub-network cluster. (D) A network view of the 11 drug pairs with prediction probability less than 0.01 and support from at least 10 clinical type publications. The edge width is proportional to number of clinical literature supports.

For the prediction test set, we used all the possible pair-wise drug combinations of 245 marketed DDCs, excluding both positive and negative set. In total 27,360 drug pairs were used as prediction test set. With the trained decision tree model, we made the prediction on the testing set and only pairs with predicted probability above 0.99 and co-occurred in at least 10 publications of clinical trial publications in PubMed were considered as candidate DDCs. As a result, 1508 drug pairs were identified and they formed a well-connected network (Figure 7-7A) and the degree distribution is approximately a Power-law Distribution (Figure 7-7B). This well-connected network indicates that those drugs in

the network are inherently promiscuous to each other and would have a higher potential to be combined with their close neighbors we further identified a condensed sub-network, highly interconnected regions in the network (Figure 7-7C) with Cytoscape (*86*) plugin MCODE (*165*). The connections between the hub drugs include some known drug combinations like hydrocortisone and dexamethasone (immunosuppressants) (*166*), morphine and tramadol (pain relievers) (*167*). Other drugs paired with morphine or dexamethasone within this sub-network could be good starting points for further experimental validation for novel drug combinations.

Among these 1508 predicted safe DDCs, 31 pairs contain at least one clinical trial evidence according to clinicaltrial.gov as pairs, including 6 pairs in phase I, 7 in phase II, 12 in phase III, and 4 in phase IV. In contrast, for the 615 drug pairs with probability less than 0.01, only 11 are supported by at least 10 publications of clinical trial types in PubMed and with a much spare network (Figure 7-7D) compared to Figure 7-7A (*p-value* of 4.19×10^{-7} of chi-square Test). When searching them against clinicaltrial.gov, only 2 of them have clinical trial records.

Besides the results presented above with side effects integrated from both sources, we also checked the prediction performance by using side effects only from drug label (i.e. SIDER) or OFFSIDES with various machine learning models. If using drug label alone, the classification performance is as follows: AUC of 0.69 for Logistic Regression model; AUC of 0.68 for Naive Bayes model and AUC of 0.54 for decision tree model; If using OFFSIDES alone, the classification performance is as follows: AUC of 0.77 for Logistic Regression model; AUC of 0.71 for Naive Bayes model and AUC of 0.57 for decision tree model. The most predictive model was the one that included information from both OFFSIDES and SIDER, followed by OFFSIDES alone, then SIDER alone, which is consistent with previous findings (*157*).

iii. Development of the rule-based model for DDC prediction

Upon proving that the SEs could be used to predict DDCs, we next aimed at constructing a rule-based method to help the decision-making in a much easier and explainable way. We summarized this method as a two-step workflow (Figure 7-1):

Step 1: Prevent unsafe co-prescriptions based on only three SEs

Here we aimed to find those side effects as markers to identify unsafe drug pairs. Of the 239 approved DDCs in the database, 41 DDCs can be annotated with the side effect features. To get the random drug pairs, we generated all the possible pair-wise drug pairs from the 41 DDCs while excluding these 41 approved DDCs. We got 949 random drug pairs and 749 (Figure 7-8) can be annotated with side effects from TWOSIDES. We performed a Fisher's Exact test for every side effect between these approved group and the random group, with 65 side effects identified as significant (p-value <0.05).

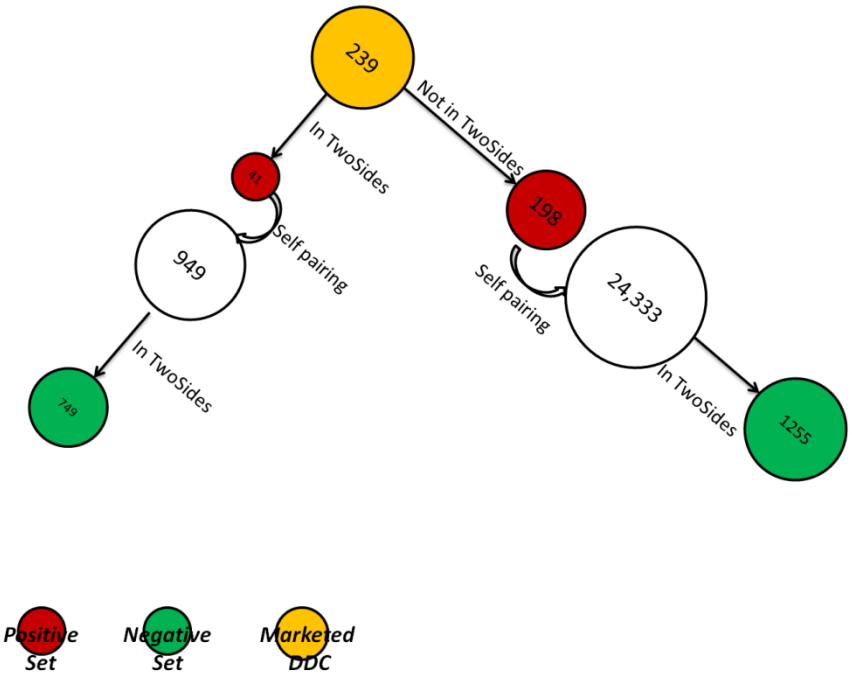

Figure 7-8. Constructions of positive sets and negative sets from the 239 DDCs in the development of the FDA black list consisting of three side effects.
Features were selected based on the positive and negative set on the left side of the figure while independent validation was done based on the positive and negative set on the right side.

Next we tested the performance of using these 65 side effects to differentiate the approved DDCs from the random set. The positive set consists of 198 approved DDCs by ruling out those 41 DDCs in the training set from the 239 approved DDCs. To build the negative set, we randomized drug pairs in the positive set and 1255 unsafe drug pairs were identified by overlapping with TWOSIDES (Figure 7-8). The number of DDCs in

positive set is approximately increased 5 times in the testing set compared to the training set. We built the features for each drug pair by checking whether zero, one or both of the drugs have any of the 65 side effect features. Similarly we applied decision tree analysis using WEKA (*168*). The AUC is 0.87 and the accuracy is 0.94. We checked the top 10 features (Table 7-1) ranked by the information gain as statistical significance, and to have biological significance we chose three side effects out of these top 10 based on a wide-spread of human organs or body systems of these side effects: pneumonia for lung or respiratory system, hemorrhage rectum for rectum or digestive system, and retinal bleeding for eye or visual system. With only these three side effects, the decision tree (Figure 7-9) achieved the AUC of 0.80 and accuracy of 0.91, supporting the superior performance of using this signature of three side effects to identify safe drug combinations. Other combinations of any three side effects from the list of 10 achieved lower performance. Based on this decision tree model (Figure 7-9), the candidate safe drug pairs should not have any of these side effects. We did a Fisher's exact test between the approved DDCs and random drug pairs to tell whether overlapping of any of the three side effects between these two groups is significantly different and the *p*-value is 2.66×10^{-33} with an odds ratio of 6.6 (Table 7-2).

Table 7-1. Top 10 side effects features from the decision tree model

Side effects
Pneumonia
Haemorrhage Rectum
Neurodermatitis
Retinal Bleeding
Allergic Alveolitis
Muscle Disorder
Vitamin B 12 Deficiency
Candida Infection
Proctitis
Infectious Mononucleosis

Table 7-2. Confusion matrix of the relationships between having the three SEs in the black list and being the unsafe co-prescription

	Share any of three SEs	Share none of three SEs
Unsafe	1254	1037
Approved combination	37	202

To rule out the possibility that the performance based on side effects as features is purely due to confounding factors like drug category, drug target, or disease indications of the drugs, we measured the drug combination classification performance with only the ATC (anatomical therapeutic chemical classification) drug category, drug targets, or disease indications. First we built a decision tree model on the same positive and negative sets with third level ATC code (*153*) as the features of pharmacology actions. It achieved an AUC of 0.62 with the top three features: G03C (i.e. Estrogens), N02A (i.e.Opioids), and C09A (i.e. Ace Inhibitors, Plain). Even with all the 100 ATC as features, the maximum AUC that can be achieved is 0.72, still less than the performance of the model based on the signature of only three side effects (i.e. AUC of 0.80). Similarly we built a decision tree model with drug targets as features and it achieved an AUC of 0.57 with the top three features: NR3C1, NR1I2, and rplD. Even with all the 296 target proteins as features, the AUC is 0.61, still less than the classification performance based on the signature of only three side effects. Finally we built a decision tree model with disease indication as features and it achieved an AUC of 0.54 with the top three features: Addison's disease, Eczema, and Prostate cancer. Even with all the 262 diseases as features, the AUC is 0.78, still less than the classification performance based on the signature of only three side effects. In sum, the decision tree model based on the signature of three side effects as features can achieve the highest performance to classify drug safety issues and it is not purely due to the co-founding factors like drug category, drug target, or disease indications.

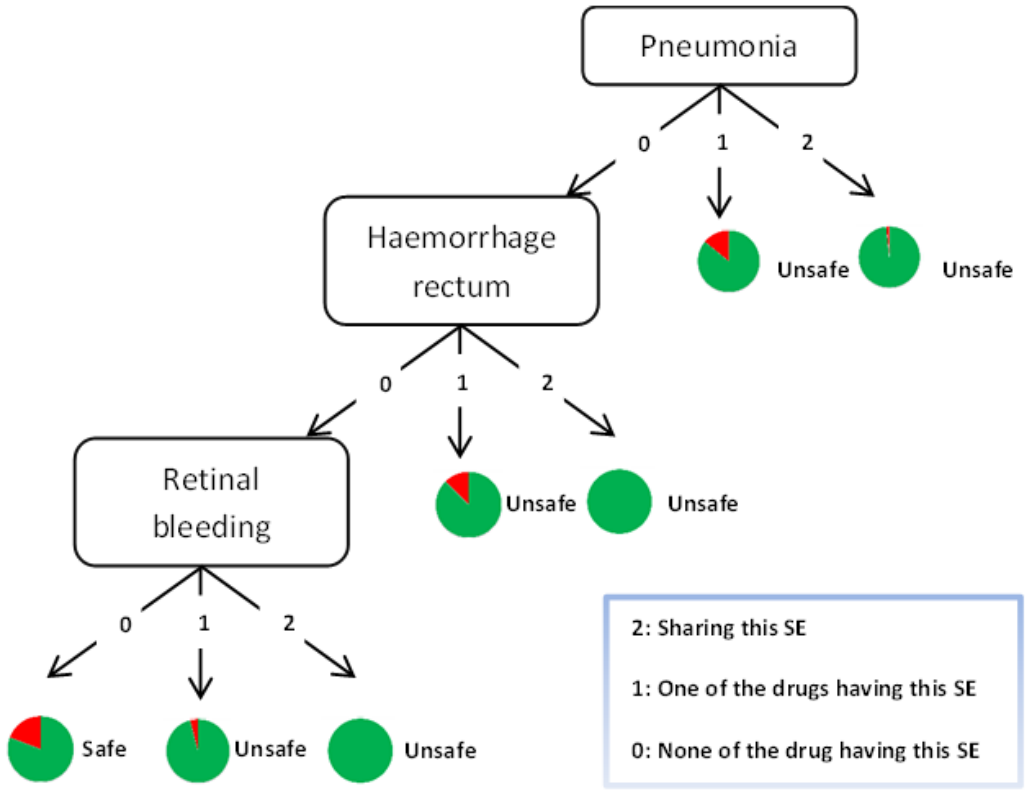

Figure 7-9. The decision tree model to decide the drug pair safety.
0, 1, and 2 indicates the number of drugs in the drug pair with such side effect. Pie charts indicate the percentage of correctly classified (green) and in-correctly classified (red) instances at each leaf. Safe represents the approved drug combinations while unsafe represents drug pairs from negative set.

Step 2: Making the DDC

In Step one, drug pairs with good safety profile can be predicted. The further step is to identify which pairs should be put together to reach a certain efficacy level for a particular disease while sustaining a low SEs profile. For those drug pairs from Step1 that can be co-prescribed or made into a combination, we only considered those pairs with shared disease indications (Figure 7-1).

We got 977 drug pairs along with at least one shared disease indication for each pair. We used literature co-occurrence as a statistical surrogate for the existence of relationships between the two drugs. 570 pairs (58%) have been supported by at least one publication of clinical trial types when searching both drug name in PubMed while 769 (79%) pairs

have been supported by at least one PubMed publication of any type. 144 pairs (15%) have been supported by at least one clinical trial from clinicaltrial.gov. For the top 20 predictions, we manually checked the clinicaltrial.gov and found 14 predictions with on-going clinical trials are indeed the combinations. The studied conditions in all of the trials agree with our suggested disease indications. Table 7-3 shows the top drug pairs proposed by 'Two Step Rule'.

Table 7-3. Top drug pairs proposed by 'Two Step Rule'.
The Status column includes Approved, indicating that predicted pair has already been approved by FDA, and Predicted, meaning not approved yet; Indication column lists all the disease(s) shared by both components. CT column and the PM column list the number of items in clinicaltrials.gov or PubMed clinical literatures that mention the drug combination.

Component_A	Component_B	Status	Indication	CT	PM
Dexamethasone	Prednisolone	Predicted	Asthma Arthritis, rheumatoid	28	255
Dexamethasone	Hydrocortisone	Predicted	Adrenal insufficiency, primary, congenital	24	473
Formoterol	Fluticasone	Predicted	Asthma	16	55
Abacavir	Zidovudine	Predicted	Infection, HIV/AIDS	9	137
Lamivudine	Emtricitabine	Predicted	Infection, hepatitis-B virus Infection, HIV/AIDS	8	77
Bimatoprost	Latanoprost	Predicted	Glaucoma	8	70
Morphine	Buprenorphine	Predicted	Pain, general Pain, post-operative Pain, musculoskeletal, unspecified	4	215
Pravastatin	Atorvastatin	Predicted	Hypercholesterolaemia	4	176
Dorzolamide	Latanoprost	Predicted	Glaucoma	4	80
Naltrexone	Buprenorphine	Predicted	Addiction, narcotic/opiate	4	29
Lisinopril	Enalapril	Predicted	Hypertension, unspecified Heart failure	3	153
Formoterol	Budesonide	Approved	Asthma Chronic obstructive pulmonary disease Bronchitis, chronic	40	190
Carbidopa	Levodopa	Approved	Parkinson's disease	35	407
Lamivudine	Zidovudine	Approved	Infection, HIV/AIDS	28	587
Timolol	Dorzolamide	Approved	Glaucoma	21	161
Levonorgestrel	Estradiol	Approved	Hormone replacement therapy	19	487
Lamivudine	Abacavir	Approved	Infection, HIV/AIDS	16	195
Ethinyl Estradiol	Levonorgestrel	Approved	Hormone replacement therapy Contraceptive, female	12	371
Tamsulosin	Dutasteride	Approved	Benign prostatic hyperplasia	6	17
Benazepril	Amlodipine	Approved	Hypertension, unspecified	5	51

For those predicted but not approved DDCs shown at the top of Table 7-3, all of them have been co-cited by at least ten clinical type literatures and three clinical trials. We also validated these co-prescriptions in FAERS, which contains millions of reports which record the drugs taken by individual patient and their adverse events. We performed a Fisher's Exact test to identify whether two drugs are significantly recorded together at the same reports. I found that the majority of the novel predicted pairs are more likely to be co-reported or co-prescribed than by chance in FAERS (Table 7-4). All the above validations indicate great promise for further investigations of those novel drug combinations. Table 7-5 also listed the top 10 drug pairs proposed by 'Two Step Rule' but are not in any clinical trials yet, which could be the candidates for developing novel fix-dose combinations.

Table 7-4. Confusion matrix of co-prescription between the five predicted pairs.
TP stands for the number of reports co-mentioned the two drugs; FP/FN stands for the number of reports only mentioned one of them; TN stands for the number of reports mentioned neither of them.

Drug A	Drug B	TP	FP	FN	TN	*p*-value
Abacavir	Zidovudine	374	1143	4208	2220659	0
Bimatoprost	Latanoprost	2	117	804	2225461	2.21E-12
Dexamethasone	Hydrocortisone	113	15529	3244	2207498	1.67E-75
Dexamethasone	Prednisolone	343	15299	26338	2184404	2.86E-30
Dorzolamide	Latanoprost	22	80	784	2225498	0
Formoterol	Fluticasone	1	99	298	2225986	2.68E-05
Lamivudine	Emtricitabine	119	6621	692	2218952	0
Lisinopril	Enalapril	6	25933	1889	2198556	8.49E-04
Morphine	Buprenorphine	15	7754	475	2218140	1.13E-22
Naltrexone	Buprenorphine	0	81	490	2225813	1
Pravastatin	Atorvastatin	24	4434	2650	2219276	4.01E-15

Table 7-5. Top 10 novel drug pairs without any clinical trials reported.

Pair #	Component_A	Component_B	Indication	CT
1	Ethinyl Estradiol	Estrogen	Hormone Replacement Therapy	0
2	Morphine	Marcaine	Pain, Post-Operative Pain, General	0
3	Lovastatin	Atorvastatin	Hyperlipidaemia Alzheimer's Disease Hypercholesterolaemia	0
4	Captopril	Enalapril Maleate	Hypertension, Unspecified Heart Failure	0

5	Lovastatin	Pravastatin	Hypercholesterolaemia	0
6	Budesonide	Beclometasone	Asthma	0
7	Abacavir Sulfate	Zidovudine	Infection, HIV/Aids	0
8	Erythromycin	Clindamycin	Acne	0
9	Fluticasone Propionate	Beclometasone	Asthma Rhinitis, Allergic, Seasonal	0
10	Norethisterone	Medroxyprogesterone	Contraceptive, Female	0

iv. Case studies

Here we looked into the proposed DDCs in Table 7-3.

Dexamethasone/Prednisolone or Dexamethasone/Hydrocortisone (DDC#1 and #2)

Prolonged use of glucocorticoid may impose variety of side effects and impact healthy anabolic processes. Elaboration of glucocorticoid drug combination, particularly with selectively acting glucocorticoid drugs or at reduced dose, could potentially help boost the therapeutic efficacy and prevent unwanted side effects or withdraw effects. This strategy has been explored and shown promises in multiple studies For examples, combination of prednisolone and low dosed dexamethasone is shown to exhibit greater anti-leukemic activity and lower drug resistance than equi-active dose of prednisolone alone (*169*) and the combination therapy using dexamethasone and prednisone has been shown to be more efficacious in patients with idiopathic sudden sensorineural hearing loss than individual glucocorticoid (*170*). As the above drugs are widely prescribed, there is also a great chance of being co-prescribed and, such as the two predicted combinations (i.e. glucocorticoids with different efficacy and potency) from our analysis could warrant further clinical testing in their overlapping indications such as asthma and rheumatoid arthritis.

Abacavir/Zidovudine or Lamivudine/Emtricitabine (DDC #4 and #5)

Combination therapy has been a key therapeutic option in the management of HIV/AIDS. For example, Abacavir and Lamivudine, a top ranked DDC predicted in this study, is one of the FDA approved drug combination. Additionally, our analysis identified several novel combinations that are not yet approved by FDA, for example, Abacavir in combination with Zidovudine, or Lamivudine in combination with Emtricitabine. Like other single antiviral agents, these drugs (Lamivudine, Emtricitabine, Abacavir), when

used on its own, cannot completely suppress viral replication thus allows for drug resistant strains to emerge. The combination of these drugs, however, can potentially impose stronger and more sustained effect than using any single drug alone.

Formoterol/Fluticasone (DDC#3)

Formoterol, a long-acting beta-adrenoceptor agonist, exerts bronchodilatation effect and is used in the management of asthma and chronic obstructive pulmonary disease (COPD). It's already been tested and used in combination with corticosteroids, such as budesonide, to treat or prevent asthma attack and/or respiratory tract inflammation. Fluticasone, another potent glucocorticoid, has been shown to have superior or similar efficacy in improving pulmonary functions in asthma patients (*171, 172*). The predicted Formoterol/Fluticasone combination could be adopted as a new and alternative option in the management of asthma or COPD along the same combination strategy of Formoterol/Budesonide which warrants further validation for its clinical efficacy or safety profile.

Dorzolamide/Lantanoprost (DDC# 9)

Both Dorzolamide and Lantanoprost are anti-glaucoma agents yet with very different MOAs - the former is a carbonic anhydrase inhibitor that exerts pharmacological function by decreasing the production of aqueous humour, yet the latter agent is a prostaglandin analogue that increases outflow of aqueous fluid. With such distinctive and hypothetically complement mechanisms, the drug combination of these two agents could potentially exert stronger efficacy in reducing intraocular pressure particularly in severe glaucoma patients.

The above two case studies (i.e. Pair # 3, #9) are the combination of agents from different categories with distinct MOAs. They could have additional and/or greater pharmacological and clinical benefits with their efficacy synergy potential, pill burden reduction, and improved compliance in patient care.

7.4 Conclusions

Evaluation of drug pair safety is a critical issue for co-prescription or making fix-dose combinations (*173, 174*). Methods have been developed to predict drug-drug interactions (DDIs) from text mining (*175, 176*), network modeling (*177*), high-throughput screening (*149*), or computational data integration (*153*). Our approach exploring the possibility of predicting new drug pairs by representing drug combinations with their clinical side effects. It is based on the hypothesis that the drugs that can be put together usually do not have overlapping serious adverse drug reactions. The key advantage of using clinical side effect information lies in that it is direct observations from human subjects with fewer translational issues compared with data from in vitro or animal studies. The "signature" set of three side effects identified from our analysis provides a practical guideline to help rule out unsafe co-prescriptions.

Using the integrated side effect data sources, we examined the effects of different machine learning methods on the prediction performance. For the prediction performance evaluation of 198 independent drug combinations with the features of the three side effects, decision tree model gives an AUC of 0.80, Naive Bayes with an AUC of 0.84 and Logistic Regression with an AUC of 0.84. The robust performance across different machine learning methods confirms that our conclusion is not biased towards a particular method. We chose the decision tree model with the aim for easy clinical implementation despite that it doesn't give the highest AUC.

One limitation of side effect based on approaches to study DDCs or DDIs is that no good resource except for TWOSIDES is currently available to capture side effects of drug pairs. In our work of deriving the three side effects as a FDA blacklist based on 65 signature side effects from TWOSIDES, we relied on the assumption that if we don't want the drug combination to have any of the three side effects, we require that neither of the drug components have such side effect. It is possible that even drug component itself doesn't have such side effect, the combination may have it due to the potential drug interaction. This may undermine our classification performance. Nevertheless, our results

demonstrate that safe drug pairs usually do not have overlap in these three serious adverse drug reactions.

The prediction performance of the three side effects is unlikely to be due to the bias in drug combinations' disease profiles. The 41 drug combinations, where the signature of the three side effects was derived from, covered 24 diseases. We used an independent dataset of 198 drug combinations to measure their prediction performance. For this independent dataset, additional 68 diseases were covered. In other words, the disease profiles for these two datasets are different and this minimizes the bias during the prediction. On the other hand we need to be cautious to extrapolate the prediction model to apply to those drugs or diseases never shown up in our dataset since the scope of the prediction performance may be limited to these 245 drugs and 92 diseases.

As discovered in the previous study (*152*), some side effects are associated with the indications of the drugs. For example, Actoplus Met is a fix-dose combination of metformin hydrochloride and pioglitazone hydrochloride. The two drug components share SEs of Anaemia; similarly, for another diabetes drug Duetact, a combination of Glimepiride and Pioglitazone Hydrochloride, these two drug components also share the SE of Anaemia. We hypothesize that for those side effects shared by approved drug combinations, they may be essential for the therapeutic effect of the drugs and they are usually not severe SEs. For example, the pharmacological effect anaemia is associated with reduced insulin consumption, which may alleviate the reliance on insulin of certain insulin resistant diabetes patients.

Dosing is another factor that has to be considered when co-prescribing drugs in the clinical practice. Here we propose a simplified model for discussion. When the concentration becomes lower, e.g., halving the dose for each component when making the DDC, the dose-related toxicity of this combination may thus be halved. However, since we require that drug components should have the same indication in our model, the efficacy may theoretically remain the same or even better due to the synergistic effect of the combination. We understand the real situation in the dosing issue is much more

complex, however. This is only an ideal model that inspires further discussion and deeper understand of the making of DDCs. Further clinical trials are needed to validate its efficacy on a particular dosing. Besides by choosing the right drug pairs, e.g., one expensive drug along with a cheaper one, with reduced doses, it may also bring economics of combining the drugs.

We suggest that our predictions may be beneficial in three areas: (i) improving the safety profiles of drug co-prescriptions in clinic; (ii) assessing potentially hazardous drug combinations in early stage of the fix-dose combination discovery in pharmaceutical industry; and (iii) potentially reducing pill burden or bringing economics of combining the drugs. While our predictions were validated *in-silico*, they should be further tested experimentally to establish their clinical implications.

Chapter 8. Conclusions

8.1 Research summary and contributions

In summary, the thesis presents a comprehensive framework of computational drug discovery, using system approaches. The thesis mainly consists of two parts: disease biomarker identification and disease treatment discoveries.

I start by introducing the research in biomarker identification for human diseases in the post-genomic era with an emphasis in system biology approaches such as using the protein interaction networks. Diagnostic biomarker is expected to detect a given type of disease in an individual with both high sensitivity and specificity; predictive biomarker serves to predict drug response before treatment is started. Both are essential before we even start seeking any treatment for the patients. In Chapter 2, I studied how the coverage of the disease genes, the protein interaction quality, and gene ranking strategies can affect the identification of disease genes; In Chapter 3, I addressed the challenge of constructing a central database to collect the system level data such as protein interaction, pathway, etc. for the biomarker discovery at the system biology level. In Chapter 4, I built case studies for biomarker identification for Diabetes by using the conclusions from Chapter 2 and 3.
The second part of the thesis mainly addresses how to find treatments after disease identification. I specifically focus on computational drug repositioning due to its low cost, few translational issues and other benefits. In Chapter 5, I described how to implement literature mining approaches to build the disease-protein-drug connectivity map and demonstrated its superior performances compared to other existing applications. In Chapter 6, I presented a valuable drug-protein directionality database which filled the research gap of lacking alternatives for the experimental CMAP in computational drug discovery field. The correlation based ranking algorithm was also extended to include the underlying topology among proteins. Chapter 5 and 6 conclude the thesis work of drug repositioning in the genomic level. In Chapter 7, I demonstrated how to study drug repositioning beyond genomic level and from one dimension to two dimensions. In specific I explored how to propose drug combination with clinical side effects as prediction features.

8.2 Future research directions

The computational workflow for drug discovery in genomics level can be generally represented as in Figure 8-1. The future research directions can be extended from each step, especially step 1, 3 and 4.

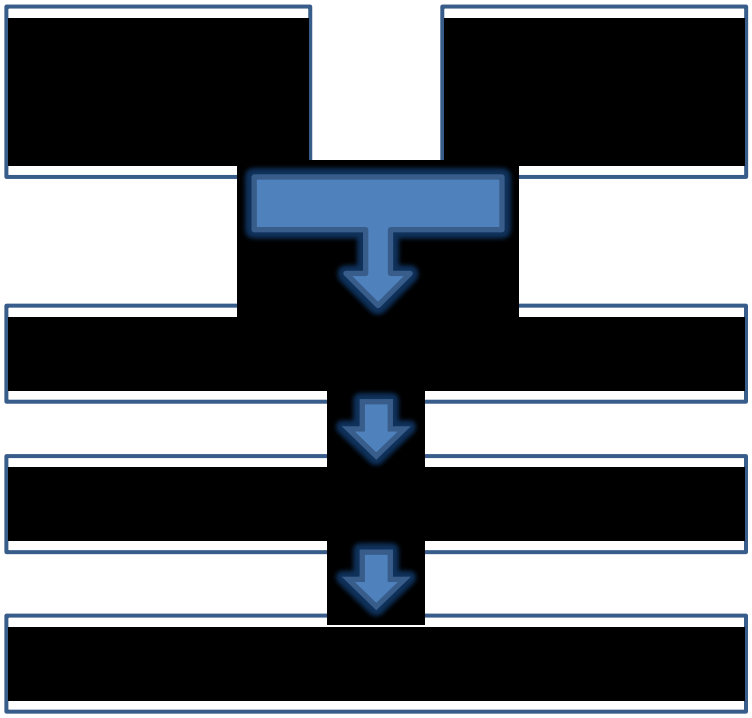

Figure 8-1.The general workflow of computational drug discovery.

i. *Research in identifying reliable disease biomarker*

How to get a consistent panel of biomarker for each disease regardless of platforms and other confounding factors is a central topic in biomarker discovery (Step 1 in Figure 8-1). The emerging of NGS techniques, specially the RNA-Seq, may be good alternatives for genomic profiling compared to microarray techniques. It has been shown to be more reliable and accurate measurement of gene expression level with the NGS techniques. The methodology of biomarker discovery in this thesis can be readily applied to the expression level from RNA-Seq experiments.

Future investigators should ideally build a probabilistic-based catalog of biomarkers (Step 1 in Figure 8-1) for each disease condition just as how DMAP are built for recording the relationships between drugs and proteins. This central database will not rely on single detection technology and should serves as a knowledge base for any disease biomarker study. Once we have a central database for disease and protein relationship, we could combine it with DMAP with the framework presented in the thesis to propose more reliable treatments.

ii. Research in disease model discovery

While the conventional "One disease, One gene, and One drug" paradigm works effectively for simple genetic disorders, it fails to produce effective drugs for complex diseases such as cancer. In complex diseases, many genes may be contributing to the disease's phenotype. Thus building a disease model (Step 3 in Figure 8-1) to explain the underlying disease mechanism is essential for developing effective disease treatments. In the thesis, I focused on breast cancer disease model for proof of concept. More disease models should be built and tested to check the robustness of the computational framework. The cost of utilizing the pathway information is that one has to put more effort in data collection and pathway construction. Ultimately a database containing disease models for major diseases is desirable. Such a disease-oriented database will provide much better resolution than traditional protein interaction databases for computational drug discovery.

iii. Research in disease ranking algorithms

In the thesis, PETS algorithm was proposed to utilize the underlying topology of the disease model to rank the potential treatments. It has shown a superior performance in breast cancer study. Future investigators need test this algorithm in more disease models. Ideally one could provide an integrative tool for wet-lab scientists to use for drug discovery by integrating the disease biomarker catalog, DMAP, and disease model database.

iv. Beyond genomics

Matching the disease expression profile with the drug perturbation expression profile is a mainstream approach in the computational drug discovery nowadays. Despite its popularity, few researchers or companies have yet proposed any drugs for FDA approval

based on this approach. The major challenge here is how to translate the discovery in molecular level to phenotypic level. We have at least two ways to address the translational issues.

One way is to build the prediction model with clinical features such as drug side effects. Since side effects are directly observed from human, the translation issues will be less of a concern. In Chapter 7, I have shown how to predict drug combination using side effects as features. Future investigators need continue to pay more attention to side effects and other measurements in clinical trials studies and utilize them to build computational models.

Another way is to utilize the Electronic medical record (EMR), a system that contains all of a patient's medical history from one practice. It contains detailed information about patients' phenotypic responses for various drugs. As the first step moving towards this direction, I built a statistical model to use FAERS to identify what drug combination is more inclined to be prescribed by doctors in Chapter 7. Utilizing the EMR can not only serve as the validation purpose of any proposed treatments, but also help uncover any off-label use by the clinicians. Models based on those data can leverage the knowledge accumulated by clinician's daily practice and will be appreciated.

With the continued research and development in computational drug discovery, and with incoming of the big data in bioinformatics, the drug discovery, especially drug repositioning, tailored to the individual patient will be realized.

References

1. J. A. Squire, TMPRSS2-ERG and PTEN loss in prostate cancer. *Nat Genet* **41**, 509 (May, 2009).
2. C. Auffray, Protein subnetwork markers improve prediction of cancer outcome. *Molecular Systems Biology* **3**, 141 (2007).
3. G. P. Nolan, What's wrong with drug screening today. *Nature Chemical Biology* **3**, 187 (2007).
4. H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, T. Ideker, Network-based classification of breast cancer metastasis. *Molecular Systems Biology* **3**, 140 (2007).
5. M. A. Pujana *et al.*, Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nature Genetics* **39**, 1338 (2007).
6. H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, T. Ideker, Network-based classification of breast cancer metastasis. *Mol Syst Biol* **3**, 140 (2007).
7. N. McCarthy, Tumour profiling: Networking, protein style. *Nature Reviews Cancer* **7**, 892 (2007).
8. R. L. Grubb Iii *et al.*, Pathway Biomarker Profiling of Localized and Metastatic Human Prostate Cancer Reveal Metastatic and Prognostic Signatures. *J. Proteome Res* **8**, 3044 (2009).
9. J. D. Durrant *et al.*, A multidimensional strategy to detect polypharmacological targets in the absence of structural and sequence homology. *PLoS computational biology* **6**, e1000648 (Jan, 2010).
10. M. A. Yildirim, K. I. Goh, M. E. Cusick, A. L. Barabasi, M. Vidal, Drug-target network. *Nat Biotechnol* **25**, 1119 (Oct, 2007).
11. A. L. Hopkins, Drug discovery: Predicting promiscuity. *Nature* **462**, 167 (Nov 12, 2009).
12. A. L. Hopkins, Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology* **4**, 682 (Nov, 2008).
13. A. L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* **12**, 56 (2011).
14. D. K. Arrell, A. Terzic, Network systems biology for drug discovery. *Clinical pharmacology and therapeutics* **88**, 120 (Jul, 2010).
15. J. Lamb *et al.*, The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929 (Sep 29, 2006).
16. A. Beyer, S. Bandyopadhyay, T. Ideker, Integrating physical and genetic maps: from genomes to interaction networks. *Nat Rev Genet* **8**, 699 (Sep, 2007).
17. M. Sirota *et al.*, Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science translational medicine* **3**, 96ra77 (Aug 17, 2011).
18. M. Y. Lan *et al.*, From NPC therapeutic target identification to potential treatment strategy. *Molecular cancer therapeutics* **9**, 2511 (Sep, 2010).
19. G. Jin *et al.*, A novel method of transcriptional response analysis to facilitate drug repositioning for cancer therapy. *Cancer research* **72**, 33 (Jan 1, 2012).
20. R. H. Shoemaker, The NCI60 human tumour cell line anticancer drug screen. *Nature reviews. Cancer* **6**, 813 (Oct, 2006).

21. J. Barretina *et al.*, The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603 (Mar 29, 2012).
22. A. Gottlieb, G. Y. Stein, E. Ruppin, R. Sharan, PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology* **7**, 496 (2011).
23. J. Li, X. Zhu, J. Y. Chen, Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. *PLoS computational biology* **5**, e1000450 (Jul, 2009).
24. D. Hristovski, C. Friedman, T. C. Rindflesch, B. Peterlin, Exploiting semantic relations for literature-based discovery. *AMIA Annu Symp Proc*, 349 (2006).
25. C. G. Begley, L. M. Ellis, Drug development: Raise standards for preclinical cancer research. *Nature* **483**, 531 (Mar 29, 2012).
26. B. S. Hendriks, Functional pathway pharmacology: chemical tools, pathway knowledge and mechanistic model-based interpretation of experimental data. *Current opinion in chemical biology* **14**, 489 (Aug, 2010).
27. P. Khatri, M. Sirota, A. J. Butte, Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology* **8**, e1002375 (Feb, 2012).
28. S. R. Chowbina *et al.*, HPD: an online integrated human pathway database enabling systems biology studies. *BMC Bioinformatics* **10 Suppl 11**, S5 (2009).
29. M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, M. Hirakawa, KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research* **38**, D355 (Jan, 2010).
30. http://www.pathguide.org/
31. C. Knox *et al.*, DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research* **39**, D1035 (Jan, 2011).
32. M. Kuhn *et al.*, STITCH 3: zooming in on protein-chemical interactions. *Nucleic Acids Res* **40**, D876 (Jan, 2012).
33. E. M. McDonagh, M. Whirl-Carrillo, Y. Garten, R. B. Altman, T. E. Klein, From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomarkers in medicine* **5**, 795 (Dec, 2011).
34. Y. Y. Li, S. J. Jones, Drug repositioning for personalized medicine. *Genome medicine* **4**, 27 (Mar 30, 2012).
35. C. P. Adams, V. V. Brantner, Estimating the cost of new drug development: is it really 802 million dollars? *Health Aff (Millwood)* **25**, 420 (Mar-Apr, 2006).
36. I. Kola, J. Landis, Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov* **3**, 711 (Aug, 2004).
37. M. S. Boguski, K. D. Mandl, V. P. Sukhatme, Drug discovery. Repurposing with a difference. *Science* **324**, 1394 (Jun 12, 2009).
38. Y. A. Lussier, J. L. Chen, The emergence of genome-based drug repositioning. *Science translational medicine* **3**, 96ps35 (Aug 17, 2011).
39. R. L. Schilsky, Personalized medicine in oncology: the future is now. *Nature reviews. Drug discovery* **9**, 363 (May, 2010).

40. F. Iorio *et al.*, Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 14621 (Aug 17, 2010).
41. J. A. DiMasi, R. W. Hansen, H. G. Grabowski, The price of innovation: new estimates of drug development costs. *Journal of health economics* **22**, 151 (Mar, 2003).
42. C. R. Chong, D. J. Sullivan, Jr., New uses for old drugs. *Nature* **448**, 645 (Aug 9, 2007).
43. T. T. Ashburn, K. B. Thor, Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews. Drug discovery* **3**, 673 (Aug, 2004).
44. J. K. Aronson, Old drugs--new uses. *British journal of clinical pharmacology* **64**, 563 (Nov, 2007).
45. M. J. Keiser *et al.*, Predicting new molecular targets for known drugs. *Nature* **462**, 175 (Nov 12, 2009).
46. M. Campillos, M. Kuhn, A. C. Gavin, L. J. Jensen, P. Bork, Drug target identification using side-effect similarity. *Science* **321**, 263 (Jul 11, 2008).
47. G. Hu, P. Agarwal, Human disease-drug network based on genomic expression profiles. *PloS one* **4**, e6536 (2009).
48. J. Lamb, The Connectivity Map: a new tool for biomedical research. *Nature reviews. Cancer* **7**, 54 (Jan, 2007).
49. H. Huang, J. Li, J. Y. Chen, Disease gene-fishing in molecular interaction networks: a case study in colorectal cancer. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference* **2009**, 6416 (2009).
50. H. Huang *et al.*, PAGED: a pathway and gene-set enrichment database to enable molecular phenotype discoveries. *BMC Bioinformatics* **13 Suppl 15**, S2 (2012).
51. X. Wu *et al.*, Network expansion and pathway enrichment analysis towards biologically significant findings from microarrays. *Journal of integrative bioinformatics* **9**, 213 (2012).
52. H. Huang *et al.*, C(2)Maps: a network pharmacology database with comprehensive disease-gene-drug connectivity relationships. *BMC Genomics* **13 Suppl 6**, S17 (2012).
53. Hui Huang, Sara Ibrahim, Thanh Nguyen, Jake Y. Chen DPP modeling in breast cancer demonstrates a new path for drug repositioning (in preparation)
54. Hui Huang, Xiaoyan A.Angela Qu, Lun Yang Drug Combination Prediction Only Based on Three Clinical Side-Effects. PLoS Comput Biol (submitted)
55. A. C. Ahn, M. Tewari, C. S. Poon, R. S. Phillips, The clinical applications of a systems approach. *PLoS Med* **3**, e209 (Jul, 2006).
56. A. N. Smith *et al.*, Mutations in ATP6N1B, encoding a new kidney vacuolar proton pump 116-kD subunit, cause recessive distal renal tubular acidosis with preserved hearing. *Nat Genet* **26**, 71 (Sep, 2000).
57. C. Giallourakis, C. Henson, M. Reich, X. Xie, V. K. Mootha, Disease gene discovery through integrative genomics. *Annu Rev Genomics Hum Genet* **6**, 381 (2005).

58. J. Chen, H. Xu, B. J. Aronow, A. G. Jegga, Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics* **8**, 392 (2007).
59. J. L. Morrison, R. Breitling, D. J. Higham, D. R. Gilbert, GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics* **6**, 233 (2005).
60. J. Y. Chen, C. Shen, A. Y. Sivachenko, Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Pac Symp Biocomput*, 367 (2006).
61. A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, V. A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**, D514 (2005).
62. M. Kanehisa *et al.*, KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**, D480 (Jan, 2008).
63. D. R. Rhodes *et al.*, Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* **9**, 166 (Feb, 2007).
64. C. J. Mattingly, M. C. Rosenstein, G. T. Colby, J. N. Forrest, J. L. Boyer, The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. *J Exp Zoolog A Comp Exp Biol* **305**, 689 (2006).
65. J. Y. Chen, S. Mamidipalli, T. Huang, HAPPI: an Online Database of Comprehensive Human Annotated and Predicted Protein Interactions. *BMC Genomics (Accepted)*, (2009).
66. P. Holmans, Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. *Advances in genetics* **72**, 141 (2010).
67. V. K. Ramanan, L. Shen, J. H. Moore, A. J. Saykin, Pathways analysis of genomic data: concepts, methods, and prospoects for future development. *Trends in Genetics* **Accepted**, (2012).
68. D. K. Slonim, From patterns to pathways: gene expression data analysis comes of age. *nature genetics* **32**, 502 (2002).
69. L. Abatangelo *et al.*, Comparative study of gene set enrichment methods. *Bmc Bioinformatics* **10**, 275 (2009).
70. A. Subramanian *et al.*, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545 (2005).
71. A. C. Culhane *et al.*, GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Res* **40**, D1060 (Jan, 2012).
72. A. Liberzon *et al.*, Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739 (Jun 15, 2011).
73. H. Eleftherohorinou *et al.*, Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. *PLoS One* **4**, e8068 (2009).
74. K. Wang, M. Li, M. Bucan, Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics* **81**, 1278 (2007).
75. H. Zhong, X. Yang, L. M. Kaplan, C. Molony, E. E. Schadt, Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *The American Journal of Human Genetics* **86**, 581 (2010).

76. S. Chowbina *et al.*, A new approach to construct pathway connected networks and its application in dose responsive gene expression profiles of rat liver regulated by 2, 4DNT. *BMC Genomics* **11**, S4 (2010).
77. A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, V. A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research* **33**, D514 (Jan 1, 2005).
78. K. G. Becker, K. C. Barnes, T. J. Bright, S. A. Wang, The genetic association database. *Nat Genet* **36**, 431 (May, 2004).
79. F. Xiao *et al.*, miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res* **37**, D105 (Jan, 2009).
80. J. Y. Chen, S. Mamidipalli, T. Huan, HAPPI: an online database of comprehensive human annotated and predicted protein interactions. *BMC Genomics* **10 Suppl 1**, S16 (2009).
81. T. Pang-Ning, M. Steinbach, V. Kumar. (Boston: Person Addison Wesley EducatioPress, 2005).
82. R. Jothi, T. M. Przytycka, L. Aravind, Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. *BMC Bioinformatics* **8**, 173 (2007).
83. R. Edgar, M. Domrachev, A. E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207 (2002).
84. S. Segditsas, I. Tomlinson, Colorectal cancer and genetic alterations in the Wnt pathway. *Oncogene* **25**, 7531 (Dec 4, 2006).
85. Madhankumar Sonachalam, Jeffrey Shen, Hui Huang and Xiaogang Wu (2012) Systems biology approach to identify gene network signatures for colorectal cancer. Frontiers in Genetics, Vol. 3, pp. 80
86. M. E. Smoot, K. Ono, J. Ruscheinski, P. L. Wang, T. Ideker, Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431 (Feb 1, 2011).
87. D. B. Allison, X. Cui, G. P. Page, M. Sabripour, Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics* **7**, 55 (2006).
88. M. Reimers, Making Informed Choices about Microarray Data Analysis. *PLoS Computational Biology* **6**, e1000786 (2010).
89. D. K. Slonim, I. Yanai, Getting started in gene expression microarray analysis. *PLoS Computational Biology* **5**, e1000543 (2009).
90. A. Subramanian *et al.*, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545 (2005).
91. D. Glez-Pena, G. Gomez-Lopez, D. G. Pisano, F. Fdez-Riverola, WhichGenes: a web-based tool for gathering, building, storing and exporting gene sets with application in gene set enrichment analysis. *Nucleic Acids Research* **37**, W329 (2009).
92. A. L. Barabasi, Z. N. Oltvai, Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* **5**, 101 (2004).
93. P. Goymer, Cancer genetics: Networks uncover new cancer susceptibility suspect. *Nature Reviews Genetics* **8**, 823 (2007).

94. M. A. Pujana *et al.*, Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nature genetics* **39**, 1338 (2007).
95. H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, T. Ideker, Network-based classification of breast cancer metastasis. *Molecular Systems Biology* **3**, 140 (2007).
96. J. Y. Chen, C. Shen, A. Y. Sivachenko. (2006), vol. 11, pp. 367-378.
97. A. C. Alibegovic *et al.*, Insulin resistance induced by physical inactivity is associated with multiple transcriptional changes in skeletal muscle in young men. *American Journal of Physiology-Endocrinology And Metabolism* **299**, E752 (2010).
98. G. Smyth, Limma: linear models for microarray data. *Bioinformatics and computational biology solutions using R and Bioconductor*, 397 (2005).
99. J. Mestres, Computational chemogenomics approaches to systematic knowledge-based drug discovery. *Curr Opin Drug Discov Devel* **7**, 304 (May, 2004).
100. http://www.pathguide.org/
101. C. Knox *et al.*, DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic acids research* **39**, D1035 (2011).
102. M. Kuhn *et al.*, STITCH 3: zooming in on protein–chemical interactions. *Nucleic acids research* **40**, D876 (2012).
103. F. Zhu *et al.*, Update of TTD: therapeutic target database. *Nucleic acids research* **38**, D787 (2010).
104. A. P. Davis *et al.*, Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks. *Nucleic acids research* **37**, D786 (2009).
105. T. Liu, Y. Lin, X. Wen, R. N. Jorissen, M. K. Gilson, BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research* **35**, D198 (2007).
106. A. L. Hopkins, Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology* **4**, 682 (2008).
107. A. Bugrim, T. Nikolskaya, Y. Nikolsky, Early prediction of drug metabolism and toxicity: systems biology approach and modeling. *Drug Discov Today* **9**, 127 (Feb 1, 2004).
108. J. Lamb *et al.*, The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929 (2006).
109. J. Li, X. Zhu, J. Y. Chen, Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. *PLoS computational biology* **5**, e1000450 (2009).
110. A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, V. A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research* **33**, D514 (2005).
111. Z. Shi, H. Li, Application of artificial neural network approach and remotely sensed imagery for regional eco-environmental quality evaluation. *Environ Monit Assess* **128**, 217 (May, 2007).
112. M. Sakamoto *et al.*, Analysis of gene expression profiles associated with cisplatin resistance in human ovarian cancer cell lines and tissues using cDNA microarray. *Hum Cell* **14**, 305 (Dec, 2001).

113. C. H. Wu *et al.*, The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic acids research* **34**, D187 (2006).
114. K. L. Blackwell *et al.*, Tamoxifen inhibits angiogenesis in estrogen receptor-negative animal models. *Clinical Cancer Research* **6**, 4359 (2000).
115. D. Bolchini, A. Finkelstein, V. Perrone, S. Nagl, Better bioinformatics through usability analysis. *Bioinformatics* **25**, 406 (February 1, 2009, 2009).
116. T. Sorlie *et al.*, Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10869 (Sep 11, 2001).
117. R. Huang *et al.*, The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Science translational medicine* **3**, 80ps16 (Apr 27, 2011).
118. F. Zhu *et al.*, Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic acids research* **40**, D1128 (Jan, 2012).
119. C. Pacini *et al.*, DvD: An R/Cytoscape pipeline for drug repurposing using public repositories of gene expression data. *Bioinformatics* **29**, 132 (Jan 1, 2013).
120. T. Barrett *et al.*, NCBI GEO: archive for functional genomics data sets--update. *Nucleic acids research* **41**, D991 (Jan 1, 2013).
121. R. A. Irizarry *et al.*, Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research* **31**, e15 (Feb 15, 2003).
122. P. Warnat, R. Eils, B. Brors, Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC bioinformatics* **6**, 265 (2005).
123. A. Yuryev *et al.*, Automatic pathway building in biological association networks. *BMC bioinformatics* **7**, 171 (2006).
124. C. Winter *et al.*, Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS computational biology* **8**, e1002511 (2012).
125. S. Kohler, S. Bauer, D. Horn, P. N. Robinson, Walking the interactome for prioritization of candidate disease genes. *American journal of human genetics* **82**, 949 (Apr, 2008).
126. A. P. Chiang, A. J. Butte, Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clinical pharmacology and therapeutics* **86**, 507 (Nov, 2009).
127. S. D. Walter, The partial area under the summary ROC curve. *Statistics in medicine* **24**, 2025 (Jul 15, 2005).
128. Donna L. Hoyert, and Jiaquan Xu, Deaths: Preliminary Data for 2011.National Vital Statistics Reports Volume 61, Number 6(October 10, 2012)
129. P. Indolfi *et al.*, Synchronous bilateral Wilms tumor: A report from the Associazione Italiana Ematologia Oncologia Pediatrica (AIEOP). *Cancer*, (Jan 10, 2013).
130. J. S. Do *et al.*, Therapeutic target validation of protein kinase C(PKC)-zeta for asthma using a mouse model. *International journal of molecular medicine* **23**, 561 (Apr, 2009).
131. D. Cheng *et al.*, The effects of protein kinase C (PKC) on the tension of normal and passively sensitized human airway smooth muscle and the activity of voltage-

dependent delayed rectifier potassium channel (Kv). *Journal of Huazhong University of Science and Technology. Medical sciences = Hua zhong ke ji da xue xue bao. Yi xue Ying De wen ban = Huazhong keji daxue xuebao. Yixue Yingdewen ban* **27**, 153 (Apr, 2007).
132. P. J. Barnes, N. M. Wilson, M. J. Brown, A calcium antagonist, nifedipine, modifies exercise-induced asthma. *Thorax* **36**, 726 (Oct, 1981).
133. S. D. Groshong *et al.*, Biphasic regulation of breast cancer cell growth by progesterone: role of the cyclin-dependent kinase inhibitors, p21 and p27(Kip1). *Molecular endocrinology* **11**, 1593 (Oct, 1997).
134. Tamoxifen Pathway in Homo sapiens. http://pathman.smpdb.ca/pathways/SMP00471/pathway?reset=true&highlight%255bDB00675%255d=true
135. E. N. Emmanouil-Nikoloussi *et al.*, Breast tumor developed in a pregnant rat after treatment with the teratogen Cycloheximide. *Hippokratia* **14**, 136 (Apr, 2010).
136. G. N. Farhat *et al.*, Sex hormone levels and risk of breast cancer with estrogen plus progestin. *Journal of the National Cancer Institute* **105**, 1496 (Oct 2, 2013).
137. H. Endogenous *et al.*, Sex hormones and risk of breast cancer in premenopausal women: a collaborative reanalysis of individual participant data from seven prospective studies. *The lancet oncology* **14**, 1009 (Sep, 2013).
138. http://www.breastcancer.org/treatment/druglist/mitomycin
139. G. Early Breast Cancer Trialists' Collaborative *et al.*, Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. *Lancet* **378**, 771 (Aug 27, 2011).
140. N. Platet *et al.*, Breast cancer cell invasiveness: correlation with protein kinase C activity and differential regulation by phorbol ester in estrogen receptor-positive and -negative cells. *International journal of cancer. Journal international du cancer* **75**, 750 (Mar 2, 1998).
141. A. Pujol, R. Mosca, J. Farres, P. Aloy, Unveiling the role of network and systems biology in drug discovery. *Trends in pharmacological sciences* **31**, 115 (Mar, 2010).
142. H. Kitano, A robustness-based approach to systems-oriented drug design. *Nature reviews. Drug discovery* **6**, 202 (Mar, 2007).
143. G. R. Zimmermann, J. Lehar, C. T. Keith, Multi-target therapeutics: when the whole is greater than the sum of the parts. *Drug discovery today* **12**, 34 (Jan, 2007).
144. T. C. Chou, Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies. *Pharmacological reviews* **58**, 621 (Sep, 2006).
145. M. A. Yildirim, K. I. Goh, M. E. Cusick, A. L. Barabasi, M. Vidal, Drug-target network. *Nature biotechnology* **25**, 1119 (Oct, 2007).
146. K. S. Smalley *et al.*, Multiple signaling pathways must be targeted to overcome drug resistance in cell lines derived from melanoma metastases. *Molecular cancer therapeutics* **5**, 1136 (May, 2006).
147. Y. Pilpel, P. Sudarsanam, G. M. Church, Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature genetics* **29**, 153 (Oct, 2001).

148. N. V. Sergina *et al.*, Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3. *Nature* **445**, 437 (Jan 25, 2007).
149. A. A. Borisy *et al.*, Systematic discovery of multicomponent therapeutics. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 7977 (Jun 24, 2003).
150. P. K. Wong *et al.*, Closed-loop control of cellular functions using combinatory drugs guided by a stochastic search algorithm. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 5105 (Apr 1, 2008).
151. T. C. Chou, Drug combination studies and their synergy quantification using the Chou-Talalay method. *Cancer research* **70**, 440 (Jan 15, 2010).
152. L. Yang *et al.*, Identifying unexpected therapeutic targets via chemical-protein interactome. *PloS one* **5**, e9568 (2010).
153. X. M. Zhao *et al.*, Prediction of drug combinations by integrating molecular and pharmacological data. *PLoS computational biology* **7**, e1002323 (Dec, 2011).
154. Y. Y. Wang, K. J. Xu, J. Song, X. M. Zhao, Exploring drug combinations in genetic interaction network. *BMC bioinformatics* **13 Suppl 7**, S7 (2012).
155. M. Duran-Frigola, P. Aloy, Recycling side-effects into clinical markers for drug repositioning. *Genome medicine* **4**, 3 (2012).
156. L. Yang, P. Agarwal, Systematic drug repositioning based on clinical side-effects. *PloS one* **6**, e28025 (2011).
157. N. P. Tatonetti, P. P. Ye, R. Daneshjou, R. B. Altman, Data-driven prediction of drug effects and interactions. *Science translational medicine* **4**, 125ra31 (Mar 14, 2012).
158. Z. Liu *et al.*, Translating clinical findings into knowledge in drug safety evaluation--drug induced liver injury prediction system (DILIps). *PLoS computational biology* **7**, e1002310 (Dec, 2011).
159. M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, P. Bork, A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology* **6**, 343 (2010).
160. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1
161. Y. Liu, B. Hu, C. Fu, X. Chen, DCDB: drug combination database. *Bioinformatics* **26**, 587 (Feb 15, 2010).
162. Jie Cheng, Qing Xie, Vinod Kumar, Mark Hurle, Johannes M. Freudenberg, Lun Yang, Pankaj Agarwal, Evaluation of Analytical Methods for Connectivity Map Data. Pacific Symposium on Biocomputing 18:5-16(2013)
163. A. Gottlieb, G. Y. Stein, Y. Oron, E. Ruppin, R. Sharan, INDI: a computational framework for inferring drug interactions and their associated recommendations. *Molecular systems biology* **8**, 592 (2012).
164. J. R. Quinlan, *C4.5 : programs for machine learning*. The Morgan Kaufmann series in machine learning (Morgan Kaufmann Publishers, San Mateo, Calif., 1993), pp. x, 302 p.
165. G. D. Bader, C. W. Hogue, An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics* **4**, 2 (Jan 13, 2003).

166. G. Cevc, G. Blume, Hydrocortisone and dexamethasone in very deformable drug carriers have increased biological potency, prolonged effect, and reduced therapeutic dosage. *Biochimica et biophysica acta* **1663**, 61 (May 27, 2004).
167. A. R. Webb, S. Leong, P. S. Myles, S. J. Burn, The addition of a tramadol infusion to morphine patient-controlled analgesia after abdominal surgery: a double-blinded, placebo-controlled randomized trial. *Anesthesia and analgesia* **95**, 1713 (Dec, 2002).
168. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.
169. M. Spenerova *et al.*, Combination of prednisolone and low dosed dexamethasone exhibits greater in vitro antileukemic activity than equiactive dose of prednisolone and overcomes prednisolone drug resistance in acute childhood lymphoblastic leukemia. *Biomedical papers of the Medical Faculty of the University Palacky, Olomouc, Czechoslovakia*, (Oct 31, 2012).
170. A. Battaglia, R. Burchette, R. Cueva, Combination therapy (intratympanic dexamethasone + high-dose prednisone taper) for the treatment of idiopathic sudden sensorineural hearing loss. *Otology & neurotology : official publication of the American Otological Society, American Neurotology Society [and] European Academy of Otology and Neurotology* **29**, 453 (Jun, 2008).
171. E. Derom, J. Van Schoor, W. Verhaeghe, W. Vincken, R. Pauwels, Systemic effects of inhaled fluticasone propionate and budesonide in adult patients with asthma. *American journal of respiratory and critical care medicine* **160**, 157 (Jul, 1999).
172. N. Adams, T. J. Lasserson, C. J. Cates, P. W. Jones, Fluticasone versus beclomethasone or budesonide for chronic asthma in adults and children. *Cochrane Database Syst Rev*, CD002310 (2007).
173. M. Pirmohamed, Drug-drug interactions and adverse drug reactions: separating the wheat from the chaff. *Wiener klinische Wochenschrift* **122**, 62 (Feb, 2010).
174. F. Montastruc *et al.*, The importance of drug-drug interactions as a cause of adverse drug reactions: a pharmacovigilance study of serotoninergic reuptake inhibitors in France. *European journal of clinical pharmacology* **68**, 767 (May, 2012).
175. J. D. Duke *et al.*, Literature based drug interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions. *PLoS computational biology* **8**, e1002614 (Aug, 2012).
176. B. Percha, Y. Garten, R. B. Altman, Discovery and explanation of drug-drug interactions via text mining. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 410 (2012).
177. M. Takarabe, D. Shigemizu, M. Kotera, S. Goto, M. Kanehisa, Network-based analysis and characterization of adverse drug-drug interactions. *Journal of chemical information and modeling* **51**, 2977 (Nov 28, 2011).

Curriculum Vitae

Hui Huang

EDUCATION

2008-2014 Indiana University, PhD in Bioinformatics, minor in Biostatistics
Five year data mining research experiences in computational drug discovery
Dissertation: Systematic biology modeling: the insights for computational drug discovery
2004-2008 University of Science and Technology of China, BS in Bioinformatics

RESEARCH EXPERIENCES

2008-2013 Indiana Center for System Biology and Personalized Medicine

- Developed computational algorithms to evaluated drug efficacy with the integrated pathway model and validated results with clinical trial evidence
- Built a comprehensive gene-set database including 10 pathway/gene signature databases.; Constructed disease specific pathway association network; Built the web interface
- Utilized computational connectivity map to study the association between cancer drugs and proteins and applied it for cancer drug repurpose
- Analyzed large scale microarray dataset for computational biomarker discovery in diabetes
- Identified disease genes with protein interaction network and computational algorithms

May 2012-August 2012 GlaxoSmithKline Computational Biology Group

- Constructed drug database with Mysql; applied machine learning methods (i.e. logistic regression, SVM, Naive Bayes, Decision Tree etc.) in R and Weka with feature selections to predict new drug combinations; Evaluate performance with AUC, Accuracy, MCC, etc.

2007-2008 Computational Biology lab, University of Science and Technology of China

- Built an ordinary differential equation model of insulin signaling pathway, simulated the model with Matlab and validated with literatures
- Designed an database to store all known kinetic parameters and predict unknowns

OTHER RESEARCH PROJECTS

2011-2012 Evaluated the mappability of Bowtie, BWA and BLAT algorithms for the NGS data of five mouse samples using 454 pyrosequencing in the NGS course project
2009-2011 Performed bioinformatics analysis for NMR and GC-MS metabolic dataset and LC-MS proteomic dataset and data mining the diet data with Oracle Data Miner for the colorectal cancer patients in Colorectal Cancer Engineering (CCE) projects
2009-2010 Data mining cancer protein interaction network and drug protein network to repurpose FDA approved breast cancer drugs for colorectal cancer use
2008-2009 Built a geography and cancer database at Introduction to Bioinformatics class and cancer biomarker database at Biological Database Management with PHP and MySQL.

SELECTED AWARDS

2008-present School of Informatics Scholarships
2012 Travel stipends from MCBIOS IX conference
2011 Travel fellowship for GENSIPS'11 from Indiana University Graduate School
2007 National Financial Aid, University of Science and Technology of China, China

2005 Outstanding Student, University of Science and Technology of China, China

METORSHIP AND TEACHING EXPERIENCE

2012 Teaching assistant in Computational System Biology courses, School of Informatics, Indiana University.

2010-2013 Graduate Mentor of over 10 undergraduate students in the Multidisciplinary Undergraduate Research Institute (MURI) Program at Indiana University-Purdue University Indianapolis (IUPUI), Indianapolis

VOLUNTEER

2007-2008 Team leader for contacting alumni graduated from English department to attend 50th anniversary of USTC

2006 Toured high school students to labs in School of Life Science during the National Science and Technology week

COMPUTER SKILLS

Programming languages: SQL, R, Perl/Python, Matlab, SPSS, SAS, PHP, C++

Data mining: Oracle Data Mining (ODM), Weka

Other software: Oracle/APEX, Drupal, Aqua Data Studio, Oracle SQL Developer

SELECTED PUBLICATIONS

1. **Hui Huang**, Sara Ibrahim, Thanh Nguyen, Jake Y. Chen Systematic drug repositioning with the drug directionality database (*in preparation*)
2. **Hui Huang**, Xiaoyan A.Angela Qu, Lun Yang Drug Combination Prediction Only Based on Three Clinical Side-Effects. (*summer internship report*)
3. Syed Aun Muhammad, **Hui Huang**, Xiaogang Wu, Safia Ahmed, X. Frank Yang, Jake Y. Chen Prioritizing Drug Targets in Clostridium botulinum Type A using Systems Biology Approach. BMC Bioinformatics (*submiited*)
4. **Hui Huang**, Xiaogang Wu, Ragini Pandey, Jiao Li, Guoling Zhao, Sara Ibrahim, Jake Y. Chen (2012) C2Maps: A network pharmacology database with comprehensive disease-gene-drug connectivity relationships. *BMC Genomics*. Vol. 13, Supplement 6, S17.
5. **Hui Huang**, Xiaogang Wu, Madhankumar Sonachalam, Sammed N. Mandape, Ragini Pandey, Karl F. MacDorman, Ping Wan, Jake Y. Chen (2012) PAGED: A pathway and gene-set enrichment database to enable molecular phenotype discoveries. *BMC Bioinformatics*, Vol. 13, Supplement 15, S2.
6. Xiaogang Wu, **Hui Huang***, Tao Wei, Ragini Pandey, Christoph Reinhard, Shuyu D. Li and Jake Y. Chen (2012) Network Expansion and Pathway Enrichment Analysis towards Biologically Significant Findings from Microarrays. *Journal of Integrative Bioinformatics*. Vol. 9, No. 2, pp. 213. (**Equally-contributed author*)
7. Xiaogang Wu, **Hui Huang***, Madhankumar Sonachalam, Sina Reinhard, Jeffrey Shen, Ragini Pandey, Jake Y. Chen (2012) Reordering based integrative expression profiling for microarray classification. *BMC Bioinformatics*, Vol. 13, Supplement 2, S1. (*Equally-contributed author)
8. Madhankumar Sonachalam, Jeffrey Shen, **Hui Huang** and Xiaogang Wu (2012) Systems biology approach to identify gene network signatures for colorectal cancer. *Frontiers in Genetics*, Vol. 3, pp. 80.
9. **Hui Huang**, Xiaogang Wu, Sara Ibrahim, Sunil Badve, and Jake Y. Chen (2011) Predicting Drug Efficacy Based on the Integrated Breast Cancer Pathway Model. The 2011 IEEE *International Workshop on Genomic Signal Processing and Statistics*

10. **Hui Huang**, Xiaogang Wu, Shuyu Li, Sara Ibrahim, Taiwo Ajumobi, and Jake Y. Chen (2010) Evaluate Drug Effects on Gene Expression Profiles with Connectivity Maps. 2nd International Workshop on Data Mining for Biomarker Discovery at the 2010 IEEE *International Conference of Bioinformatics and Biomedicine*.
11. **Hui Huang**; Jiao Li; Chen, J.Y.; , "Disease gene-fishing in molecular interaction networks: A case study in colorectal cancer," *Engineering in Medicine and Biology Society*, 2009. Annual International Conference of the IEEE , vol., no., pp.6416-6419, 3-6 Sept. 2009 doi: 10.1109/IEMBS.2009.5333750