Systems Toxicology: mining chemical-toxicity signaling paths to enable network medicine

A Thesis

Submitted to the Faculty

of

Drexel University

by

Kaushal D Desai

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

May 2012

Dedications

To my parents, my wife and my son.

Acknowledgments

From the first data mining course at Drexel University to the final version of this thesis, I owe an immense debt of gratitude to my supervisor, Dr. Xiaohua (Tony) Hu. His pragmatism and passion for scientific research have inspired me to push my boundaries during the doctoral program. Under Dr.Hu's mentorship and guidance, I was able to gain an in-depth understanding of data and text mining approaches that were relevant to the research questions addressed in the thesis. His enthusiasm, openness and honest feedback at various stages during the program have helped me grow as a researcher.

I would also like to express my sincere gratitude to Dr. Anastasia Christianson. Anastasia co-supervised my doctoral thesis along side her full-time position at AstraZeneca Pharmaceuticals. Our regular meetings were the ideal opportunity for me to discuss emerging research, develop a good understanding of the domain and define scientific gaps that required a systems biology approach. Anastasia has never turned down a request for help or guidance, some times even on a weekend or when she was on vacation. Her inputs have had a significant impact on the quality of this thesis. For this, I will be indebted to her forever.

My gratitude also extends to Drs. Jiexun Jason Li, Yuan An and Hualou Liang for their valuable feedback and guidance as committee members.

I would like to thank both my university and my employer, for giving me the opportunity to interact with reknowned experts in relevant scientific domains. My years at these two esteemed institutions have helped me develop a cross-disciplinary perspective to information research.

My father, his life long dream to see his kids achieve academic excellence and all the hardships he has endured to educate his kids are the very raison d'être' for this doctoral journey. My words seem inadequate as I attempt to express gratitude for the decades of sacrifices and hardship my parents have endured to see their children succeed. I also sincerely

appreciate their encouragement and support during the many months leading up to the thesis defense.

I am eternally indebted to my wife, Krupali for her patience and support under difficult circumstances as I tried to cope with a full time job, business travel and the doctoral program. A sincere note of thanks to my son, Atharva for giving up many hours of play time so I could focus on my thesis. I would also like to sincerely thank my parents-in-law for their support during the doctoral program.

This thesis is dedicated to these great individuals and to all those who have helped me in my educational journey.

Table of Contents

LIST OF TABLES vii

LIST OF ILLUSTRATIONS viii

ABSTRACT x

1. INTRODUCTION 1

1.1 Drug-induced toxicity Evaluation: Need for *In silico* Approaches 2

1.2 Characteristics of Biological Networks 4

1.3 Research Objectives 8

2. BACKGROUND AND RELEVANT WORK 11

2.1 *In silico* Toxicity Evaluation Methods: Recent Advances 12

2.2 Advances in Network Biology 17

2.3 Discovering Signaling Paths in Biological Networks 22

3. A NOVEL SYSTEMS BIOLOGY APPROACH FOR DETECTING TOXICITY-RELATED HOTSPOTS 27

3.1 Drug Toxicity Signaling Paths (DTSP) Detection Algorithm 28

3.2 Discovering DTSPs for Drug-induced Neutropenia 35

3.3 Results and Discussion 42

3.4 Limitations 49

3.5 Conclusions 49

4. INTEGRATING PROTEIN INTERACTION AND GENE EXPRESSION INFORMATION TO GAIN INSIGHTS INTO TOXICITY MECHANISMS 51

4.1 Advantages of using a Gene Expression Measure 51

4.2 Computing Edge Weights using Gene Expression Measure 55

4.3 Results and Discussion 56

4.4 Limitations 58

4.5 Conclusions 58

5. DTSP ALGORITHM: COMPARATIVE EVALUATION 61

5.1 Comparison: Toxicity-inducing vs Control Drugs 61
5.2 Comparative Evaluation of algorithm accuracy 70
5.3 Results and Discussion 74

6. CONCLUSIONS AND FUTURE WORK 82

6.1 Thesis Contributions 82
6.2 Recommendations for Future Work 83

Bibliography 86
Appendix A: Graph Theoretic Definitions 94
VITA 96

List of Tables

1. Types of Adverse Drug Reactions with examples 3
2. Properties of Biological Networks 7
3. Protein end nodes for path detection 37
4. Drugs associated with non-immune Neutropenia 38
5. List of proteins that occur in at least one DTSP across all analyzed drugs 46
6. Pathway association for DTSP proteins 48
7. Proteins involved in DTSPs discovered for each drug 59
8. Percentage of compounds with altered gene expression levels in each therapeutic category 60
9. Drugs not known to cause non-immune neutropenia 62
10. DTSPs common to both groups and regulation of toxicity-related proteins on those paths 70
11. CTD data status as of May 2011 74
12. Comparative Evaluation of DTSP algorithms 77

List of Illustrations

1. The connectivity map concept, reproduced from Lamb et al. [Lamb 2007]. a. Gene expression profiles derived from the treatment of cultured human cells with a large number of perturbagens populate a reference database. Perturbagens are ranked by a connectivity score that represents the direction and strength of enrichment of a query signature with each reference profile. b. PPARγ agonists are connected with diet-induced obesity in Rats 16

2. Direct versus module-assisted approaches for functional annotation (reproduced from Sharan et al. 2008 [Sharan et al. 2007]). The scheme shows a network in which some proteins have known annotations. Proteins with the same function are shown to have the same color. Unannotated proteins are in white. The direction of edges indicates influence of annotated proteins on unannotated ones 18

3. Algorithm pseudocode for color coding 25

4. Network model schematic for drug-toxicity signaling paths 29

5. Algorithm pseudocode for DTSP detection 30

6. Edges or protein interactions that are common to the detected set of paths are assigned a higher relevance score 33

7. STRING Database: Evidence types used to confidence score for each interaction (reproduced from Jensen et al. 2009 [Jensen et al. 2009]) 40

8. As the path length increases towards 10, the number of discovered paths involving additional proteins decreases towards zero 44

9. Plot shows the number of proteins that occur at least once across various drugs in the analysis set 45

10. Algorithm Pseudocode for detecting Drug-Toxicity Signaling Paths using microarray data 53

11. Venn Diagram shows the number of proteins common and exclusive to paths discovered for toxicity-inducing and control drugs. Geneset enrichment analysis revealed top 10 canonical pathways associated with path proteins discovered for each set of drugs. Some known pathways (like MAPK, cancer pathways and GnRH signaling pathway) were found to be common across the groups. Some others (like drug metabolism – cytochrome P450, Focal adhesion) were exclusively associated with path proteins discovered for drugs that cause non-immune neutropenia 66

12. Average Path expression score in each set reveals the greater extent of gene expression change (treated vs. control) in path proteins discovered for the Toxicity-inducing drugs 68

13. CTD integrates curated data for chemical-gene interactions, chemical- disease and gene-disease relationships (colored circles) with select public datasets (gray circles; pathways from the KEGG and Reactome databases and GO annotations). Solid lines describe directly curated or integrated relationships and dashed lines describe inferred relationships. (Figure and Legend reproduced from the BioInformatics Primer, Pathway Interaction Database, Mattingly 2011)[Mattingly, et al. 2006] [Davis, et al. 2011] 73

14. The CTD curation and integration paradigm (reproduced from Davis et al. 2011[Davis et al. 2009]. a. Biocurators capture three types of data from literature for chemicals (C), genes (G) and diseases (D): C-G interactions, G-D relationships and C-D relationships. These three relationships generate the chemical-gene-disease triad. b. The integration of these three data sets enables users to infer novel connections between chemicals-disease and gene-diseases (dashed arrow)..... 80

Abstract

Systems Toxicology: mining chemical-toxicity signaling paths to enable network medicine

Kaushal Desai

Xiaohua Hu, Ph.D.

Systems toxicology, a branch of toxicology that studies chemical effects on biological systems, presents exciting knowledge discovery challenges for the information researcher. The exponential increase in availability of genomic and proteomic data in this domain needs to be matched with increasingly sophisticated network analysis approaches. Improved ability to mine complex gene and protein interaction networks may eventually lead to discovery of drugs that target biological sub-networks ('network medicine') instead of individual proteins.

In this thesis, we have proposed and investigated the use of a maximal edge centrality criterion to discover drug-toxicity signaling paths inside a human protein interaction network. The signaling path detection approach utilizes drug and toxicity information along with two novel edge weighting measures, one based on edge centrality for detected paths and another using differential gene expression between tissues treated with toxicity-inducing drugs and a control set. Drugs known to induce non-immune Neutropenia were analyzed as a test case and common path proteins on discovered signaling paths were evaluated for toxicological significance. In addition to investigating the value of topological connectivity for identification of toxicity biomarkers, the gene expression-based measure led to identification of a proposed biomarker panel for screening new drug candidates.

Comparative evaluation of findings from the DTSP approach with standard microarray analysis method showed clear improvements in various performance measures including true positive rate, positive predictive value, negative predictive value and overall accuracy. Comparison of non-immune Neutropenia signaling paths with those discovered for a control set showed increased transcript-level activation of discovered signaling paths for toxicity-inducing drugs.

We have demonstrated the scientific value from a systems-based approach for identifying toxicity-related proteins inside complex biological networks. The algorithm should be useful for biomarker identification for any toxicity assuming availability of relevant drug and drug-induced toxicity information.

CHAPTER 1: INTRODUCTION

The problem of mining for meaningful information inside complex networks has recently attracted the information researcher's attention in a wide variety of domains, including but not limited to the network of hyperlinks on the internet [Barnett 2005] , network of social interactions among human participants [Shaikh et al. 2007], diffusion of knowledge inside organizations [Owen-Smith and Powell 2004], literature citation networks [Chen et al. 2008] and molecular interactions in complex organisms [Alfarano et al. 2005]. In the life sciences domain, the human genome project and subsequent advances in high-throughput technology were expected to transform medical research through elucidation of components that formed the cellular machinery. However, it was soon realized that the 'parts list' that emerged from human genome sequencing was far from a 'wiring diagram' and 'circuit logic' required to understand complex linkages between the genotype (i.e. organism's genetic makeup), the phenotype (i.e. organism's observable traits or characteristics) and the environment [Quackenbush 2007]. Molecular interactions between individual parts or constituents including genes, proteins and metabolites need to be examined at the level of pathways, cells, tissue and organ to ultimately understand the physiology of the entire organism or system. The field of systems biology has emerged to address this gap using a holistic approach that blends biomedical science, computational modeling and high-throughput experimentation to provide an understanding of cell signaling, developmental biology, cell physiology and metabolic networks [Oprea et al. 2007].

Systems biology approaches have leveraged network science for three primary goals – Network Inference, Network Analysis and Network modeling. Biological network inference refers to the problem of inferring the structure of biological networks and the state of network elements to construct an interaction graph underlying the system. Biological network analysis refers to the use of graph theory to analyze a known (complete or incomplete) interaction

graph and to extract new biological insights and predictions from the results. Dynamic biological network modeling aims to describe how known interactions among defined elements determine the time course of the state of the elements, and of the whole system, under different conditions. A dynamic model of the cellular machinery allows researchers to study changes in the system's behavior due to external perturbations like drug administration [Albert 2007]. Systems and Network biology are therefore important in terms of their potential applications to drug discovery and development.

A steady increase in the number of drug candidates failing in late-stage clinical development over the past decade has been concurrent with the assumption of a 'one gene, one drug, one disease' paradigm [Hopkins 2008]. Also, recent evidence has challenged the paradigm of single target intervention in drug discovery and development. Studies have shown that phenotypes are robust to single gene-knockout and only about 19% genes are essential across a number of model organisms. It is therefore, conceivable that the robustness of phenotype can be understood in terms of redundant function and alternative compensatory signaling routes inside complex biological networks [Hopkins 2007]. Polypharmacology, defined as the specific binding of a drug to multiple targets, and its effect on biological networks and phenotypes is therefore important to understand in the context of the need to evaluate drug safety as well as drug efficacy.

1.1 Drug-induced toxicity evaluation: need for In silico approaches

The ability to understand and reliably predict adverse effects of drug administration on biological systems before the drug is administered in humans continues to be a major challenge for pharmaceutical research and development. WHO defines an adverse drug reaction as "a response to a drug that is noxious and unintended and occurs at doses normally used in man for the prophylaxis, diagnosis or therapy of disease, or modification of physiological function" [Edwards and Aronson 2000]. Definition and examples of different

types of adverse drug reactions is outlined in Table 1. In the context of this thesis, it is important to note the difference between Type A (Dose-related) and Type B (Non-dose related) adverse drug reactions. While Type A drug reactions are directly related to the pharmacological action of the drug, Type B reactions include immunological or hypersensitivity type reactions, that are not related to the pharmacological action of the drug.

TABLE 1. Types of Adverse Drug Reactions with examples

Type of reaction	Mnemonic	Features	Examples
A: Dose-related	Augmented	• Common • Related to a pharmacological action of the drug • Predictable • Low mortality	• Toxic effects: Digoxin toxicity; serotonin syndrome with SSRIs • Side effects: Anticholinergic effects of tricyclic antidepressants
B: Non-dose-related	Bizarre	• Uncommon • Not related to a pharmacological action of the drug • Unpredictable • High mortality	• Immunological reactions: Penicillin hypersensitivity • Idiosyncratic reactions: Acute porphyria Malignant hyperthermia Pseudoallergy (eg, ampicillin rash)
C: Dose-related and time-related	Chronic	• Uncommon • Related to the cumulative dose	• Hypothalamic-pituitary-adrenal axis suppression by corticosteroids
D: Time-related	Delayed	• Uncommon • Usually dose-related • Occurs or becomes apparent some time after the use of the drug	• Teratogenesis (eg, vaginal adenocarcinoma with diethylstilbestrol) • Carcinogenesis • Tardive dyskinesia
E: Withdrawal	End of use	• Uncommon • Occurs soon after withdrawal of the drug	• Opiate withdrawal syndrome • Myocardial ischaemia (β-blocker withdrawal)
F: Unexpected failure of therapy	Failure	• Common • Dose-related • Often caused by drug interactions	• Inadequate dosage of an oral contraceptive, particularly when used with specific enzyme inducers

SSRIs=serotonin-selective reuptake inhibitors.

The use of computational methods for evaluation of toxic potential for new chemical entities is termed as *In silico* toxicity evaluation. A variety of *in vivo (inside a living organism, e.g. animal studies)* and *in vitro (outside a living organism, e.g. test tube or assay-based evaluation)* toxicity screens are available for screening drug candidates. More recently, high-throughput technologies have provided the ability to measure changes in mRNA (messenger RNA) levels for thousands of gene transcripts on a single chip. This, combined with advances in proteomics (ability to measure proteins as products of translation using

protein expression, MS and other protein-based measurement technologies) has led to an explosion in gene and protein-level information inside publicly available databases. While current approaches for analyzing genomic and proteomic data have been successful to some extent, it is expected that a 'systems' level approach that integrates these diverse data types will provide insights that are likely to improve the success of the drug development process [Ganter et al. 2006]. It is expected that viewing drug action through the lens of network biology may provide insights into the safety and efficacy of novel pharmaceutical agents.

The high-level objective of this dissertation is to improve *In silico* toxicity evaluation using network biology.

1.2 Characteristics of Biological Networks

A network can be defined as a graphical structure with nodes as their fundamental units. The interconnections between nodes in a network are represented by weighted or unweighted edges. Each node may represent a physical or conceptual entity from the specific domain of study. In case of complex biomolecular networks, the nodes may represent genes, proteins or other cellular components connected by edges that represent interactions between cellular components including chemical reactions and biophysical interactions. A network model of biological entities is used to represent complex interactions that result in biological changes as a result of environmental stimulus at the cellular, tissues/organ and organism level.

In order to understand the characteristics of real world networks, many theoretical network models have been proposed. Network models and their parameters provide an overview of the global structure or topology of these important cellular networks. The most general level of network analysis comes from global network measures that allow us to characterize and compare the given network topologies (i.e. the configuration of the nodes and their connecting edges). Global measures such as the degree distribution (the degree of a node is the number of

edges it participates in) and the clustering coefficient (defined in section 2.1) have recently been thoroughly reviewed in the context of cellular networks and in proteomics.

The small-world model is defined as one where any two nodes in the network can be connected through a much shorter path than would be expected in a random network of similar size and number of connections. Metabolic networks have been found to be small-world networks and additionally, the network diameter does not appear to vary between different organisms [Alm and Arkin 2003]. Small-world networks have low average path lengths and high clustering coefficients.

Scale-free networks are characterized by a connectivity distribution that decays as power law.

$$P(k) = Ak^{-\gamma}$$

where A is the normalization constant and the degree exponent γ is between 2 and 3 [Albert 2007, Barabasi and L. 1999, Jeong et al. 2000]. Essentially, this means that there are a small but finite number of highly connected nodes in the system, forming the so-called 'hubs'. It has been shown that many biological networks follow a power-law distribution, including protein families, super-families, folds, short DNA words and even pseudogene families [Luscombe et al. 2002]. Like small-world networks, scale-free networks have low average path lengths and high clustering coefficients.

The non-random nature of biological networks has to do with the biological functions of nodes and edges. Several studies in yeast have revealed correlations between the topology and composition of a network and important biological properties of nodes [Jeong et al. 2001]. The well-connected hubs are largely represented by evolutionary conserved proteins because the interactions impose certain structural constraints on sequence evolution [Fraser et al. 2002].

It is important to study biological phenomena at a systems level because certain properties of biological networks can only be observed at that level. Table 2 defines properties like modularity, robustness, redundancy etc. with a biological example from the process of hemostasis and coagulation.

Recently, another topological feature of the network has received attention – betweenness, which measures the total number of non-redundant shortest paths going through a certain node or edge. Betweenness was originally introduced to measure the centrality of the nodes in the network [Yu et al. 2007]. Girvan and Newman proposed a network partitioning algorithm, based on the shortest-path algorithm in graph theory, that iteratively removes edges with the highest betweenness until the network breaks down into individual clusters [Girvan and Newman 2002]. Multiple authors have proposed improvements over the Garvin and Newman approach, including the use of 'edge clustering coefficient' [Radicchi et al. 2004], combining clique detection, superparamagnetic (SPC) clustering and Monte Carlo optimization (MC) to search for functional modules in yeast protein network [Spirin and Mirny 2003]. Chapter 3 describes the implementation of an edge betweenness centrality measure to discover signaling paths associated with drug-induced toxicity.

While large-scale characteristics of biological networks (e.g. small-world, scale-free) provide insight into topological properties at a network-wide level, it has been observed that the most important biological processes such as signal transduction, cell-fate regulation, transcription and translation involve more than four but much fewer than hundreds of proteins [Spirin and Mirny 2003]. It is believed that most relevant processes in biological networks correspond to the meso scale (5-25 genes/proteins). Algorithms for identification of clusters (sets of proteins having many more interactions among themselves than with the rest of the network) have led to discovery of protein complexes and functional modules inside interaction networks.

TABLE 2. Properties of Biological networks

Component	Definition	Hemostatic example
Robustness	Ability to maintain function in the presence of changes in the system	Asymptomatic protein C deficiency or factor V_{Leiden}
Tolerance	One abnormal component within a module will not bring down the system	Protein C levels of 30–40% that are asymptomatic. Factor XIII level >2% is asymptomatic
Modularity	Subsystems are physically or functionally insulated so failure of one module does not spread and lead to system-wide failure	Platelet mechanisms and coagulation pathways
Redundancy	Proteins work in more than one area (multiple components with equivalent function for back-up)	Protein C/protein S to regulate factors V and VIII versus antithrombin to regulate coagulation enzymes
Adaptation	Ability to cope with environmental changes	Increased factor VII and von Willebrand factor levels to compensate for pregnancy bleeding
System control	Negative feed-back or feed-forward mechanisms	Thrombin activation of factors V and VIII
Structural stability	Intrinsic mechanisms to promote stability	von Willebrand factor binding of factor VIII to maintain local clot formation
Flexibility	System can have components that are decreased (or increased) and the system still functions	Asymptomatic protein C and protein S deficiency
Dynamic	Regulatory systems control with time, space and amplitude	Fibrinolysis system is slower to activate than coagulation system
Protocols	Rules designed to manage relationships of isolated modules and processes smoothly and effectively	Fibrin monomer concentration generates fibrin clot formation
Graceful degradation	Slow degradation of system's function after damage rather than catastrophic failure	Mild to moderate disseminated intravascular coagulation without catastrophic failure

Functional modules consist of proteins that participate in particular cellular processes while binding to each other under specific conditions (e.g. after drug administration, during specific cell cycle phases, inside specific cellular compartments etc.). Protein/Module function prediction under drug administration conditions may help reveal the mechanism of action for the drug and identification of its toxic effects early during drug development.

1.3 Research Objectives

This thesis focuses on the following research questions –

1) Can the use of a network model provide insights into mechanisms of drug-induced toxicity?
2) How can toxicity related proteins or 'hot spots' be discovered inside biological networks?
3) What is the biological significance of discovered 'hotspots' with respect to the drug-induced toxicity being evaluated?

Evaluation of a candidate drug for potential toxicities during early stages of drug research is challenging due to absence of experimental evidence in humans. Animal studies don't always provide an accurate estimate of probability of occurrence for a particular effect in humans and also cannot represent human cellular interactions completely. The use of animal testing also needs to be minimized for ethical reasons. The emergence of novel high-throughput technologies has therefore, led to emphasis on '*In silico*' (computational) approaches to toxicity evaluation. Use of biological networks for drug-induced toxicity evaluation has been challenging for various reasons. First, the scientific information required to construct various elements of such a network (genes, proteins, phenotypes etc.) is dispersed across multiple public databases and inside an ever-growing volume of published literature [Dietmann et al. 2006]. Use of standardized vocabularies and taxonomies for various biological annotations has recently enabled genome-wide, cross-platform integration of biomolecular data into system-wide biological networks. Second, analysis of the integrated biological network requires computationally intensive mining algorithms that can handle very large networks. The exponential increase in availability of genomic and proteomic data in this domain needs to be matched with increasingly sophisticated network analysis approaches. As an example, Steffen et al. discovered MAPK signaling paths inside a filtered yeast protein-protein

interaction (PPI) network consisting of 5560 interactions and 3725 proteins [Steffen et al. 2002]. Compared to this, the human subset of STITCH database (version 1.0) consists of ~18600 proteins and ~1432500 unique protein-protein interactions. Mining such a large network to elucidate toxicity-related proteins, requires sophisticated algorithmic approaches that can run in reasonable time with manageable computational burden. Third, the identified protein or set of proteins should have toxicological relevance to the effect being evaluated. Analysis of networks may yield insights into a wide variety of physiological processes. However, it is important to be able to distinguish between a physiological effect that is 'normal' from effects that can be directly associated with external stimuli like drug administration. Similarly, external stimuli can typically lead to a wide variety of clinical effects. It is therefore important for network algorithms to be able to associate a particular systemic change to a specific clinical manifestation, namely drug-induced toxicity.

If the above challenges can be addressed appropriately, a network-based model may capture the complexity of human physiological networks more accurately compared to toxicity screens that are based on studying the effects of drug administration on one or a few proteins and their interactions. Dynamic programming and fixed parameter tractability offers solutions to problems that not very long ago, were considered computationally intractable. The challenge for an information researcher is to utilize domain knowledge and develop algorithms that can leverage advances in applied mathematics to discover knowledge inside complex networks. The primary motivation for this thesis is to devise analytical methodologies that can mine such large biological networks to provide insights into drug effects. This thesis proposes and implements two algorithms that leverage multiple sources of information and mine a biological network for protein modules or 'hot spots' that may be associated with a particular drug-induced toxicity. The algorithms are then compared with findings from

microarray data analysis, the prevalent approach for identification of biomarkers of drug-induced toxicity. Finally, thesis contributions and directions for future work are summarized in chapter 6.

CHAPTER 2: BACKGROUND AND RELEVANT WORK

Many drug candidates fail in clinical trials and are withdrawn because of unforeseen effects of human metabolism, such as toxicity and unfavorable pharmacokinetic profiles. Early pre-clinical elimination of such compounds is important but not always possible, due to lack of accurate and reliable predictive models. Three broadly defined resources have been considered as components of pharmaceutical *In silico* toxicity evaluation systems [Bugrim et al. 2004] –

First, chemical structure-activity relationships (SAR) computational models are built within groups of structurally similar compounds and aimed at effective prediction based on experimental data (toxicity end points, metabolic products and intermediates). The SAR approach leverages empirical rules for modeling metabolism and toxicity based on chemical structure of lead compounds.

The second set of technologies relevant to an *In silico* toxicity evaluation system is collectively referred to as 'toxicogenomics' platforms. Toxicogenomics has been defined as an integration of genomics and toxicology [Gomase and Tagore 2008]. It is a branch of science that aims to study the interaction between the cell's genome, chemicals in the environment and disease. Toxicogenomics studies classify toxicities based on gene transcriptional patterns observed on high-throughput gene expression platforms using tissues relevant to specific toxicological endpoints. Patterns from new chemicals and tissue samples can be compared with those in 'reference' databases to extrapolate the probability of toxicity [Bugrim et al. 2004].

The third source of information relevant to *In silico* toxicity prediction is function pathway databases and enzymatic reaction repositories. These systems biology databases capture information around chemical reactions, metabolic cascades and signal transduction routes across a variety of transporters, metabolizing enzymes and protein interaction.

Advances in structure-activity computational models and 'omics' approaches to toxicity evaluation are relevant to this work and will therefore be discussed in section 2.1. Individually, none of these technologies is sufficient for the task of predicting the ADME (Absorption, Distribution, Metabolism and Excretion)/ Tox (toxicity) behavior of new chemical entities [Bugrim et al. 2004]. For example, empirical structure-based models do not deal with the underlying mechanisms of toxicity and metabolism, genomics-based systems and patterns resulting from analysis of data on a single platform have limited value in the absence of validation from other experimental and observational sources and pathway databases do not reflect the spatial or temporal patterns of pathway activity, and therefore reflect the functional potential of a cell or tissue rather than the real activity in response to stimuli. Because the understanding of metabolism and toxicity is a systems-level problem, the solution requires integration of data from various sources into one comprehensive model. Section 2.2 and 2.3 review two areas that address the need for a 'systems' level approach, namely network biology which aims to analyze complex networks created through integration of multiple sources of biomolecular data and automated signaling path detection, a set of algorithms aimed at automated discovery of signaling paths inside large biomolecular networks.

2.1 In silico Toxicity Evaluation Methods: Recent advances

2.1.1 Advances in structure-activity (SAR) based ADME/Tox modeling

In silico ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicology) modeling has received increasing attention in the post-genomic era. From some of the early reviews of systems biology it has been proposed as an approach to understanding pharmacodynamic and toxicological properties of chemical entities [Kitano 2002, Ekins 2006].

Expert systems based approaches have aimed to go beyond the limitations of traditional quantitative structure-activity relationship (QSAR) models for prediction of drug activity.

Expert systems fall into four broad categories – automated QSAR, knowledge-based systems, automated rule derivation and decision tree-based algorithms [Dearden 2003]. More recent advances in ADMET modeling are exemplified by software systems like TOPKAT™, CASE™, DEREK™, HazardExpert™ etc. Both TOPKAT™ and CASE™ are automated QSAR expert systems where toxicity prediction is based on use of predictive models (e.g. linear regression for continuous toxicity endpoints like LD_{50} and logistic regression for discreet endpoints like mutagenicity and carcinogenicity). While TOPKAT uses mainly topological, sub-structural and electronic descriptors in its predictive model, CASE creates its own structural alerts by breaking down each molecule into possible fragments from two to ten heavy atoms. DEREK™, HazardExpert™ and OncoLogic™ are knowledge-based expert systems that relie on an expert-curated rulebase for prediction of toxicity-related endpoints. Various endpoints including carcinogenicity, mutagenicity, teratogenicity, irritation, ACD (allergic contact dermatitis), acute toxicity, CYP450 2D inhibition and cellular toxicity can be predicted with varying degrees of accuracy using models provided by various *in silico* ADMET platforms.

Another interesting approach related to prediction of drug effects in new chemical entities was termed as 'biological spectra similarity' by Fliri et al [Fliri et al. 2007]. The authors estimated similarity between 1064 drugs based on six descriptor sets starting with two 'signal input' (compound structure) descriptors, two intermediate cellular level (protein-ligand displacement and Medline based protein-drug co-investigation) descriptors and two organism level (MedDRA-based literature mining and COSTART based literature mining for adverse events) descriptors as clinical manifestations of drug effects. Drug-protein association similarity and protein-adverse events association similarity were computed using hierarchical clustering to elucidate compounds with similar structure and effect relationships. For a new set of compounds, the authors found that compounds with high similarity based on 'structure

information spectra' also showed high similarity along the 'effects information spectra'. The authors concluded that structure-effect correlations become quantifiable only within certain structural boundaries and that the ability to compare broad drug effects on cellular function with broad drug effects on organisms is a key to the design of medicines. Paolini et al. constructed a polypharmacology interaction network using structure-activity relationships (SAR) integrated from diverse sources [Paolini et al. 2006]. A global mapping of chemical structure, protein sequence and disease indication enabled construction of a ligand-target matrix that could be used as a probabilistic model to predict pharmacology from a large knowledgebase. The authors could predict potential activity of a compound against the protein space with 56.7% accuracy, representing a 153-fold enrichment over random.

A major limitation of QSAR-based approaches is the scarcity of experimental data outside a small chemical space of well-studied chemical structures and endpoints like carcinogenicity, mutagenicity etc. In order for a QSAR to model biological data well and to be predictive, all the compounds involved should act by the same mechanism, since the physico-chemical and structural descriptors used in the QSAR are deemed to reflect mechanism of action [Dearden 2003]. Moreover, expert systems are aimed at prediction of toxicity elucidation and don't directly contribute to elucidation of molecular mechanisms of toxicity. High-throughput technologies like genomics and proteomics may help address this limitation.

2.1.2 Advances in 'Omics' approaches to *In silico* toxicity evaluation

'Omics' approaches consist of genomics, proteomics and metabonomics for evaluation of drug-induced toxicity. High-throughput genomics data is abundantly available from gene expression or microarray platforms while proteomics data is also more recently being made available in the public domain in the form of protein binding/interaction and protein expression data.

Waring et al. investigated whether gene expression profiles can be used for *In silico* toxicity evaluation, by profiling gene expression in rats treated with 15 different known hepatotoxins [Waring et al. 2001]. Results showed a strong correlation between the histopathology, clinical chemistry and gene expression profiles induced by these agents, thereby confirming that gene expression data may provide insights into drug-induced toxicity similar to other well-established methods.

Lamb et al. proposed gene expression signature similarity as a means to connect diseases, genetic perturbation and drug action [Lamb et al. 2006]. The authors constructed a 'Connectivity map' using gene expression profiles for 164 distinct small molecule perturbagens inside four different types of cell lines, 6 hours and 12 hours after compound treatment. Using a non-parametric, rank-based pattern matching strategy, the authors were able to identify compounds from the same class (HDAC inhibitors) using a class-specific query signature, as shown in Figure 1. The connectivity map identified compounds with negative connectivity as well as compounds with similar effects. Most of the signatures were generated from a single cell type. Signatures linked diseases to genes and genes to drug effects only on the basis of gene expression changes, without necessarily arriving at a toxicity-related functional module and without use of proteomic information that confirms the post-translational changes that may validate gene expression-based observations.

The task of predicting the extent to which various proteins play a role in drug-induced toxicity may be considered analogous in its intent to the task of assigning new functional association to proteins with unknown function. In this context, approaches collectively referred to as 'guilt by association' are relevant to this work. Guilt by association has been broadly defined as the use of contextual information for *in silico* inference of protein function [Aravind 2000]. Various types of contextual information have previously been used in order to predict protein function, including phyletic profile (pattern of occurrence of orthologs of a particular gene in a set of

genomes under comparison), sequence homology and gene neighborhood along the chromosome [Huynen et al. 2000]. More recent approaches that integrate genomic and proteomic information and utilize biological networks may also be broadly categorized as 'guilt by association' [Dittrich et al. 2008].

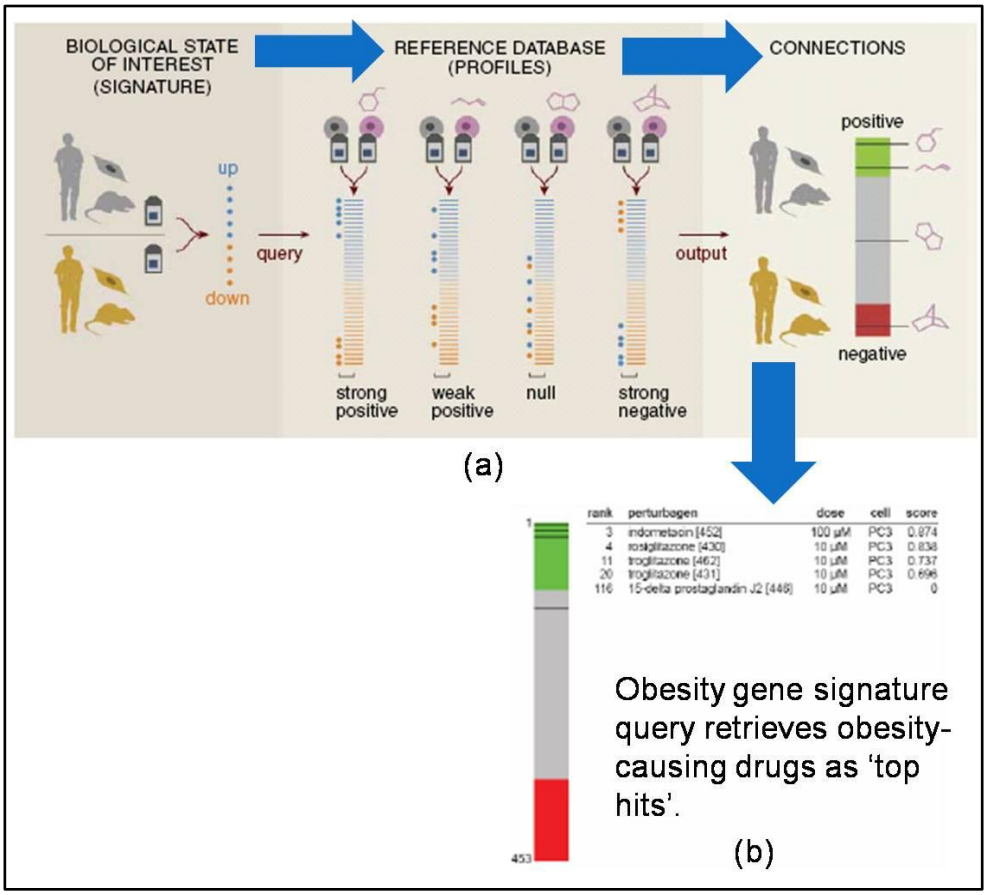

Figure 1. The connectivity map concept, reproduced from Lamb et al. [Lamb 2007]. a. Gene expression profiles derived from the treatment of cultured human cells with a large number of perturbagens populate a reference database. Perturbagens are ranked by a connectivity score that represents the direction and strength of enrichment of a query signature with each reference profile. b. PPARγ agonists are connected with diet-induced obesity in Rats.

However, algorithms that can mine large biological networks to elucidate the role of specific proteins in drug-induced toxicity are yet to be developed. The next section reviews network-based approaches in detail.

2.2 Advances in Network Biology

Two areas of application for network biology are relevant to this thesis – first, algorithms that predict protein function based on their topological and other characteristics inside biological networks are relevant because as mentioned in the research objectives, one of the primary goals of this thesis is to discover toxicity-related proteins inside a large biological network. Second, approaches that analyze disease-gene-drug networks are important because one of the goals of this thesis is to discover relationships between genes/proteins and drug-related effects that manifest as adverse events after drug administration. Section 2.2.1 and 2.2.2 review these two research areas in network biology.

2.2.1 Network-based prediction of protein function

Current approaches to prediction of protein function have relied on network topology (connectivity of unannotated and annotated proteins in the network) as well as their known functional association [Sharan et al. 2007, Chua et al. 2006, Letovsky and Kasif 2003, Hu et al. 2007].

Sharan et al. distinguished between two types of approaches, namely **Direct annotation schemes**, which infer the function of a protein based on its connections in the network and **Module-assisted schemes**, which first identify modules of related proteins using different approaches and then annotate each module protein based on known function of its members, as shown in Fig. 2.

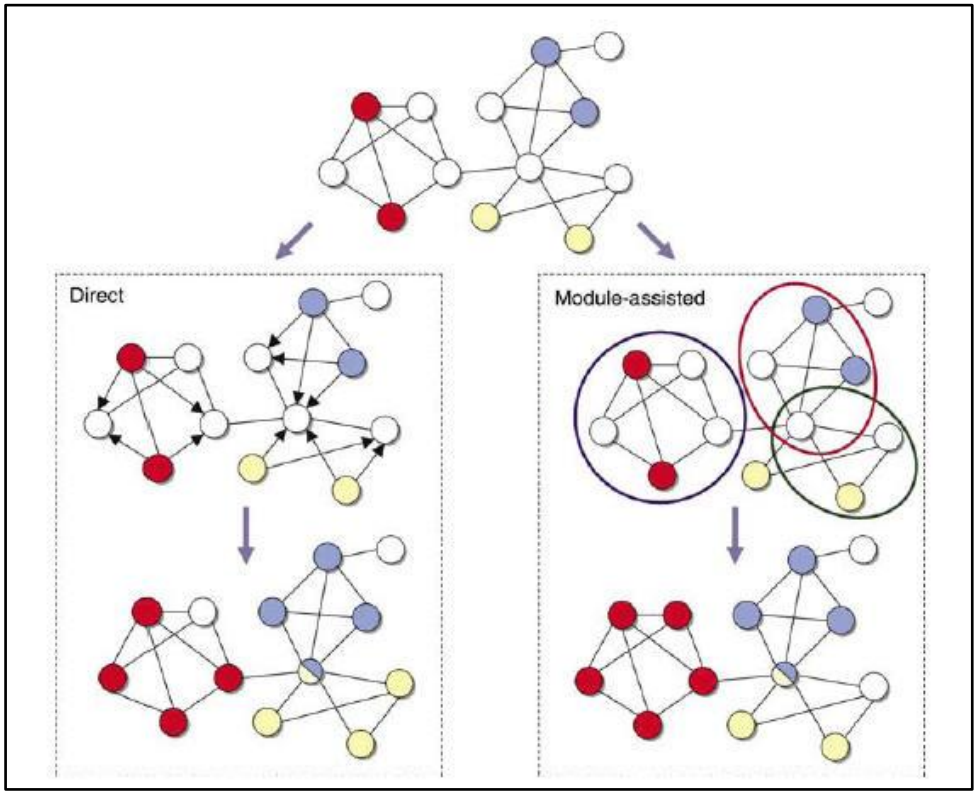

Figure 2. Direct versus module-assisted approaches for functional annotation (reproduced from Sharan et al. 2008 [Sharan et al. 2007]). The scheme shows a network in which some proteins have known annotations. Proteins with the same function are shown to have the same color. Unannotated proteins are in white. The direction of edges indicates influence of annotated proteins on unannotated ones.

Direct Annotation approaches include simple neighborhood counting, where the function of a protein is determined by known function of its immediate neighbors. The basic neighborhood counting methods don't provide a statistical significance level or a level of confidence for predicted functional annotations [Schwikowski et al. 2000]. Hishigaki et al. proposed a variation to the simple neighborhood counting by computing a γ^2-like score for function assignment for all n-neighboring proteins [Hishigaki et al. 2001]. Within the n-neighborhood, proteins at different distances from p are treated in the same way. Chua et al tried to tackle the problem by investigating the relation between network distance and functional similarity [Chua et al. 2006]. Considering just the 1- and 2-neighborhoods of a

protein, a functional similarity score was computed to assign different weights to proteins according to their distances from the target protein.

Graph theoretic methods take into account the full topology of the network and mainly consist of two types of approaches - cut-based and flow-based. Vazquez et al. aimed at assigning a function to each unannotated protein so as to maximize the number of edges between proteins with the same function [Vazquez et al. 2003]. The authors computed a score that counted the number of interacting proteins with the same functional assignment and associated it with any given configuration of functions for the whole set of unannotated proteins. The objective of the algorithm was to minimize the score so as to find a configuration with the least number of interactions between proteins not annotated with the same function. Function prediction therefore became a global optimization problem where multiple solutions were possible. The authors therefore, developed an 'objective' way to assing multiple functions to some proteins, depending on whether multiple solutions were available to achieve the minimum score configuration. Karaoz et al. applied a similar approach but considered one function at a time and compared two types of edge weighting approaches, one that captured only qualitative functional links between proteins and one that integrated gene co-expression so that the weight of an edge in the integrated network was the absolute value of the correlation coefficient of the gene-expression profiles of the pair of interacting proteins [Karaoz et al. 2004]. Deng et al. applied a Markov Random Fields approach to assign function to unannotated proteins based on function of annotated proteins and presence of interaction between proteins with known function [Deng et al. 2003].

Module-assisted methods differ in their approach for detecting functional modules inside the network. Some module detection algorithms are solely based on network topology while others utilize additional data sources, such as gene expression measurements or deletion phenotypes. Bader et al. used a core clustering coefficient to add interaction weights to edges

in the network and applied a greedy search algorithm to discover modules [Bader 2003]. In order to predict the size of a cluster or module, the authors developed a mathematical model for the expected number of closed loops in a clustered network. Comparing observations on the actual network with those from randomized networks, the analytical model predicted a cluster size of 15 ± 2 proteins. Hierarchical clustering approaches have also been used with a variety of similarity measures for module identification. It has been shown that pairwise distances may not be appropriate as a similarity measure because they tend to be identical between many protein pairs [Arnau et al. 2005]. King et al. partitioned the nodes of a given graph into distinct clusters, depending on their neighboring interactions, with a cost-based local search algorithm that updates a list of already explored clusters that are forbidden in later iteration steps. Clusters with either low functional homogeneity, cluster size or edge density were filtered out [King et al. 2004]. It has been postulated that module members may have similar shortest-path profiles. This criteria has been applied to identify modules inside small subnetworks, along with other properties like centrality measures [Dunn et al. 2005]. Several methods use expansion of complex seeds to find protein complexes. The SEEDY algorithm constructs complexes by adding proteins to a given seed, as long as the reliability of the most reliable path from a candidate to the seed does not fall below a certain threshold [Bader 2003]. Wu and Hu proposed an algorithm called *Commbuilder* that accepts a seed protein, gets the neighbors (*FindCore* component), finds the core of the community to build and expands the core (*ExpandCore* component) to find the eventual community [Hu and Wu 2007]. *Findcore* performs a naive search for maximum clique from the neighborhood of the seed protein by recursively removing vertices with the lowest in-community degree until all vertices in the core set have the same in-community degree. *ExpandCore* performs a breadth-first expansion to first create a set containing the core and all neighbors of the core. It then adds to the core a vertex that either meets the quantitative definition of a community in a strong sense (i.e.

number of edges connecting the vertex to vertices within the core is greater than the number of edges connecting the vertex to vertices outside the core) or the fraction of in-community degree over a relaxed (user-defined) affinity threshold of the size of the core. The affinity threshold is 1 when the candidate vertex connects to each of the vertices in the core set.

While the above approaches are useful for predicting protein functional association, identification of proteins that may play a role in drug-induced toxicity requires incorporation of background knowledge about the drug and the toxicity under consideration. Drug administration perturbs the biological network at specific 'nodes' (referred to as drug target proteins) that may be several levels upstream (or several nodes apart in terms of distance) from toxicity-related 'end nodes' in the network. Mapping the cascade of proteins associated with drug response and identification of network 'hotspots' is therefore, a central challenge in toxicity evaluation.

2.2.2 Analysis of Disease-Gene-Drug Networks

Network biology has recently been used to mine gene-protein-disease-drug networks [Yildirim et al. 2007, Goh et al. 2007, Xu and Li 2006, Ekins et al. 2005]. Yildrim et al. constructed a bipartite network of all FDA approved drugs and their protein targets using data from the drugbank database [Yildirim et al. 2007]. The resulting network was analyzed to elucidate an overabundance of 'follow on' drugs (drugs that 'hit' previously known targets) and a trend towards greater polypharmacology through functionally diverse targets for emerging drugs. The authors also found significant differences between 'palliative' and 'etiological' drugs with respect to the shortest distance between the drug targets and their respective gene products.

Goh et al. constructed a human disease network using gene-disease associations found in the NCBI- OMIM (Online Mendelian Inheritance in Man) database [Goh et al. 2007]. Genes associated with similar disorders showed higher likelihood of physical interactions between

their products and higher expression profiling similarity of their transcripts, supporting the existence of distinct disease-specific functional modules. While essential human genes were found to be more likely to encode 'hub' proteins and were expressed more widely in most tissues, majority of the non-essential genes were located on the periphery of the human disease network.

Identification of functional modules [Chen, et al. 2006], 'hubs' or 'hot spots' [Begley, et al. 2004,] in protein-protein interaction networks is considered the first step towards understanding the organization and dynamics of cell function [Hu and Wu 2007]. *In silico* toxicity evaluation and prediction approaches are yet to benefit from insights gained from a systematic analysis of biological networks.

2.3 Discovering signaling paths in biological networks

With the advent of genomic and proteomic data, various approaches have been proposed to discover biologically meaningful substructures such as dense groups of interacting proteins and loop structures. Linear pathways play an important role in signal transduction inside biological networks. They are easy to understand and analyze and, as demonstrated by Ideker et al. for galactose metabolism in Yeast, they can serve as seed structure for experimental investigation of more complex mechanisms [Ideker et al. 2001, Hüffner et al. 2008].

Steffen et al. constructed a network of 5560 non-redundant protein-protein interactions among more than 3725 proteins using yeast-two-hybrid technique and ran an exhaustive search to identify all possible linear paths up to length eight starting at any membrane protein and ending at transcription factors. The search was applied to an unweighted interaction graph, considering all interactions equally reliable. Microarray data was then used to rank all paths according to the degree of similarity in the expression profiles of pathway proteins. The algorithm, called *NetSearch*, was able to reproduce many of the essential elements of known MAPK pathways in Yeast [Steffen et al. 2002]. Kelly et al., developed an efficient algorithm

(called PATHBLAST) to detect simple paths in a graph that is based on finding acyclic orientations in the graph's edges [Kelley et al. 2003]. More recently, network algorithms for detecting signaling paths have improved upon earlier approaches in two ways –

a. Assignment of reliability score to interactions (use of weighted graph) and

b. Use of a powerful algorithmic technique by Alon et al. [Alon et al. 1995] called color coding, to find high-scoring paths efficiently.

A more efficient algorithm like color coding is necessary for detection of signaling paths because the problem of finding linear pathways in a large network is NP Hard or computationally intractable (the travelling salesman problem is reducible to it). NP-Hardness implies an inherent combinatorial explosion in the solution space that leads to running times growing exponentially with the input size. This means that large instances of NP-hard problems cannot always be solved efficiently to optimality. *Heuristics* drop the demand for useful running time guarantees and are tuned to run fast with good results on typical instances. *Approximation algorithms* trade the demand for optimality for a provably efficient running time behavior, while still providing provable bounds on the solution quality. However, these methods have serious drawbacks. In many applications, it is not acceptable that there is a chance that the algorithm might take exceptionally long in corner cases; and the approximation guarantees that are typically obtained are rather weak and it has been shown that a guarantee such as 10% error is often not attainable [Hüffner 2007]. *Parameterized complexity*, as accomplished with the color coding approach, offers a third alternative where the structural complexity is measured with a nonnegative parameter (denoted by k) such that growth of the running time is determined by both, the input size and the parameter k. Algorithms that leverage parameterized complexity, confine the combinatorial explosion to the parameter such that instances solve fast whenever the parameter is small. As explained by Hüffner et al. [Hüffner 2007], Fixed Parameter Tractability (FTP) is different from 'polynomial time solvable

for fixed *k*'; an algorithm running in $O(n^k)$ time demonstrates that a problem is polynomial-time solvable for any fixed *k*, but does not show fixed-parameter tractability.

Color coding is a randomized algorithm for finding simple paths and simple cycles of a specified length *k* within a given graph. Consider a weighted interaction graph in which each vertex is a protein and each edge (*u*, *v*) represents an experimentally observed interaction between proteins *u* and *v*, and is assigned a numerical value *p*(*u*, *v*) representing the probability that *u* and *v* interact. Each simple path in this graph can be assigned a *score* equal to the product of the values assigned to its edges. Among paths of a given length, those with the highest scores are plausible candidates for being identified as linear signal transduction pathways. Scott et al. extended the color coding algorithm to accurately recover well-known MAP Kinase and ubiquitin-ligation pathways [Scott et al. 2006]. The implementation however, was limited to path lengths of 10 vertices and moreover, required some hours of run time. Hüffner et al. used a dynamic programming approach to enable detection of paths up to a length of 20 vertices in some hours. The task of finding pathways of length 10 could be accomplished in a few seconds [Hüffner et al. 2008].

Following Hüffner et al., an implementation of the color coding algorithm can be summarized as follows:

The network is modeled as an undirected graph where each protein is a vertex and each edge is weighted by the negative logarithm of the interaction probability for the two proteins it connects. In other words, in order to work with an additive weight rather than a multiplicative one and to formalize the problem of high probability pathway candidate detection to a NP-hard problem called MINIMUM-WEIGHT PATH, each edge is assigned a weight $w(v,u) = -\log p(u,v)$. The weight of a path is defined as the sum of the weights of the edges and the length of the path as the number of vertices it contains.

```
Algorithm: COLORCODING(G = (V,E))
1  repeat a sufficient number of times:
2  for each  v ∈ V:
3      color v randomly;
4  if TRIAL(G):
5      return true
6  return false
```

Figure 3. Algorithm pseudocode for color-coding

MINIMUM-WEIGHT PATH

Input: An undirected edge-weighted graph $G = (V, E)$ with n:=$|V|$ and m:=$|E|$ and an integer k.

Task: Find a length-k path in G that minimizes the sum over its edge weights.

The primary goal with color coding is to randomly color the vertices of an input graph with k colors and then search for *colorful* paths, that is, paths where no color occurs twice. Given a fixed coloring of vertices, Hüffner et al. find the minimum-weight colorful path using dynamic programming: Assume that for some $i < k$ a value $W(v, S)$ is computed for every vertex $v \in V$ and cardinality-i subset S of vertex colors; this value denotes the minimum weight of a path that uses every color in S exactly once and ends in v. Clearly, this path is simple because no color is used more than once. This value can now be used to compute the values $W(v, S)$ for all cardinality -(i + 1) subsets S and vertices $v \in V$ because a colorful length-(i+1) path that ends in a vertex $v \in V$ can be composed of a colorful length-i path that does not use the color of v and ends in a neighbor of v. Mathematically,

$$W(v, S) = \min_{e=\{u,v\} \in E} (W(u, S \setminus \{\text{color}(v)\}) + w(e))$$

The coloring of input graph is random and hence many coloring trials have to be performed to ensure that the minimum-weight path is found with a high probability. One of

the strengths of the color coding approach is that it can be adapted to many practically relevant variations in problem formulation, including the problem of finding paths with maximum product of edge weights as an optimization criterion. As mentioned by Sharan et al., automated signaling path detection approaches have not been applied towards prediction of protein function [Sharan et al. 2007]. Chapter 3 and 4 demonstrate how the use of color coding technique in combination with appropriate edge-weighting criteria can be used to understand the role of network proteins in drug-induced toxicity.

CHAPTER 3: A NOVEL SYSTEMS BIOLOGY APPROACH FOR DETECTING TOXICITY-RELATED HOTSPOTS

This chapter describes the implementation of a novel systems biology approach for detecting toxicity-related proteins or sets of proteins inside a biological network. The algorithm, called 'Drug Toxicity Signaling Path' (DTSP) detection, utilizes an edge centrality measure in conjunction with the color coding algorithm to identify toxicity-specific modules inside protein interaction networks. Section 3.1 provides a conceptual description of the algorithm, followed by implementation details in section 3.2 and a discussion of results and conclusions in section 3.3 and 3.5 respectively.

The rationale for use of protein interaction data to implement the algorithm can be summarized as follows –

At least three different types of molecular interaction networks have previously been considered for network biology research: the protein-protein interaction network, the transcription regulatory network and the small-molecule metabolism network. Technologies like Mass Spectrometry have led to unprecedented generation of protein-protein interaction data at the cost of a large fraction of false-positives as well as false-negatives [Alm and Arkin 2003]. Proteins are important players in executing the genetic program and also act as targets for modulating biological processes and pathways after drug administration. When carrying out a particular biological function or serving as molecular building blocks for a particular cellular structure, proteins rarely act individually. Understanding the network of proteins is important because drug-induced toxicity is a result of interactions between proteins, with DNA, RNA and small molecules that form molecular machines. These machines are modular, involve both static and dynamic assemblies of macromolecules, and transmit as well as respond to intra- and extracellular signals [Ideker and Sharan 2008]. This thesis, therefore, proposes a novel

analytical methodology to mine protein-protein interaction networks and discover toxicity-related hotspots.

3.1 Drug Toxicity Signaling Path (DTSP) Detection Algorithm

Beneficial, deleterious or neutral functional effects of drugs can be mediated via binding to the desired therapeutic target and/or to other molecular targets such as G-protein-coupled receptors (GPCR), ion channels, or transporters on the cell membrane, or to intracellular targets such as enzymes and nuclear hormone receptors [Valentin and Hammond 2008]. The effects of an external stimulus are propagated through a complex biomolecular network of protein-protein interactions, starting at the drug target protein and ending in proteins associated with specific effects of the drug. In order to understand the molecular mechanism of toxicity and to identify toxicity-specific functional modules, network analysis algorithms should be able distill the set of proteins involved in signaling paths that link the drug target proteins to toxicity-related proteins in the PPI network. In this thesis, such paths are defined as **d**rug **t**oxicity **s**ignaling **p**aths (hereinafter "**DTSP"**).

Figure 4 shows a schematic representation of DTSPs for a drug and a toxicity of interest. A **DTSP** is any linear path that links one of the drug's target proteins to a toxicity-associated protein within the PPI network. For a summary of graph-theoretic definitions used in this thesis, refer to Appendix A.

Mathematically, a DTSP can be defined as follows –

Let $D_T = \{t_0, t_1, t_2, \ldots\ldots t_k\}$ be the set of drug target proteins in the protein interaction network. Let $T_p = \{p_0, p_1, p_2, \ldots\ldots p_k\}$ be the set of toxicity-related proteins in the protein interaction network. A Drug-Toxicity Signaling path is a term used to refer to any path that starts at an element of set D_T and ends at an element of set T_p.

$$\text{DTSP} = \{ x_0, x_1, x_2, \ldots\ldots x_k\} \text{ where } x_0 \in D_T \text{ and } x_k \in T_p.$$

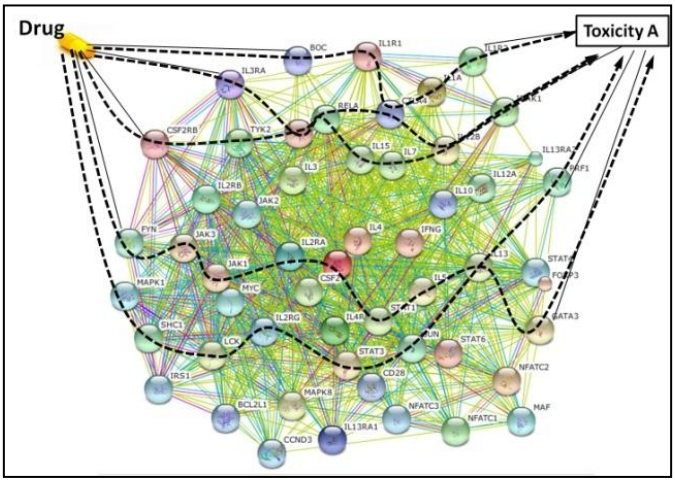

Figure 4. Network model schematic for drug-toxicity signaling paths

As described in section 2.3, automated signaling path detection algorithms and variations thereof have recently been proposed and evaluated for Yeast biomolecular networks [Hüffner et al. 2008, Noga et al. 1995]. Our algorithm extends the algorithm proposed by Steffen et al.[Steffen et al. 2002] to make it amenable to discovery of toxicity-specific proteins inside large protein interaction networks and utilizes the color coding algorithm proposed by Alon et al.[Alon et al. 1995]. Hüffner [Hüffner et al. 2008] implemented the color coding technique as the FASPAD algorithm using dynamic programming, to address the NP-hard problem of discovering signaling paths inside biomolecular networks [Hüffner et al. 2007].

Figure 5 outlines the algorithm pseudo code for DTSP detection. The steps in our algorithm are as follows:

Step 1 - Discover high confidence paths

Various types of evidence may be used to estimate confidence or reliability of each interaction in the network. The first step in the DTSP algorithm involves use of a heuristic approach to harvest a family of high-confidence paths in the PPI network, as shown in Figure 4. The color coding algorithm is adapted to use the drug's protein targets as fixed start nodes and proteins known to be associated with the toxicity as fixed end nodes. Toxicity-related

proteins may be members of pathways associated with the toxic phenomenon and can be identified from published literature as well as pathway annotation databases.

ALGORITHM: DTSPDETECTION($G = (V,E)$, KEGGPaths ToxD, start, dest)

For each Drug D Є ToxD

% *Find high confidence paths: length 3 to 10 for each drug*

For Pathlength = n : m

G_R = FindHiConfPaths(G=(V,E), start, dest , NumPaths , %similarity , Pathlength)

End

% *Compute edge centrality of each interaction in the network of high confidence paths*

I_{EDGE} = ComputeEdgeCentrality(G_R)

% *Find paths with high edge centrality (DTSP candidates)*

P = FindMaxCentralityPaths(G_R , I_{EDGE})

% *Permutation test to evaluate statistical significance of DTSPs*

For each path P_i Є P

For permutations = 1: p

Assign RandomEdgeWeights(P_i)

ComputePathScore(P_i)

End

P_{SIG} = computePval(P_i)

$KEGG_{annot}$ = FindSigPathways(P_{SIG})

End

% *Find statistically significant KEGG Pathway association for DTSPs*

Function $KEGG_{annot}$ = FindSigPathways(P_{SIG})

ES_{path} = CalculateEnrichScore(PathProtAnnotations, KEGGPaths)

$KEGG_{annot}$ = ComputeSig(PathProtAnnotations, ES_{path}, Permutations)

Return $KEGG_{annot}$

Return P_{DTSP} , $KEGG_{annot}$

Figure 5 (continued)

Pseudocode variable definitions:

ToxD - Set of Drugs known to cause the toxicity

D - a drug known to cause the toxicity

start – Set of Drug Target Proteins (start nodes for signaling path detection)

dest – Set of End nodes for the color coding algorithm

NumberPaths – Input parameter for color coding, specifying the maximum number of paths to be discovered.

%similarity – Threshold for the maximum allowed % similarity between two discovered paths.

Pathlength – Input parameter to algorithm, specifying the length of each detected path.

P – Set of high edge centrality paths discovered for a drug.

P_i - a member of the path set P.

P_{SIG} – Set of statistically significant paths or DTSPs.

PathProtAnnotations – KEGG pathway annotations for proteins inside DTSPs

KEGGPaths – Set of canonical KEGG pathways and associated proteins.

ES_{path} – Enrichment Score for a particular KEGG pathway

$KEGG_{annot}$ – Set of KEGG pathways associated with proteins involved in DTSPs.

Figure 5. Algorithm pseudocode for DTSP Detection

As an example, proteins involved in hematopoiesis could be considered as end-nodes when discovering paths relevant to blood disorders.

Edge probabilities represent the strength of evidence for each interaction in the PPI network and detect paths that maximize the product of edge probabilities. Strength of evidence can be based on the amount and type of evidence available to verify the existence of each edge or interaction. Types of evidence may include experimental observation, consensus among

multiple databases, evidence gathered from research literature using text mining or gene-based analysis like gene fusion or neighborhood analysis and approaches similar to those used traditionally as 'guilt by association' (reviewed in chapter 2). The color coding algorithm is then applied to this network, to discover paths that consist of high confidence interactions. The resulting family of high confidence paths is used in the subsequent steps to identify possible drug-toxicity signaling paths, as described in step 2.

Given the fact that the problem of discovering all paths between two nodes in a large network is NP-hard (Non-deterministic polynomial-time hard), a heuristic approach is developed for detection of high confidence paths, as described in section 3.2. The color coding algorithm, as implemented by Hüffner et al. takes various input parameters, including the length of detected paths, the maximum threshold for similarity between two discovered paths, the total number of paths to be discovered and the start/end nodes for detected paths in the network.

Step 2 - Compute edge weights using an appropriate measure

The family of high confidence paths discovered in step 1 constitute theoretical paths, some of which may be actual drug toxicity signaling paths. Others may just be topological interaction chains that link the drug's protein targets to toxicity-related proteins in the network without being actually involved in the mechanism of the toxicity. Identification of drug-toxicity signaling paths from among these high confidence paths, therefore requires a criteria that is

a. Relevant for discovery of drug toxicity signaling paths from a toxicological perspective and

b. Enables mining of the network of high confidence paths to yield protein 'hot spots'.

In order to accomplish this, two novel edge weighting criteria have been proposed and implemented as a part of this thesis. The first approach applies background knowledge of a set

of drugs known to cause the toxicity to discover proteins that are most central to linking a set of drugs to toxicity-related proteins via their respective protein targets.

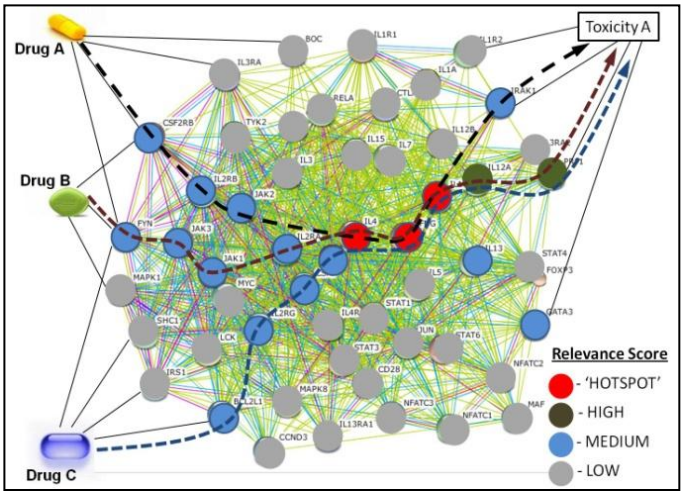

Figure 6. Edges or protein interactions that are highly central to the detected set of paths are assigned a higher relevance score.

This criterion for edge weighting, termed as the edge centrality measure, is described in detail in section 3.2. The second approach uses drug-specific gene expression data to compute edge weights that reflect the extent to which various protein interactions in the high confidence graph involve genes that are highly differentially regulated after drug administration. The gene expression measure and findings from its application have been described in detail in chapter 4.

Each interaction in the smaller network of high confidence paths is weighted using one of the measures described above and the resulting network is used to detect drug-toxicity signaling path candidates in step 3.

Step 3 - Automated Signaling Path (DTSP) Detection

Step 3 in our algorithm uses the network of reliable paths as its starting point, with each edge being assigned a new weight using one of the two measures described in step 2. As shown in Figure 6, the color coding technique is re-applied to this weight graph so as to detect drug toxicity signaling path (DTSP) candidates or paths with a high product of edge weights. The

input parameters to the color coding algorithm remain the same. At this step in the algorithm, the detected path candidates are yet to be evaluated for statistical and biological significance.

Step 4 - Evaluate statistical significance of detected paths

Evaluation of statistical significance for detected network paths has been carried out previously using a randomized network approach [Lindfors et al. 2009]. Briefly, edge weights in our network are randomly shuffled across the network to create a simulated edge weighted PPI network with the same overall distribution of edge weights. The ranking statistic (product of edge probabilities) is then calculated for each path detected in step 3 above. The simulation is repeated 10,000 times to yield a distribution of the statistic for each path. The threshold for statistical significance of $p < 0.01$ is applied after correcting for multiple test comparisons using an appropriate correction measure.

Step 5 - Evaluate the biological relevance of detected paths

Functional attributes of the set of path proteins (common to DTSPs across the set of drugs) and their ability to modulate disease pathways can be analyzed using geneset enrichment methods. Subramaniam et al. have proposed and implemented the geneset enrichment analysis (GSEA) approach, aimed at extracting biological insight from genome wide information [Subramanian et al. 2005]. Briefly, given an *a priori* defined set of genes *S* (e.g. pathways associated with DTSP proteins), the goal of GSEA is to determine whether the members of *S* are randomly distributed throughout *L* (e.g. the set of all proteins in the network) or primarily found at the top (among the most commonly occurring pathway annotations associated with DTSP proteins) or bottom (among the rarely occurring pathways associated with DTSP proteins). The expectation would be that pathways related to the phenotypic distribution (e.g. pathways related to the set of detected drug toxicity signaling paths) will tend to show the latter distribution. As described by Subramaniam et al. [Subramanian et al. 2005], there are three key elements of the GSEA method:

Step 1: Calculation of Enrichment score. The enrichment score is the maximum deviation from zero encountered in a random walk, when a running-time statistic is increased if a gene in *S* (i.e. having a particular functional/pathway annotation) is encountered by walking down the list in *L* (i.e. the list of all common path proteins in DTSPs for drugs known to cause the toxicity) and decreased if the gene encountered is not in *S*. The Enrichment Score (ES) corresponds to a weighted Kolmogorov-Smirnov-like statistic.

Step 2: Estimation of significance level of ES. The statistical significance of ES is estimated using an empirical phenotype-based permutation test procedure. The phenotype labels are permuted and the ES is recomputed for multiple permutations to generate a null distribution. The confidence intervals for this distribution are computed and the actual ES is compared to these thresholds to establish statistical significance.

Step 3: Adjustment for Multiple Hypothesis Testing. When the entire set of pathway annotations is evaluated for statistical significance for all path proteins, the estimated significance level is adjusted to account for multiple hypothesis testing.

Geneset enrichment analysis can point to biological processes that may be modulated downstream from the drug's target proteins as reflected through the identification of proteins involved in DTSPs. Statistically significant enrichment of specific pathways and functional annotations among the set of DTSP proteins may highlight their relevance to the drug-induced toxicity of interest.

3.2 Discovering DTSPs for drug-induced non-immune Neutropenia

3.2.1 Non-immune Neutropenia: considerations for selecting the toxicity under evaluation

In order to evaluate the ability of our algorithm to detect toxicity-specific proteins inside a PPI network, compounds known to induce Neutropenia were considered. Neutropenia resulting from drug administration can be fatal when severe. In the United States, labels of over 40

currently marketed prescription drugs include a warning of a risk for neutropenia and/or agranulocytosis [Multiple 2005]. Neutropenia can be a 'type A' (dose-dependent, typically associated with the drug's mechanism of action) or a 'type B' (not related to pharmacological action, typically an immunological or hypersensitivity reaction) adverse effect [Edwards and Aronson 2000]. The thesis uses non-immune mediated (type A) neutropenia to evaluate the approach because our algorithm leverages information on the drug's target proteins that may be directly associated with the drug's mechanism of action.

3.2.2 Drugs, Drug Targets and Toxicity-related proteins

Our algorithm utilizes existing knowledge about drugs that are known to be associated with the toxicity of interest. Three types of evidence were used to create a list of drugs known to be associated with non-immune neutropenia. First, literature mining was performed on Pubmed abstracts to parse sentences involving neutropenia and a drug term. The drugs were ranked by frequency of co-occurrence with neutropenia or one of its subtypes and the list was manually curated to remove irrelevant associations. Second, the Comparative Toxicogenomics database [Mattingly et al. 2006] was searched to create a list of compounds associated with neutropenia. Third, a toxicologist reviewed the combined list from both analyses to provide a list of drugs associated with non-immune neutropenia. Table 3 shows a list of drugs identified with this approach. Information from the STITCH database was used to identify each drug's target proteins [Kuhn et al. 2008]. STITCH database provides chemical-protein interaction information integrated from multiple public databases, including DrugBank {Wishart, 2006 #104}, SuperTarget and Matador {Gunther, 2008 #127} .

As per our definition of DTSP, proteins associated with the pathophysiology of interest had to be included as 'end' nodes in our path detection algorithm. A review of literature pertaining to hematopoietic regulation reveals the key role of a set of proteins involved in neutrophil production and maturity [Kaushansky 2006, von Vietinghoff and Ley 2008, Daniel et al. 2009].

These proteins are involved in lineage commitment to a specific cell type during the differentiation of hematopoietic stem cells into one of the multiple blood cell types, including platelets, erythrocytes, monocytes, neutrophils, T cells, NK cells and B cells. In addition to their relevance to neutrophil biology, some of these proteins (e.g. CFU-GM) have recently been proposed as possible early markers of bone marrow toxicity [Pessina et al. 2005].

The proteins listed in Table 4 were therefore considered as 'end' nodes in the PPI network.

3.2.3 Rationale for use of an Edge Centrality Measure

Consider a hypothetical situation where all high confidence paths discovered in step 1 of the algorithm have exactly one interaction in common. This interaction could be considered the most central to achieving topological connectivity between the drug's targets and the toxicity of interest because all reliable paths discovered in step 1 involve this protein interaction. The interacting proteins would be of interest from a biological perspective because their function could yield insights into toxic mechanisms that link the drug to its adverse effect. Edge centrality is important from a biological perspective because interactions common to high confidence paths across toxicity-inducing drugs may point to common mechanisms of toxicity.

TABLE 3. Protein 'end nodes' for path detection.

Ensembl Protein ID	Description	HGNC symbol
ENSP00000225474	Granulocyte colony-stimulating factor Precursor	CSF3
ENSP00000228280	Kit ligand Precursor (C-kit ligand)(Stem cell factor)	KITLG
ENSP00000231454	Interleukin-5 Precursor (IL-5)(T-cell replacing factor)	IL5
ENSP00000296871	Granulocyte-macrophage colony-stimulating factor Precursor (GM-CSF)	CSF2

Edge centrality in this context, is defined as the extent to which a particular edge is involved in all drug-toxicity signaling paths for a particular drug. As a general case across many drugs that cause the same toxicity, edge centrality of each interaction can be computed to identify those edges in the network that are most central to linking the drug's targets to its toxicity of interest, as shown in Figure 7. Edge centrality of an edge ***e*** is defined as the number of reliable paths that pass through ***e***. Each edge in the set of high confidence paths is assigned a weight equal to its edge centrality.

TABLE 4. Drugs associated with Non-immune Neutropenia

Drug	Therapeutic category[DrugBank]
Vancomycin	Anti-Bacterial Agents , Glycopeptide antibacterials
Zidovudine	Anti-HIV Agents, Antimetabolites, Nucleoside and Nucleotide Reverse Transcriptase Inhibitors, Reverse Transcriptase Inhibitors
Econazole	Antifungal Agents
Miconazole	Antifungal Agents
Ketaconazole	Antifungal Agents
Isoniazid	Antitubercular Agents, Fatty Acid Synthesis Inhibitors
Rifampicin	Antitubercular antibiotics, Antituberculosis Agents,Enzyme Inhibitors,Leprostatic Agents,Nucleic Acid Synthesis Inhibitors
Clotrimazole	Local Anti-Infective agents,Antifungal Agents, Growth Inhibitors
Doxorubicin	Antibiotics, Antineoplastic Agents
Methotrexate	Abortifacient Agents, Abortifacient Agents, Nonsteroidal Antimetabolites, Antimetabolites, Antineoplastic Agents, Antirheumatic Agents, Dermatologic Agents, Enzyme Inhibitors, Folic Acid Antagonists, Immunosuppressive Agents, Nucleic Acid Synthesis Inhibitors
Etoposide	Antineoplastic Agents, Phytogenic Nucleic Acid Synthesis Inhibitors
Irinotecan	Phytogenic Antineoplastic Agents, Enzyme Inhibitors, Parasympathomimetics, Prodrugs, Radiation-Sensitizing Agents
Paclitaxel	Phytogenic Antineoplastic Agents, Tubulin Modulators
Tamoxifen	Hormonal Antineoplastic Agents, Bone Density Conservation Agents, Estrogen Antagonists, Selective Estrogen Receptor Modulators
Vinblastine	Phytogenic Antineoplastic Agents, Tubulin Modulators
Promazine	Antiemetics, Antipsychotic Agents, Antipsychotics, Dopamine Antagonists, Neuroleptics, Phenothiazines

Table 4 (continued)

Desferrioxamine	Chelating agent, Iron Chelating Agents, Siderophores
Furosemide	Diuretics, Sodium Potassium Chloride Symporter Inhibitors
HCTZ	Antihypertensive Agents, Diuretics, Sodium Chloride Symporter Inhibitors

This smaller network of reliable paths is then used to detect drug-toxicity signaling paths. In step 2 of our algorithm the edge centrality of each interaction/edge in the family of high confidence paths is computed.

3.2.4 DTSP algorithm implementation

The color coding algorithm was adapted so that it could be applied towards discovery of signaling paths relevant to the toxicity of interest. Toxicity relevance for detected paths was achieved through use of hematopoiesis-related proteins (shown in Table 3) as destination nodes in the algorithm. Drug relevance for detected paths was achieved through use of drug target proteins as 'source' nodes in the algorithm. The primary source of protein interaction data was the STITCH database (http://stitch.embl.de/). All protein-protein interactions from the STITCH database were downloaded and the human subset was used as our PPI network.

Details on the STITCH database and calculation of combined score for confidence of each edge have been published elsewhere [Kuhn et al. 2008]. Briefly, a consolidated set of chemicals was derived from PubChem and relations between chemicals were derived from similar activity profiles in the NCI60 cell lines, from pharmacological actions assigned to chemicals in the Medical Subject Headings (MeSH) and from the literature. In order to link the derived chemical-chemical associations to the protein world, a variety of databases of chemical-protein interactions were imported. Experimental evidence of direct chemical-protein binding was derived from the PDSP K_i Database and the protein data bank [Berman et al. 2000]. The STITCH database contains interaction information for over 68000 chemicals,

including 2200 drugs, and connects them to 1.5 million genes across 373 genomes and their interactions contained in the STRING database [Jensen et al. 2009].

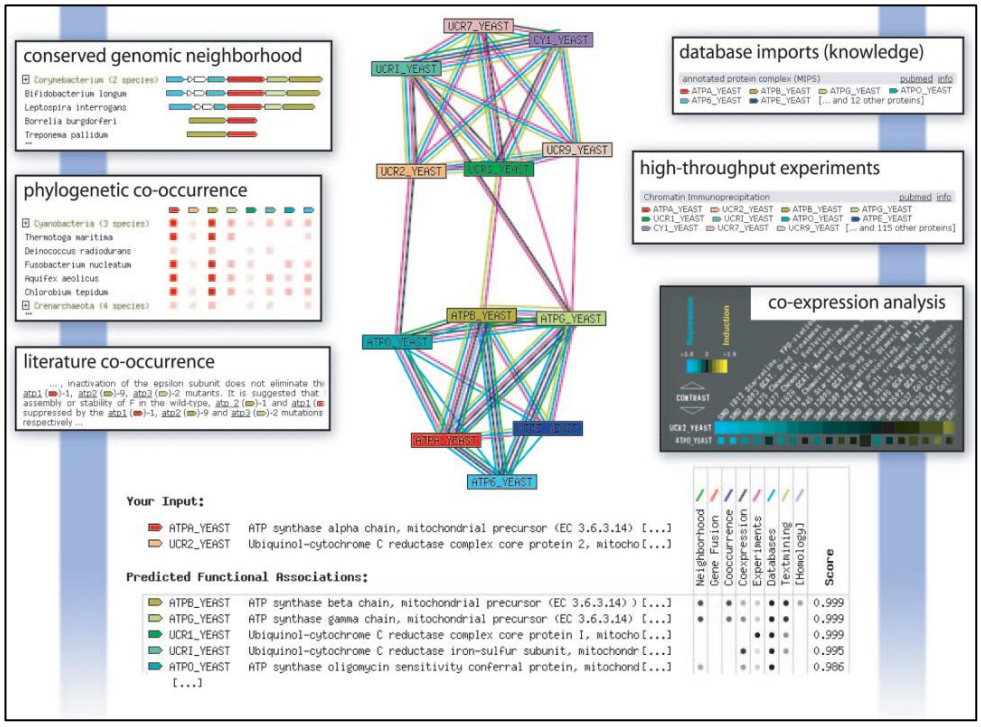

Figure 7. STRING Database: Evidence types used to confidence score for each interaction (reproduced from Jensen et al. 2009 [Jensen et al. 2009])

Many of the protein-protein interactions in the STRING database are imported from other databases but STRING also contains a large body of predicted associations that are produced *de novo*. Completely sequenced genomes are periodically imported and searched for three types of genomic context associations:

a. **Conserved genomic neighborhood**: The conservation of proximity of genes along the genome between distantly related species may predict interaction between the protein products {Dandekar, 1998 #581}.

b. **Gene fusion events**: Some interacting protein pairs are encoded by two independent genes in some organisms whereas they are encoded by a single gene in other organisms.

Knowledge of gene fusion events may therefore be used to ascertain whether two proteins may be interacting with each other [Valencia and Pazos 2003].

c. **Co-occurrence of genes across genomes**: Similar Loci for genes corresponding to a protein pair across genomes may point to a possibility of interaction.

All three searches aim to identify pairs of genes which appear to be under common selective pressures during evolution (more so than expected by chance), and which are therefore thought to be functionally associated. Another important source of protein association information is published literature. As shown in Figure 8, STRING database assigns confidence score to interactions based on systematic extraction of associations from PubMed, by searching for recurrent co-mentioning of gene names in abstracts. This search relies on gene names and synonyms parsed from SwissProt as well as from organism-specific databases. It utilizes a benchmarked scoring system based on the frequencies and distributions of gene names in abstracts. Finally, protein-protein interactions are also derived from functional genomics data: evidence of co-regulation of genes across diverse experimental conditions is imported from the ArrayProspector server [Jensen et al. 2004].

The command-line version of FASPAD implementation [Hüffner et al. 2007] was used to discover high confidence paths in the PPI network using a heuristic approach. As a first step in the algorithm, signaling paths were detected with maximum product of path reliabilities for path lengths up to 10. The rationale for selection of a maximum path length of 10 was based on the observation that the number of new proteins involved in discovered paths decreased drastically as the path length approached 10, as shown in Figure 9. For each path length, 100 high confidence paths were harvested to detect a total of 800 candidate signaling paths for each drug. FASPAD parameters were set so that there was no more than 70% similarity between two paths of the same length. For the second step in our algorithm, the network created from high confidence paths was used to compute the edge centrality value for each interaction. As

described in section 3.1, the edge weights were then re-assigned to each interaction using its edge centrality value, so that the value on each interaction represented the extent to which the interaction was central to paths that connect the drug target proteins to toxicity-related proteins in the network. The third step in the algorithm involved detection of paths with high edge central interactions, or the set of candidate 'Drug – Toxicity signaling paths' for each drug. The color coding algorithm was applied again, with the same parameters as earlier but with edge weights reassigned based on edge centrality calculation for each interaction.

The resulting set of signaling paths was then evaluated for statistical significance as described in section 3.1. The product of edge weights for each path was used as the test statistic. For 20,000 random permutations of edge weights inside the network, the product of edge weights for each discovered path was calculated. Paths were then filtered using a $p < 0.01$ threshold and only paths with statistically significant value were retained. This process was repeated for each drug listed in Table 4. Finally, Geneset Enrichment analysis (GSEA) implemented in WebGestalt [Zhang et al. 2005] was run to identify significant KEGG pathway associations for DTSP proteins.

Briefly, the procedure calculated a ratio enrichment score for each KEGG pathway as k/k_e, where k was the actual number of genes in the DTSP set of proteins that belonged to a particular KEGG pathway and k_e is the expected number of genes for that pathway. Using a statistical significance threshold of $p<0.01$ for hypergeometric test and Bonferroni multiple testing correction, a set of canonical pathway associations for each drug were identified.

3.3 Results and Discussion

3.3.1 Impact of the signaling path detection approach

A key motivation for this approach was to evaluate whether the huge network of human protein-protein interactions can be condensed to a small set of proteins that may constitute a 'bottleneck' linking drug targets with the toxicity-related end nodes in the network. In order to

evaluate this, we compared the total number of immediate neighbors of our end nodes with the number of immediate neighbors that were involved in at least one DTSP. At level 1 (immediate neighborhood of toxicity-related proteins), we found 2337 proteins connected to the four 'end nodes' related to non-immune neutropenia. All of these immediate neighbors could theoretically be involved in paths terminating in end nodes. However, for the non-immune neutropenia causing drugs analyzed in this study, only 227 proteins out of the 2337 were found to be involved in a drug-toxicity signaling path. Within the set of immediate neighbors for toxicity-related proteins, these 227 proteins appeared to be more relevant to understanding the mechanism of non-immune neutropenia compared to the rest. Taking this observation to the next level of connectivity, out of 16339 proteins that were linked to the 2337 immediate neighbors, our algorithm found 270 proteins in DTSPs detected for drugs analyzed. The value of our approach was evident from its ability to identify a smaller set of proteins that were relevant to both the toxicity of interest as well as the drugs known to be associated with the toxicity.

3.3.2 Path length heuristic and discovery of new proteins

One important consideration in our algorithm was the choice of path length when harvesting signaling paths. As the path length increased towards 10, longer paths seemed to be created using proteins involved in shorter paths. This is demonstrated in Figure 9, where the number of new proteins involved in discovered paths decreases substantially as the path length increases. To the extent that the algorithm is aimed at identification of protein/s relevant to the toxicity, there seems to be limited value in detecting paths of higher lengths. This is an aspect of this algorithm that needs to be investigated in more detail.

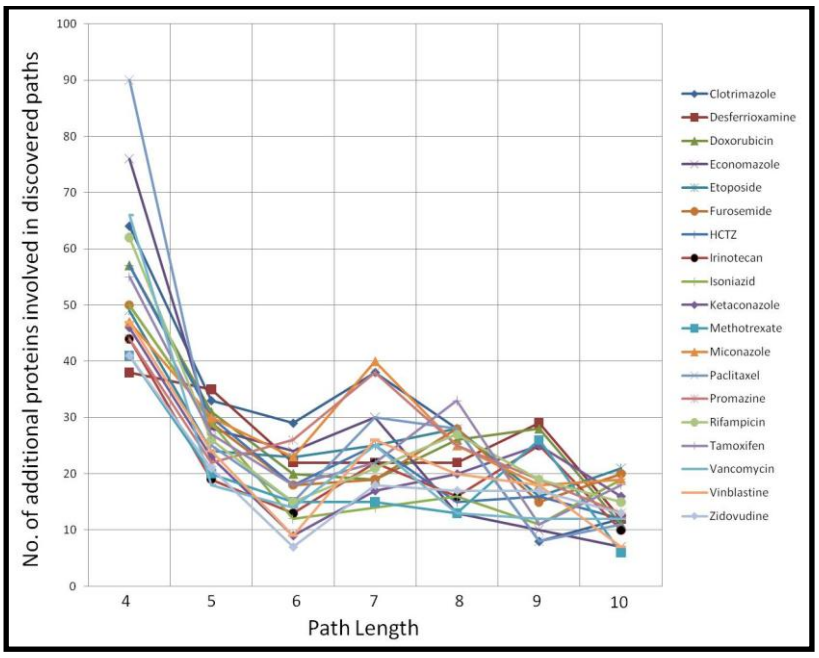

Figure 8. As the path length increases towards 10, the number of discovered paths involving additional proteins decreases towards zero

3.3.3 Network 'hotspots' and their toxicological relevance

As conceptually visualized in Figure 7, proteins involved in at least one DTSP for all drugs get a high interestingness score due to their ability to provide topological connectivity between a target protein and a toxicity-related protein for all drugs that are known to be associated with the toxicity. We therefore analyzed the list of DTSPs to find proteins that were common across all drugs analyzed in this study. As shown in Figure 10, 119 proteins were involved in a DTSP for a single drug, 12 proteins were common to 4 Drugs (i.e. involved in one or more DTSPs for four drugs) and 9 proteins were found to be involved in one or more DTSPs for all drugs. Table 5 lists the nine proteins that occur in one or more DTSPs across all drugs.

A brief summary of their relevance to non-immune neutropenia is provided below –

A genetic variant of CSF2RB (Ensembl id: ENSP00000262825) has been patented in its application as a marker associated with adverse hematological response to drugs [Athanasiou and GERSON 2006]. GM-CSF exhibits a number of overlapping biological activities in

hematopoiesis, which are all mediated via binding of GM-CSF to the GM-CSF receptor [Barreda et al. 2004]. The STAT3 (signal transducer and activator of transcription 3, acute-phase response factor) protein (Ensembl id: ENSP00000264657) plays a key role in apoptosis and cell differentiation. Deletion of STAT3 in rats has been shown to cause abnormalities in myeloid cells [Welte et al. 2003].

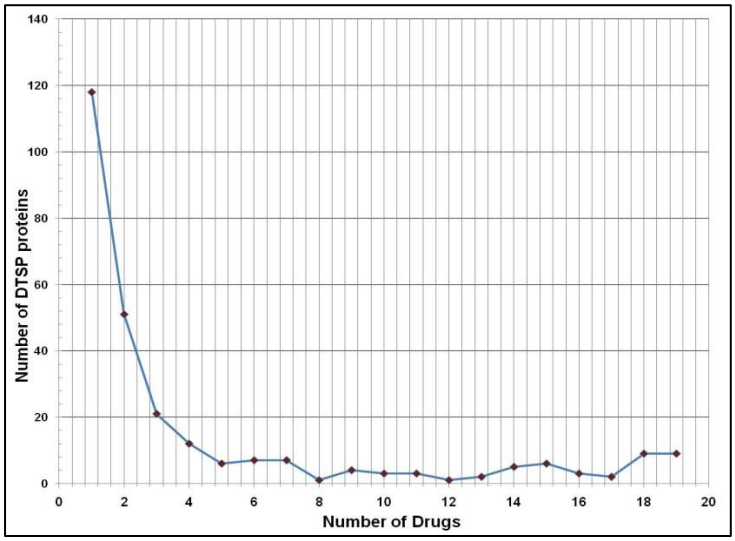

Figure 9. Plot shows the number of proteins that occur at least once across various drugs in the analysis set.

The Janus Kinase JAK1 (Ensembl id: ENSP00000294423) and JAK2 (Ensembl id: ENSP00000371067) pathways in addition to the STAT pathway constitute a major signaling mechanism engaged by the G-CSFR (Granulocyte Colony Stimulating Factor) receptor. When activated by the GCSFR, Jak tyrosines phosphorylate STAT complexes which then translocate to the nucleus where they activate transcription [Touw and Bontenbal 2007]. The Tyrosine protein phosphatase non-receptor type 6, PTPN6 protein (Ensembl id: ENSP00000326010) is a key regulator of neutrophil apoptosis [Simon 2003]. The growth factor receptor bound protein 2 (Ensembl id: ENSP00000339007) functions as an adapter protein in the MAPK transduction pathway which plays a key role in hematopoiesis [Geest and Coffer]. The SHC-transforming

protein 1 (Ensembl id: ENSP00000357432) is another signaling adapter that couples activated growth factor receptors to MAPK signaling pathway.

The KIT stem cell precursor (Ensemble id: ENSP00000370749) is the receptor of the stem cell factor that is directly involved in myeloid cell differentiation. Finally, the Tumor Necrosis Factor precursor (Ensemble id: ENSP00000372790) is a cytokine that binds to tumor necrosis factor (TNF) receptor. TNF may have inhibitory effects on granulocyte-macrophage progenitors and on committed and primitive hematopoietic progenitors in vitro [Drutskaya et al. 2005].

TABLE 5. List of proteins that occur in at least one DTSP across all analyzed Drugs.

Protein Ensembl Id	HGNC Symbol	Name	Biological relevance
ENSP00000262825	CSF2RB	Colony Stimulating Factor Receptor, Beta	Barreda et al. 2004 [Barreda et al. 2004]
ENSP00000264657	STAT3	Signal Transducer and Activator of transcription	Welte et al. 2003[Welte et al. 2003]
ENSP00000294423	JAK1	Janus Kinase 1	Touw et al. 2007[Touw and Bontenbal 2007]
ENSP00000326010	PTPN6	Tyrosine Protein phosphatase type 6 (HCP, PTP1C)	Simon 2003[Simon 2003]
ENSP00000339007	GRB2	Growth Factor Receptor Bound 2	Geest et al. 2009[Geest and Coffer 2009]
ENSP00000357432	SHC-1	SHC-transforming protein 1	Geest et al. 2009[Geest and Coffer 2009]
ENSP00000370749	KIT precursor	Stem Cell precursor	Kaushanky 2006[Kaushansky 2006]
ENSP00000371067	JAK2	Janus Kinase 2	Touw et al. 2007[Touw and Bontenbal 2007]
ENSP00000372790	TNF precursor	Tumor Necrosis Factor precursor	Drutskaya et al. 2005[Drutskaya et al. 2005]

3.3.4 Elucidating class-specific mechanisms of toxicity

In order to understand whether proteins involved in paths discovered for one class of drugs were different from another, we compared proteins common to anti-infective and anti-cancer DTSPs i.e. proteins involved in one or more DTSPs for all drugs that belong to the same therapeutic class. A set of 12 proteins were found to be involved in one or more DTSPs for all anti-cancer drugs. A different set of 10 proteins were found to be involved in one or more DTSPs for all anti-infective drugs. Geneset enrichment analysis revealed distinct biological processes being regulated by each set. We found that the 'anti-infective' set of proteins were also involved in biological processes like response to bacterial stimulus. This observation seems to imply that anti-infective mode of action may be the desirable action for drugs in this class, but proteins involved in anti-infective related processes also link the drug targets and/or the indication to hematopoiesis-related proteins in the network. The observation also seems important in light of the fact that manifestations of agranulocytosis are secondary to infection [Flanagan and Dunk 2008].

3.3.5 Pathway Association for DTSP proteins

The robustness of a phenotype may be understood in terms of alternative compensatory signaling routes inside complex biomolecular networks. To this end, proteins involved in the set of DTSPs for each drug may yield insights into downstream effects of the drug on multiple biological and disease pathways. Geneset enrichment provided a set of KEGG pathways with high ratio enrichment score for each drug. Using this approach for each drug, we identified a set of KEGG pathways that were statistically overrepresented in DTSPs for that drug. Table 6 shows a list of KEGG pathways that were overrepresented for all drugs in this analysis.

A key observation from enrichment analysis was that proteins associated with cancer-related pathways were found within DTSPs discovered for all drugs associated with non-immune neutropenia in this analysis. This finding applied to all drug classes analyzed with our

algorithm, including anti-cancer, anti-infective as well as other disease-related drugs. Though an observation made on a small number of drugs, this is in line with the clinical observation that cytotoxic cancer treatments predispose patients to clinically significant changes in neutrophil counts, often leading to severe neutropenia [Crawford et al. 2004]. Neutrophils being the first line of defense against infection and as the first cellular component of the inflammatory response to nascent infections, it is conceivable that proteins involved in inflammatory response and immune system disorders are also members of drug-toxicity signaling paths discovered using our algorithm.

TABLE 6. Pathway Association for DTSP proteins

KEGG pathway Id	KEGG Pathway Name
hsa05221	Acute myeloid leukemia
hsa04920	Adipocytokine signaling pathway
hsa04662	B cell receptor signaling pathway
hsa04062	Chemokine signaling pathway
hsa05220	Chronic myeloid leukemia
hsa05210	Colorectal cancer
hsa04060	Cytokine-cytokine receptor interaction
hsa05213	Endometrial cancer
hsa05120	Epithelial cell signaling in Helicobacter pylori infection
hsa04664	Fc epsilon RI signaling pathway
hsa04510	Focal adhesion
hsa05214	Glioma
hsa04910	Insulin signaling pathway
hsa04630	Jak-STAT signaling pathway
hsa04010	MAPK signaling pathway
hsa04650	Natural killer cell mediated cytotoxicity
hsa04722	Neurotrophin signaling pathway
hsa05223	Non-small cell lung cancer
hsa05212	Pancreatic cancer
hsa05200	Pathways in cancer
hsa05215	Prostate cancer
hsa05211	Renal cell carcinoma
hsa04660	T cell receptor signaling pathway
hsa04620	Toll-like receptor signaling pathway
hsa04930	Type II diabetes mellitus

3.4 Limitations

This study was aimed at understanding downstream regulation of drug-induced toxicity at a 'systems' level. Improvements on the proposed algorithm are certainly possible. First, we mined a non-directional PPI network for signaling paths. This implies that directionality of the interaction chains cannot be inferred from this analysis. Some background knowledge of the signaling cascade from the CSF receptor to activation of Colony Stimulating factor (CSF) suggests that proteins like JAK1, JAK2 and STAT3 are mediators of CSF activation but this cannot be concluded from our analysis. Second, we have used a range of path lengths and the maximum number of detected paths as parameters.

In Figure 9, we have shown that the choice of heuristic path length seems appropriate. However, the choice of maximum paths discovered needs further investigation.

Further analysis could reveal the extent to which the set of detected paths covers the available topological space. Further work is also needed to incorporate additional evidence, gene expression data, for example, to identify putative signaling paths from among the larger set we have discovered through this analysis. Integration with gene expression data may also reveal 'dose-specific' paths that reflect pharmacodynamic changes after drug administration. We believe that the use of multiple sources of evidence to assign interaction confidence scores may have resulted in fewer false negatives and that, integration with gene expression data may also reduce the potential for false negatives in our analysis. Finally, a large set of drugs known to induce non-immune neutropenia needs to be analyzed, preferably from a wider variety of therapeutic categories, to understand the specificity and sensitivity of our approach.

3.5 Conclusions

This study explored downstream effects common to drugs known to induce non-immune neutropenia. Using our approach, we have identified proteins that constitute 'bottlenecks' in interaction chains that link a neutropenia-inducing drug with toxicity-related endpoints. Further

analysis is required to identify pathways involved in pathophysiology of non-immune neutropenia for various drug classes. The edge centrality based approach provides an alternative to other approaches that can be applied at the systems level, either using local or global properties of PPI networks. We have demonstrated the value of our approach using non-immune neutropenia as a test case. The algorithm may be applied towards detection of protein interaction hotspots for any toxicity where the drugs known to induce the toxicity and the physiological processes involved in the toxic mechanism are known.

CHAPTER 4: INTEGRATING PROTEIN INTERACTION AND GENE EXPRESSION INFORMATION TO GAIN INSIGHTS INTO TOXICITY MECHANISMS

In chapter 3, the drug-toxicity signaling path (DTSP) algorithm was used to identify toxicity-related hotspots inside protein interaction networks through a combination of color coding and edge centrality as a measure for filtering detected paths [Desai et al. 2011]. The approach relied on background knowledge of a set of drugs and proteins related to the drug-induced toxicity. One of the recognized limitations of the edge-centrality based approach is its inability to leverage dynamic drug-specific gene expression data to further inform the filtering of detected signaling paths. This chapter extends the earlier approach by incorporating a gene expression-based measure for filtering discovered paths based on analysis of gene expression data collected *in vitro* after treatment with the set of drugs under consideration. The resulting set of path proteins is analyzed, leading to a biomarker panel that may be used for screening of drug candidates.

All steps involved in the drug-toxicity signaling path detection algorithm as well as implementation of the algorithm for non-immune neutropenia have been described in section 3.1 and 3.2. These details will therefore, not be repeated in this chapter. The advantages of using a gene-expression based measure are described in section 4.1 and implementation of the measure is described in section 4.2. Section 4.3 discusses results from the analysis, followed by Limitations and Conclusions in section 4.4 and 4.5 respectively.

4.1 Advantages of using a gene-expression based measure

As described in chapter 2, contextual information in the form of evolutionary conservation of expression patterns across organisms when used for guilt by association (GBA) analysis, has yielded functionally related groups of genes under homeostatic conditions [Quackenbush 2003]. Gene co-expression based GBA approaches have also looked for genes whose

expression patterns mimic those of known disease-associated genes[Walker et al. 1999]. While such approaches are useful for assigning disease association more generally, they may not be sufficient for identifying proteins associated with drug-induced toxicity. This is because genes discovered using such approaches may be located multiple levels downstream from protein targets of drugs associated with a particular toxicity.

Integrating gene expression data with protein interaction networks has many advantages. First, to the extent that gene expression data can reflect dynamic changes that happen at various time points after drug administration, this could enable identification of signaling paths that consist of genes that are differentially expressed between treated and control conditions. Second, clustering expression data into groups of genes that share profiles is a proven method for grouping functionally related genes, but does not order pathway components according to physical or regulatory relationships [Steffen et al. 2002]. Protein Interaction information can complement gene expression data to help identify differentially expressed genes that may lie downstream from the drug's known target proteins and upstream from the known toxicity-related endpoints in the network.

Commercial pathway analysis tools provide the ability to construct networks based on integration of evidence from multiple sources and the ability to analyze the resulting network using algorithms like shortest paths [Ekins 2006]. Drug administration perturbs the biological network at specific 'nodes' (referred to as drug target proteins) that may be several levels upstream from the toxicity-related 'endpoints' in the network. Mapping the cascade of proteins associated with drug response is therefore a central challenge in toxicity evaluation.

ALGORITHM: INTEGRATEEXPRESSION(*G* = (*V*,*E*), ToxD, startVert, dest)
For each Drug D ∈ ToxD

% Find high confidence paths: length 3 to 10 for each drug

For Pathlength = *n* : *m*

G_R = FindHiConfPaths(G = (*V*,*E*),startVert,dest, Paths ,%similarity, Pathlength)

End

% Compute Gene Expression Score for high confidence interactions (based on gene expression for interacting proteins)

PathExpScore = ComputeIntExpScore(G_R , D)

% Detect signaling paths that maximize path differential expression score

P = FindMaxDiffExpPaths(G_R , PathExpScore)

For permutations = 1: *p*

For each interaction P_{int}

Assign RandomEdgeWeights(P_{int})

ComputePathScore(P)

End

End

P_{SIG} = computePval(P)

% Compute common pathproteins for each therapeutic category

$P_{cancerprot}$ = intersect ($Prots_{PSIG}$)

Return P_{SIG} , $P_{cancerprot}$

Pseudocode variable definitions (also refer to other variables defined in Figure 5):

P – Set of paths with highest differential gene expression between treated and control.

$P_{cancerprot}$ – Path proteins found on DTSPs for anti-cancer treatments.

Figure 11 (continued)

Function ComputeIntExpScore(G_R , D)

% Extract and Process Raw gene expression CEL files for n Treated samples and % m vehicle scans for each Treated sample. Compute Average Expression Score for % each probeset and each protein. Finally, compute interaction expression score.

RawExp = GetCEL(D)

GCTExp = ExpressionFileCreator(RawExp,'QuantileNorm','medianScale', HgU.cdf)

For probesets = 1: p_n

For each Experiment *i*

$$\text{SampleExpScore}(i) = \frac{treatedExpi}{(\sum_{i=1}^{ni} vehicleExpi)/ni}$$

Return SampleExpScore

$$\text{ProbesetExpScore} = \ln(\sum_{i=1}^{ns} SampleExpScore/ns)$$

End

For each Path p in G_R

For each protein *i* in p

$$\text{ProtExpScore(i)} = \sum_{i=1}^{np} ProbesetExpScore_i/n_p$$

End

$$\text{IntExpScore(p)} = \sum_{i=1}^{nprot} ProtExpScore_i/nprot$$

Return IntExpScore

Return IntExpScore

Pseudocode variable definitions:

RawExp – Raw CEL file data for a single sample.

GCTExp – Matrix of probeset-Sample with expression values inside each cell.

SampleExpScore – Ratio of treated/control expr. values for all HT_HGU133 probesets.

ProbesetExpScore – Log ratio of average gene expression score across all samples

ProtExpScore – Average gene expression score for a protein based on mapped probesets.

IntExpScore – Interaction expression score, i.e. average differential expression score for interacting proteins

Figure 10. Algorithm Pseudocode for detecting Drug-Toxicity Signaling Paths using microarray data

4.2 Computing Edge weights using Gene Expression Measure

Gene expression was used to identify paths that may be 'active' in the presence of a drug known to induce the toxicity. A differential gene expression measure was applied as follows –

We hypothesized that the most relevant toxicity-related paths would consist of genes that showed high differential expression between treatment (with drugs associated with the toxicity) and control. In other words, if all proteins in a detected path consisted of genes that were unperturbed after treatment, that path was unlikely to be associated with the drug and its toxicity and would therefore be ranked lower with our algorithm. Also, the algorithm should consider the magnitude of up-regulation and down-regulation as equally important. So, a 2-fold up-regulation would be considered as important as a 2-fold down-regulation. We therefore, using principles inherent in a ranking measure proposed earlier [Zhang and Gant 2008], calculated the average absolute log ratio of treated vs. control for each protein in the set of reliable paths to assign weights for each interaction based on differential gene expression of proteins participating in that interaction. This smaller network of reliable paths was then used to detect drug-toxicity signaling paths in step 3 of the algorithm (described in section 3.1).

To compute the gene expression measure, we downloaded gene expression CEL files from the connectivity map data provided by the MIT Broad Institute. Details of the connectivity map data have been published elsewhere [Lamb 2007]. Briefly, the connectivity map (also known as cmap) is a collection of genome-wide transcriptional expression data from cultured human cells treated with bioactive small molecules. Build 02 of this publicly available collection provides access to more than 7000 expression profiles representing 1309 compounds. For the 19 drugs analyzed in this study, we found 41 instances or experiments in the connectivity map collection, i.e. one or more treated samples (dose: 10µM, duration: 6 hours) and multiple vehicle scans in MCF7 cell lines. The 'ExpressionFileCreator' module in GenePattern was used to convert CEL files into gct format expression files [Reich et al. 2006]. Robust multichip

average (RMA) measure was used for conversion and quantile normalization was applied. The absolute average log ratio values were calculated for each probeset and each ensemble protein was assigned the average of all probesets that mapped to the protein. Mapping between affy probeset ids and ensemble ids was obtained from Ensembl database [Flicek et al. 2008]. Each edge or interaction in the network of reliable paths was then assigned the expression measure value as the average of absolute log ratio value for its interacting proteins.

The third step in the DTSP algorithm involved detection of paths consisting of high differential expression interactions, or the set of candidate 'Drug – Toxicity signaling paths' for each drug. We used the color coding algorithm again, with the same parameters as earlier but with edge weights assigned based on the gene expression measure for each interaction.

4.3 Results and Discussion

4.3.1 Difference in path proteins discovered with different measures

In order to compare DTSPs discovered using differential gene expression with DTSPs discovered using edge centrality (published earlier [Desai et al. 2011]), the percentage difference in number of proteins involved in DTSPs obtained using the two measures was computed. An average of 45% proteins was different between the two measures. This finding seems to point to the fact that edge central proteins may not be substantially differentially expressed under treatment conditions. This may be attributable to the presence of alternative compensatory paths in protein interaction networks and the fact that highly edge central interactions by definition are likely to receive positive as well as negative regulatory inputs from a relatively large number of interactions.

4.3.2 Toxicological relevance of detected paths

As neutropenia is defined as a clinically significant reduction on neutrophils and as the end nodes (listed in table 3) in our algorithm are involved in neutrophil production, we considered paths that down-regulated the **end nodes** (treated < control) as more relevant than filtered paths

that up-regulated the **end nodes** (treated > control), irrespective of the direction of regulation for intermediate nodes (nodes between the two end nodes for each path). Also, to eliminate redundancy among down-regulated paths, we manually curated the list to identify the shortest path/s that down-regulated each end node. In other words, if there was one shorter path and one longer path that terminated in the same end node, we only considered the shorter path. The paths can be used to understand the relevant mechanisms of toxicity for a specific drug. For example, the highest ranked path for Clotrimazole may be interpreted as follows -

In tissue samples treated with Clotrimazole, Cyp3A4 (node 1) inhibition may lead to down-regulation of SRGN (node 2), the serglycin haematopoeitic cell granule proteoglycan which in turn, may serve as a mediator of granule mediated apoptosis [Ma et al. 2008]. SRGN down-regulates the end node 'KITLG', stem cell factor (end node) that is able to augment the proliferation of both myeloid and lymphoid hematopoietic progenitors in bone marrow cells [Niemann et al. 2004].

The above path was one of four paths for Clotrimazole, the other four having length four and ending in the second down regulated end node, IL5.

The entire list of specific paths for each drug is pending further interpretation in the context of non-immune neutropenia. Table 7 shows a list of path proteins for each drug.

4.3.3 Elucidating class-specific mechanisms of toxicity

In order to understand whether proteins involved in paths discovered for one therapeutic category of drugs were different from another, we compared proteins common within and between (involved in one or more DTSPs) anti-infective, anti-cancer and Other DTSPs. Only KITLG was common (>50%) within the three therapeutic categories suggesting some commonality of compounds that induce neutropenia irrespective of therapy area. In addition, the following were common within a therapy area: IL5 and TNFa in anti-infective, BAX, CSF3, IL6 and IL8 in anti-cancer and FOS, IL6, IL8, JAK and JUN within other.

4.3.4 Potential neutropenia biomarkers

In addition to identifying common DTSPs, the data within Table 8 was evaluated to identify a potential neutropenia biomarker panel irrespective of therapy area. KITLG was altered in 76% of all compounds evaluated and could be further investigated as a neutropenic biomarker. Of the compounds where KITLG was not altered, TNFA was altered in the anti-infective compounds and CSF3 (GM-CSF) in anti-cancer compounds. Therefore, a biomarker panel containing at least KITLG, TNFA and CSF3 should be further investigated for detecting compounds that induce neutropenia.

4.4 Limitations

The DTSP algorithm and use of gene expression data has provided additional insights into signaling paths that may be down regulated in non-immune neutropenia. Other extensions may improve the algorithm further. Specifically, knowledge on protein localization and tissue specific protein expression may add further specificity to detected paths. Experimental validation of the proposed biomarker panel may provide evidence to support the use of the panel for screening drug candidates.

4.5 Conclusion

We have developed a new algorithm for detection of toxicity-related hotspots inside protein interaction networks and evaluated its ability to detect proteins relevant to non-immune neutropenia. This work has demonstrated that better network analysis algorithms can provide valuable insights into mechanisms of toxicity and at the same time, deliver potential biomarkers that may be experimentally validated and used as a tool for screening drug candidates.

TABLE 7. Proteins involved in DTSPs discovered for each drug

Category	Drug	Path Proteins (HGNC Symbol)
Anti-infectives	Clotrimazole	CD44, CYP3A4, GZMB, IFNG, IL5, IL11, KITLG, PRF1, SRGN
	Desferrioxamine	CSF3, EPO, EPOR, FOS, IFNG, IL1B, IL5, IL6, IL8, JAK1, JUN, KIT, KITLG, MAPK11, MAPK14, RELA, SELP, STAT1, STAT6, TNFA
	Economazole	ADCYP1, FOS, TNFA
	Isoniazid	BOC, CA3, JAK2, KITLG, MPO, PTPN6
	Ketaconazole	ABCB1, BCL2, CD79A, CCL2, CSF3, CYCS, EP300, EPOR, IL1A, IL2, IL4, IL5, IL6, IL6ST, IRS1, ITGAM, JAK1, JAK2, KIT, KITLG, LIF, MAPK8, PIK3R1, PRL, RELA STAT1, TNFA, TP53
	Miconazole	BAX, CSF3, HMOX1, IFNG, KITLG, POR
	Rifampicin	ABCB1, BCL2, CCND1, CCND3, CSF2, CYP1A2, GST3, IFNG, IL2, IL2RA, IL5, IL7, IL8, JAK1, JUN, MAPK1, MAPK8, NR3C1, NR0B2, RBL2, STAT1, STAT3, STAT5B, TNFA
	Vancomycin	CCL5, CD4, IL1A, IL5, IL6, IL10, IL11, IL12B, JUN KITLG, STAT3, TNFA
	Zidovudine	CD4, EPO, FOS, IL5, IL10, IL12A, JAK2 KITLG, STAT3
Anti-cancer[1]	Doxorubicin	BCL2, CASP8, CSF3, FADD, IL4, IL6, IRS1, KIT, KITLG, MAPK1, PIK3R1, TNFA TNFRSF1A
	Etoposide	ABL1, BAX, BCL2, BCL2L1, CASP3, CSF2RB, CSF3, DIABLO, EP300, IFNG, IL2, IL5, IL6, IL8, MAPK8, MDM2, PTPN6, RB1,STAT3 TP53
	Paclitaxel	EPOR, FOS, IL12A, JAK2, KIT, KITLG
	Tamoxifen	BAX, BCL2, CRH, CSF2, CSF3, FOS, IFNG, IL1A, IL2, IL3, IL5, IL6, IL8, JUN, KITLG, MAPK8, TGFB1, TNFA, TP53
	Vinblastine	CD4, CSF3, IL8, IL10
Other	Furosemide	AGT, APAF1, BCL2, FOS, IL1B, IL2, JAK2, JUN, KITLG, LRP1, OXT, PPP1R12A, RB1, REN, RHOA, SERPINE1, SHC1, SLC12A2, SPP1, TNFA
	Hctz	ACE, AGTR1, CSF2, EDN1, FOS, IL5, IL6, JAK2, JUN, KITLG, REN, SERPINE1, SLC33A1, SOCS3, STAT3
	Promazine	CHRM1, CSF3, EGFR, HTR2C, IL3, IL6, JAK2, KIT, KITLG, PIK3CA, SHC1, STAT1

1 – Irinotecan was also analyzed in this therapeutic category but did not yield statistically significant DTSPs.

TABLE 8. Percentage of compounds with altered gene expression levels in each therapeutic category

	BAX	BCL2	CSF2	CSF3	FOS	IFNG	IL1	IL2	IL5	IL6	IL8	JAK1	JAK2	JUN	KIT	KITLG	MAPK1	MAPK8	STAT1	STAT3	TNFA
anti-infectives (n=9)	11	22	11	33	33	44	33	22	67	33	22	33	33	33	22	78	11	22	33	33	56
anti-cancer (n=5)	40	60	20	80	40	40	20	40	40	60	60	0	20	20	40	60	20	40	0	20	40
other (n=3) %	0	33	33	33	67	0	33	33	33	67	0	0	100	67	33	100	0	0	33	33	33
	18	35	18	47	41	35	29	29	53	47	29	18	41	35	29	76	12	24	24	29	47
Of anti-infectives, TNF was altered by all drugs that did not have altered KITLG, and 56% of all anti-infectives																					
of oncology drugs, CSF3 (GM-CSF) was altered by all drugs that did not have altered KITLG and 80% of all oncology compounds																					
All other drugs, KITLG and JAK2 was altered 100% of time, FOS was altered with 67% of the compounds																					
KITLG is typically altered by compounds (76%) that induce neutropenia irrespective of therapy area																					
100% of compounds were detected using the following combination: KITLG, CSF3 and TNF																					

CHAPTER 5: DTSP ALGORITHM - COMPARATIVE EVALUATION

Chapters 3 and 4 describe two unique analytical approaches for detecting signaling paths and toxicity hotspots inside large protein interaction networks. While the edge centrality approach (outlined in chapter 3) utilizes information on path involvement and edge centrality for each protein in the network, the gene expression integration approach (outlined in chapter 4) leverages information on dynamic, drug-specific changes that occur inside the network as a result of drug administration, as measured using *in vitro* gene expression experiments. This chapter describes two implementations aimed at comparative evaluation of the DTSP approach. The first evaluation compares paths detected for the set of drugs known to cause non-immune neutropenia with another set of paths, discovered for drugs NOT known to cause drug-induced neutropenia. The second evaluation compares both DTSP algorithms with the current standard approach for biomarker identification in toxicogenomics – namely, microarray data analysis.

Section 5.1 outlines the rationale for comparison of results with a 'control' set of drugs (i.e. drugs not known to cause non-immune neutropenia) and compares/contrasts results for the two set of drugs. The DTSP algorithm described in chapters 3 and 4 was used on this set of drugs along with the gene-expression based measure for assigning edge weights. Description of the algorithm implementation is therefore, not repeated in this chapter.

5.1 Comparison: Toxicity-inducing Drugs vs Control Drugs

5.1.1 Analysis Methods

Biological changes inside the network may be caused by many different external stimuli. A set of biological changes may lead to multiple phenotypes. For example, an administered drug and an environmental allergen may both cause an up-regulation in inflammation-related pathways inside the network and this in turn, could lead to many clinical manifestations, including the drug-induced toxicity. It would therefore, be interesting to compare the paths

discovered for toxicity-inducing drugs with paths for drugs that are **not** known to cause drug-induced neutropenia. Is it likely that some of the same proteins may be involved in both sets of paths ? What proteins, in paths discovered for drugs related to non-immune neutropenia, are also likely to be modulated in relation to other clinical effects and will therefore be 'hotspots' detected for a set of drugs not involved in non-immune neutropenia ? What paths are common to both sets of drugs and how are they regulated under 'toxic' and 'nontoxic' conditions?

In order to address the above questions, the algorithm was re-run on a set of drugs that are not known to cause drug-induced neutropenia. The choice of drugs used in this analysis was based on two criteria, namely manual curation by a toxicologist to confirm non-association with drug-induced non-immune neutropenia and availability of drug-specific gene expression data so as to enable implementation of the algorithm. Based on these criteria, a toxicologist collaborator performed manual curation to arrive at the list of drugs shown in Table 9.

TABLE 9. Drugs not known to cause non-immune neutropenia

Drug	Therapeutic Category	Description
Acacetin	Antineoplastic Agents	Acacetin, a flalvinoid compound, has been studied for its anti-proliferative effects in human non-small cell lung cancer
Bemegride	CNS stimulants, AntiConvulsants, Respiratory system agents	A CNS stimulant that is used to induce convulsions in experimental animals. It has been used as a respiratory stimulant and in the treatment of barbiturate overdose.
Benparidol	Antipsychotic agent	A butyrophenone with general properties similar to those of HALOPERIDOL. It has been used in the treatment of aberrant sexual behavior. (From Martindale, The Extra Pharmacopoeia, 30th ed, p567)
Ethotoin	AntiConvulsant	Ethotoin is a hydantoin derivative and anticonvulsant.
Ethionamide	Leprostatic Agents Antitubercular Agents Fatty Acid Synthesis Inhibitors	A second-line antitubercular agent that inhibits mycolic acid synthesis. It also may be used for treatment of leprosy. (From Smith and Reynard, Textbook of Pharmacology, 1992, p868)
Etomidate	Hypnotics and Sedatives Anesthetics, Intravenous	Imidazole derivative anesthetic and hypnotic with little effect on blood gases, ventilation, or the cardiovascular system. It has been proposed as an induction anesthetic.
Etodolac	Hypnotics and Sedatives Anesthetics, Intravenous	Etodolac is a non-steroidal anti-inflammatory drug (NSAID) with anti-inflammatory, analgesic and antipyretic properties. Its therapeutic effects are due to its ability to inhibit prostaglandin synthesis. It is indicated for of rheumatoid arthritis and osteoarthritis.
Galantamine	Parasympathomimetics Cholinesterase Inhibitors Nootropic Agents	A benzazepine derived from norbelladine. Galantamine is a cholinesterase inhibitor that has been studied as a treatment for Alzheimer's disease and other central nervous system disorders.

Table 9 (continued)

Indapamide	Antihypertensive Agents Diuretics	A benzamide-sulfonamide-indole. It is called a thiazide-like diuretic but structure is different enough (lacking the thiazo-ring) so it is not clear that the mechanism is comparable.
Isosorbide	Vasodilator Agents Nitrates and Nitrites Nitric Oxide Donors	Isosorbide mononitrate is a drug used principally in the treatment of angina pectoris1 and acts by dilating the blood vessels so as to reduce the blood pressure. It is sold by AstraZeneca under the trade name Imdur.
Mestranol	Estrogens	The 3-methyl ether of ethinyl estradiol. It must be demethylated to be biologically active. It is used as the estrogen component of many combination ORAL contraceptives.
Nalbuphine	Narcotics Analgesics, Opioid Narcotic Antagonists	A narcotic used as a pain medication. It appears to be an agonist at kappa opioid receptors and an antagonist or partial agonist at mu opioid receptors.
Nabumetone	Antineoplastic Agents	Nabumetone is a nonsteroidal anti-inflammatory drug (NSAID) of the arylalkanoic acid family (which includes diclofenac). Marketed under the brand name Relafen, it has been shown to have a slightly lower risk of gastrointestinal side effects than most other non-selective NSAIDs.
Orlistat	Enzyme Inhibitors Anti-Obesity Agents	Orlistat is a drug designed to treat obesity. Its primary function is preventing the absorption of fats from the human diet, thereby reducing caloric intake. Orlistat works by inhibiting pancreatic lipase, an enzyme that breaks down triglycerides in the intestine. Without this enzyme, triglycerides from the diet are prevented from being hydrolyzed into absorbable free fatty acids and are excreted undigested.
Papaverine	Vasodilator Agents Phosphodiesterase Inhibitors	An alkaloid found in opium but not closely related to the other opium alkaloids in its structure or pharmacological actions. It is a direct-acting smooth muscle relaxant used in the treatment of impotence and as a vasodilator, especially for cerebral vasodilation. The mechanism of its pharmacological actions is not clear, but it apparently can inhibit phosphodiesterases and it may have direct actions on calcium channels.
Pimozide	Antipsychotic agent	A diphenylbutylpiperidine that is effective as an antipsychotic agent and as an alternative to haloperidol for the suppression of vocal and motor tics in patients with Tourette syndrome.
Raloxifane	Antihypocalcemic Agents Osteoporosis Prophylactic Estrogen Antagonists	A second generation selective estrogen receptor modulator (SERM) used to prevent osteoporosis in postmenopausal women. It has estrogen agonist effects on bone and cholesterol metabolism but behaves as a complete estrogen antagonist on mammary gland and uterine tissue.
Selegiline	Central Nervous System Agents Antiparkinson Agents	A selective, irreversible inhibitor of Type B monoamine oxidase. It is used in newly diagnosed patients with Parkinson's disease. It may slow progression of the clinical disease and delay the requirement for levodopa therapy. It also may be given with levodopa upon onset of disability. (From AMA Drug Evaluations Annual, 1994, p385) The compound without isomeric designation is Deprenyl.

Hereinafter, the set of drugs not known to cause non-immune neutropenia (listed in Table 9) will be termed as 'control drugs' and the set of drugs known to cause non-immune neutropenia (listed in Table 4) will be described as 'toxicity-inducing' drugs. All steps outlined in chapters 3 and 4 were applied to this analysis and the methods will therefore not be repeated here.

Sections 5.1.2 summarizes results from a comparative analysis of DTSPs discovered for toxicity-inducing and control drugs.

5.1.2 Results and Discussion

Some characteristics of DTSPs for 'control' drugs were found to be equivalent to their 'toxicity-inducing' counterparts. On average, 127 unique proteins were involved in statistically significant paths connecting drug targets for control drugs with toxicity-related proteins in the network. On average, 147 unique proteins were involved in statistically significant paths connecting drug targets for toxicity-inducing drugs with toxicity-related proteins in the network. This may imply that the density of interactions among the set of proteins involved on DTSPs for each group is more or less similar within the two groups. The average number of paths were also found to be similar (424 for toxicity-inducing drugs as opposed to 519 for control drugs) in both groups. As the algorithm is designed to identify paths maximising the product of gene expression score and the number of such paths desired is an input parameter to the algorithm, we find about the same number of paths in both groups. The algorithm ranks detected paths based on gene expression change relative to other paths for the same drug. If a path is found to be highly ranked for a drug, this may not necessarily imply that paths discovered for 'control' drugs have the same extent of differential expression change (Treated vs. Control) compared to paths discovered for the 'toxicity-inducing' drugs. Section 5.1.2.2 describes the difference in extent of gene expression change between the two groups.

5.1.2.1 Toxicity-inducing vs Control drugs: Common path proteins and pathways

If a protein is found to be involved in a statistically significant signaling path for a toxicity-inducing drug, it may point to the protein's topological relevance to the drug-induced toxicity, depending on the extent of gene expression change and functional association between the protein and patho-physiological processes involved in the drug-induced toxicity. However, if

the same protein is also found to be involved on paths for control drugs, this can be interpreted in many ways. First, it would be important to know the extent to which the path proteins are differentially expressed under each treatment condition. If a protein has topological relevance but unequal gene expression change under 'control' and 'toxicity-inducing' conditions, this may point to its role in drug-induced toxicity. On the contrary, if the protein is topologically relevant (i.e. involved on paths for both toxicity-inducing and control groups) but exhibits the same direction or lack of gene expression change irrespective of the drug treatment, this could make the protein less relevant to the drug-induced toxicity. If the protein is involved in paths for one set of drugs (either toxicity-inducing or control) but not the other, it could still be relevant to the toxicity depending on the extent and type (up-regulation or down-regulation) of gene expression change observed in each case. For example, if a path protein associated with increase in inflammation is found to be down-regulated after control drug administration, this may point to its role in inhibiting inflammatory cytokines thereby preventing any neutropenia related adverse effect after drug administration. This protein would be relevant to understanding toxic mechanisms even if it was not found to be involved in any statistically significant paths for the toxicity-inducing drugs. Interpretation of findings solely on the basis of protein involvement on paths within each group is therefore likely to be less accurate compared to a detailed analysis of the extent of gene expression change and the effect of this change on toxicity-related end nodes in the network.

Key observations from identifying protein/s that are involved on at least one path for both groups and those that are only found on paths for one of the groups are as follows –

- Out of the 424 unique proteins found to be involved in at least one path for the 'toxicity-inducing' drugs, 272 proteins were also found to be involved in at least one path for the 'control' set of drugs. 152 proteins were only found on paths for the toxicity-inducing set while 247 proteins were found only on paths for the control drugs.

- Further analysis of the above set of proteins was carried out to understand whether specific canonical pathways were being modulated by each group of proteins. Geneset Enrichment for statistically significant KEGG pathway association was carried out using the WEBGESTALT application as described in section 3.1.5. The intention with geneset enrichment analysis was to understand whether some pathways were being modulated solely in the 'toxicity-inducing' set while other pathways were being modulated in both set of drugs. Figure 12 shows the top 10 pathways that were modulated by each set of proteins.

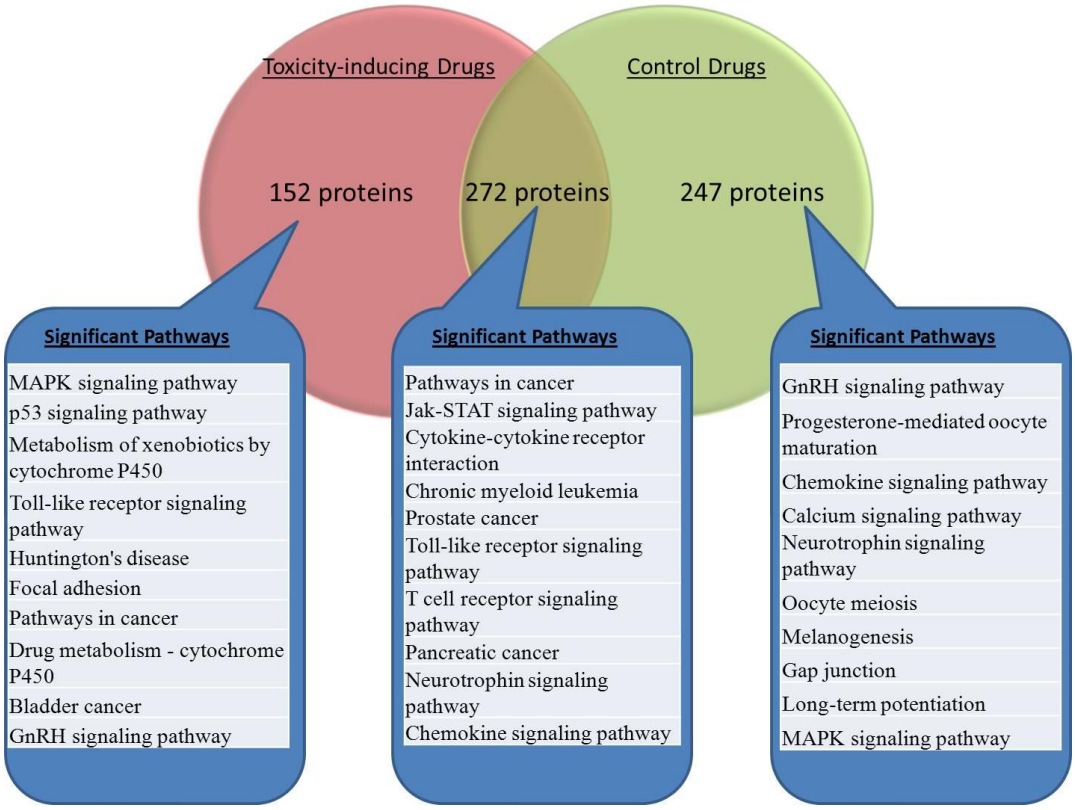

Figure 11. Venn Diagram shows the number of proteins common and exclusive to paths discovered for toxicity-inducing and control drugs. Geneset enrichment analysis revealed top 10 canonical pathways associated with path proteins discovered for each set of drugs. Some known pathways (like MAPK, cancer pathways and GnRH signaling pathway) were found to be common across the groups. Some others (like drug metabolism – cytochrome P450, Focal adhesion) were exclusively associated with path proteins discovered for drugs that cause non-immune neutropenia.

- Some known pathways like MAPK signaling and cancer-related pathways showed statistically significant association in both groups. This implies that distinct components of these pathways are involved in paths discovered for each group. This isn't unexpected given observations that distinct set of proteins within the MAPK signaling cascade have opposing effects on apoptosis [Xia et al. 1995]. While it cannot be concluded solely on findings from this analysis, it is possible that some proteins associated with MAPK signaling activate processes that lead to drug-induced neutropenia while certain other proteins also associated with MAPK signaling exhibit the exact opposite effect under control treatment conditions. MAPK is a key signaling pathway known to modulate the therapeutic effects of chemotherapy drugs[Boldt et al. 2002]. Detailed analysis of these path proteins with additional data and knowledge of up and down regulation of relevant proteins in the context of other changes may therefore, help understand the underlying toxicological processes better.

5.1.2.2 Toxicity-inducing vs. Control drugs: Extent of genomic regulation

As described in section 5.2, both topological relevance and extent of gene expression change in path proteins are important to understand when interpreting the results from this analysis. Topological relevance of proteins discovered for each set of drugs have been analyzed in section 5.2.

In order to compare the extent of gene expression change in path proteins for the two set of drugs, an average of path expression scores for all paths was computed across all statistically significant paths for drugs in the toxicity-inducing and control set. Figure 13 shows a boxplot of the average score in each group.

The boxplot implies that proteins involved in DTSPs for drugs in the 'toxicity-inducing' set exhibit greater extent of gene expression change (treated vs. control) compared to those for drugs in the 'control' set. In other words, the set of statistically significant paths discovered for

toxicity-inducing drugs may be ‘active’ under drug administration conditions while the set of paths discovered for control drugs are relatively ‘inactive’ from a gene expression standpoint.

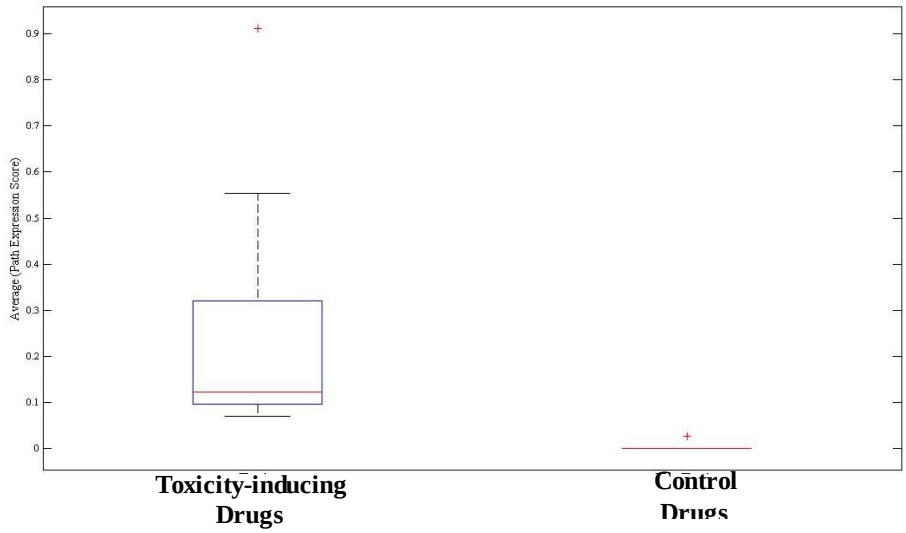

Figure 12. Average Path expression score in each set reveals the greater extent of gene expression change (treated vs. control) in path proteins discovered for the Toxicity-inducing drugs.

The advantage of using the DTSP algorithm on protein interaction networks and integrating gene expression data is clear from this observation. Topological relevance of a protein provides a rationale for studying its gene expression change and the extent of gene expression change indicates that one set of path proteins is more likely to yield a downstream effect in the form of a drug-induced toxicity.

Some DTSPs were found to be common to both set of drugs. If a statistically significant path discovered for the toxicity-inducing drugs also exists in the control set, it would be important to know whether the extent and direction of gene expression regulation for such paths is different between the two set of drugs. If the end nodes on such paths are differentially regulated between the two sets of drugs, this may point to their role in drug-induced non-immune Neutropenia. This is because Neutropenia is defined as a clinical significant **reduction** in Neutrophil count. Paths where the toxicity-related end node proteins are down-regulated, are

therefore important because they may lead to reduction in Neutrophil production. The 'toxicity-inducing' set of DTSPs were therefore, compared with all paths discovered for the control set to identify four common DTSPs. The extent of gene expression regulation in the end nodes (known hematopoiesis-related proteins) was calculated (ratio of treated and control) for each common path. Table 10 shows the four paths and the gene expression ratio for the end nodes for each path.

Colony stimulating factor CSF3, an end node for paths (depicted as a chain of HGNC symbols) IL1B-MAPK14-STAT1-CSF3 and REN-EDN1-ACE1-CSF3 was marginally down-regulated (treated/control) for the first path in both toxicity-inducing as well as control group while in the second path, CSF3 was found to be marginally up-regulated in both, the toxicity-inducing and the control set. The longer path that started at CCND1 also led to marginal up-regulation of its end node CSF2 in both sets of drugs. DTSPs common to both groups and regulation of toxicity-related proteins on those paths.

TABLE 10. DTSPs common to both groups and regulation of toxicity-related proteins on those paths

Drug-Toxicity Signaling Path (DTSP)							Toxicity-inducing Drugs Ratio (Trt/Ctl)	Control Drugs Ratio (Trt/Ctl)
IL1B	MAPK14	STAT1	CSF3				0.939631803	0.944454842
REN	EDN1	ACE1	CSF3				1.016675682	1.112032776
CHRM1	EGFR1	SHC1	KITLG				0.857261358	1.095764547
CCND1	STAT3	SRC5	GJA1	MAPK3	FOS1	CSF2	1.078424425	1.015210632

KITLG (KIT Ligand) was the only end node that was found to be marginally down-regulated in the toxicity-inducing set and marginally up-regulated in the control set. This implies that down-regulation of KITLG via EGFR and SHC1 proteins may lead to reduced

Neutrophil production due to toxicity-inducing drugs. This observation needs further experimental confirmation to evaluate the range of expression change in KITLG and corresponding change in Neutrophil production.

The role of KITLG as a key network protein involved in drug-induced non-immune neutropenia as hypothesized in Section 4.3.4 and its inclusion as a potential biomarker is therefore, supported with this comparative analysis of paths discovered for the toxicity-inducing and control drugs.

5.2 Comparative Evaluation of algorithm accuracy

The integration of gene expression and protein interaction information may seem to offer a more comprehensive model compared to using either 'omics' platform by itself. However, a statistical evaluation of the DTSP algorithm is necessary in order to understand its predictive ability compared to the most prevalent toxicogenomics method. To achieve this, microarray data analysis was performed using the same gene expression dataset that was used earlier to compute the gene expression measure. Results from all algorithms were compared against results from microarray data analysis. The choice of microarray data analysis for comparison was driven by two factors:

1. As the first high-throughput 'omics' platform that developed and evolved after the human genome project, gene expression-based measurements are relatively mature compared to protein expression and metabonomics platforms. Affymetrix Genechip arrays have been known to deliver repeatable results when performed under quality-controlled experimental conditions. The other 'omics' platforms (proteomics and metabonomics) are relatively newer and drug-specific data for these platforms are not widely available.

2. Microarray gene expression data are publicly available and extensively used for toxicogenomics applications [Roth et al. 2011, Mongan and Hamadeh 2011, Afshari et al. 2011, Hamadeh et al. 2002].

The following section describes steps implemented to analyze microarray gene expression data.

5.2.1 Microarray Data Analysis

Analysis of raw data collected using a microarray platform involved the following steps:

1. Raw CEL files for one or more treated samples and more than one control sample for each treated sample were downloaded for each 'toxicity-inducing' drug from the connectivity map server. CEL files (treated and 1-4 vehicles scans) for 51 experiments were downloaded for the 19 toxicity-inducing drugs analyzed in this study. For the 18 control drugs, CEL files (treated and 1-4 vehicles) for 41 experiments were downloaded. There were one or more treated samples for each drug (dose: 10µM, duration: 6 hours) and multiple vehicle scans from HT_HgU133 chip array in NCI MCF7 cell lines.
2. The 'ExpressionFileCreator' module in GenePattern was used to convert CEL files into gct format expression files [Reich et al. 2006]. Robust multichip average (RMA) measure was used for conversion and quantile normalization was applied.
3. The average log ratio of treated/average(vehicle) values were calculated for each probeset and gene expression data for all drugs were combined into a single gct file with class labels 'toxicity-inducing' and 'control'.
4. The gene expression dataset was then analyzed using the 'ComparativeMarkerSelection' module in GenePattern. The 22277 probesets on HT_HgU133 chip were ranked using the 'Signal to Noise' (SNR) test statistic –

$$\mathbf{SNR = (\mu_{toxic} - \mu_{control})/(\sigma_{toxic} + \sigma_{control})}$$

where μ_{toxic} – Mean value for a probeset in the 'toxicity-inducing' set

$\mu_{control}$ - Mean value for a probeset in the 'control' set

σ_{toxic} – Standard deviation for a probeset in the 'toxicity-inducing' set

$\sigma_{control}$ – Standard deviation for a probeset in the 'control' set

Use of the SNR test statistic ensured that probesets with largest difference in mean expression value between the two sets and least standard deviation within the two sets were ranked higher.

5.2.2 Choice of Benchmark Database

In order to compare outputs from the two DTSP algorithms with results from microarray data analysis, benchmark data associating specific network proteins with the drug-induced toxicity were required. The Comparative Toxicogenomics Database (CTD), with its collection of chemical–gene, chemical–disease and gene–disease relationships manually curated and inferred from published literature was used for this purpose. The database is dedicated to promoting the exploration and development of testable hypotheses about the effects of the environment on human health [Davis et al. 2011].

A significant number of relationships in the CTD result from computation of an inference score. The inference score reflects the degree of similarity between CTD chemical–gene–disease networks and a similar scale-free random network. Many biological networks, such as disease and metabolic networks, have been shown to be scale-free random networks. The score takes into account the connectivity of the chemical, disease and each of the genes used to make the chemical disease inference. The higher the score, the more likely the inference network has non-uniform connectivity as observed in scale-free random networks.

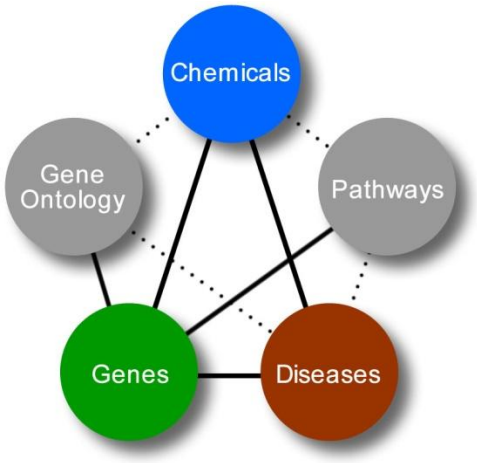

Figure 13. CTD integrates curated data for chemical-gene interactions, chemical- disease and gene-disease relationships (colored circles) with select public datasets (gray circles; pathways from the KEGG and Reactome databases and GO annotations). Solid lines describe directly curated or integrated relationships and dashed lines describe inferred relationships. (Figure and Legend reproduced from the BioInformatics Primer, Pathway Interaction Database, Mattingly 2011)[Mattingly et al. 2006] [Davis et al. 2011]

As summarized in Table 11, CTD provides ~325,000 curated chemical-gene interactions and ~1,500,000 curated and inferred Gene-Disease relationships. A major strength of CTD is that these core data are manually curated from the literature by professional biocurators [Salimi and Vita 2006], ensuring accuracy. CTD does use text mining to triage the literature, but each reference (abstract or full-text) is read by biocurator to identify interactions and relationships, and all curated data is supported by its source citation. The manual curation approach at CTD allows biocurators to validate every interaction and relationship, ensure that the correct chemical name and gene symbol is used, and generate detailed descriptions of the types of interaction. Data are uploaded to the database monthly. The CTD was therefore used as a benchmark to compare results from microarray analysis with results from the two DTSP algorithms proposed in this thesis.

TABLE 11. CTD data status as of May 2011 (summarized from BioInformatics Primer, Mattingly 2011)

Description	Data Count
Chemical-Gene curated Interactions	325,342
Gene-disease relationships	1,575,076
- Curated	13,187
- Inferred	1,561,889
Chemical-disease relationships	334,448
- Curated	11,378
- Inferred	323,070

5.3 Results and Discussion

Hereinafter, the DTSP Algorithm described in chapter 3 will be referred to as the 'DTSP_Edgecent' (for Edge Centrality measure) algorithm and the DTSP algorithm described in chapter 4 (combining protein interaction network and gene expression measure) will be referred to as 'DTSP_GeneExp' algorithm. The two algorithms were compared with results from microarray data analysis described in section 5.2.1. Comparison of results was carried out as follows:

1. Results from each algorithm (path proteins in case of the two DTSP algorithms and probesets in case of microarray data analysis) were mapped to genes in order to enable comparison with neutropenia-related genes provided by the Comparative Toxicogenomics database.
2. Genes common to the three sets, i.e. the human subset of the STITCH database, the HT_HgU133 chip genes and CTD database genes were identified. This was the larger set of common genes from which each algorithm mined a smaller set of 'toxicity-related' genes.

3. From within this common set, the DTSP_Edgecent algorithm identified 182 unique genes involved on paths discovered for the toxicity-inducing drugs. The DTSP_GeneExp algorithm identified 417 unique genes involved on paths discovered for the toxicity-inducing drugs. The top 500 genes (ranked by SNR statistic described in section 5.2.1) from microarray data analysis were compared with results from the two DTSP algorithms using the CTD benchmark.

Comparison of the two DTSP algorithms with results from microarray data analysis was undertaken using gene-neutropenia-drug relationships from the CTD database as a reference. The CTD database contained less than 25 manually curated relationships between 18 genes and neutropenia-related disorders (including severe congenital 1/2 autosomal dominant, Poikiloderma with Neutropenia disorder and chronic idiopathic subtypes) and not all gene-neutropenia relationships were found in the context of drug administration. The number of inferred relationships in the database was much higher, each with an inference score that ranged from about 1.9 (indicating weak association) to 140 (indicating strong association).

In order to explore changes in accuracy estimates with use of different inference score thresholds for identifying a 'positive set' inside the CTD benchmark, three different 'positive' set of genes were considered –

Set 1: All Neutropenia genes with direct (manually curated) evidence in CTD

Set 2: List of top 500 neutropenia genes in CTD (direct and inferred evidence)

Set 3: All Neutropenia genes with direct and inferred evidence for association with the set of toxicity-inducing drugs used for the DTSP algorithm.

Results from comparative evaluation of each algorithm using Set 1 is summarized below:

Out of the 18 unique genes that were associated with various subtypes of neutropenia, 11 were associated with the higher-granularity disease 'neutropenia', 5 with congenital subtype or cyclic subtype of the disease, one with 'granule deficiency' and one with 'Poikiloderma with

Neutropenia'. Out of the 11 genes associated with 'Neutropenia', 4 were associated with genetic polymorphisms in some cases associated with genes that degrade specific chemotherapy drugs. Given that in vitro gene expression data used in this analysis was primarily aimed at detecting gene expression changes from drug administration using *in vitro* tissue samples and evidence for genetic polymorphism may not imply a functional effect under drug administration conditions, it was considered appropriate to compare algorithm results with the remaining set of genes. Out of the 7 remaining genes in the 'positive' set, 5 genes (71%) were detected using the DTSP-GeneExp algorithm, 2 genes (29%) were detected using the DTSP-Edgecent algorithm and 1 gene (14%) was ranked among the top 500 with microarray data analysis.

For the 'positive' sets 2 and 3 in CTD, the following measures were computed

a. **True Positives (TP)**: Number of genes identified as toxicity-related using an algorithm that are also associated with the toxicity inside CTD.
b. **False Positives (FP)**: Number of genes identified as toxicity-related using an algorithm that are not associated with the toxicity inside CTD.
c. **True Negatives (TN)**: Number of genes identified as NOT toxicity-related using an algorithm that are NOT associated with the toxicity inside CTD.
d. **False Negatives (FN)**: Number of genes identified as NOT toxicity-related using an algorithm that are associated with the toxicity inside CTD.
e. **Sensitivity** = TP/(TP + FN)
f. **Specificity** = TN/(FP + TN)
g. **Positive Predictive Value** = TP/(TP + FP)
h. **Negative Predictive Value** = TN/(TN + FN)
i. **False Discovery Rate** = FP/(FP + TP)
j. **Accuracy = (TP + TN)/(TP + TN + FP + FN)**

For set 2, the top 500 genes (ranked by inference score) were considered as the CTD 'positive' set and for set 3, all 4004 genes (with direct and inferred evidence) were considered as the CTD 'positive' set. Table 12 shows measures computed for both sets.

TABLE 12.Comparative Evaluation of DTSP algorithms

	CTD (Top 500)			CTD - all 'neutropenia' genes (4004)		
Measure	**Microarray (top 500)**	**DTSP EdgeCent**	**DTSP GeneExp**	**Microarray (top 500)**	**DTSP EdgeCent**	**DTSP GeneExp**
True Positives	37 (7.4%)	80 (44%)	141 (34%)	231 (46%)	136 (74%)	296 (71%)
False Positives	463 (93%)	102 (56%)	276 (55%)	269 (54%)	46 (25%)	121 (29%)
True Negatives	10274 (96%)	10635 (96%)	10461 (97%)	6964 (65%)	7187 (65%)	7112 (66%)
False Negatives	463 (4.3%)	420 (3.7%)	359 (3.3%)	3773 (35%)	3868 (35%)	3708 (34%)
Sensitivity	0.074	0.16	0.282	0.032	0.034	0.074
Specificity	0.96	0.99	0.97	0.93	0.99	0.98
PPV	0.074	0.44	0.34	0.46	0.75	0.71
NPV	0.96	0.96	0.97	0.35	0.65	0.66
FDR	0.93	0.56	0.66	0.538	0.25	0.29
Accuracy	0.92	0.95	0.94	0.36	0.65	0.66

Both DTSP algorithms showed a significant improvement over standard microarray data analysis for the two sets. A five-fold improvement in 'true positives', two-fold improvement in 'false positives' and marginal improvements in 'true negatives' was observed with the DTSP algorithms with set 2 (CTD Top500),. However, due to high proportion of 'False positives' (partly attributable to a stringent cut-off for number of genes in the 'true positive' set), the sensitivity with all three methods was impacted. The two DTSP algorithms showed distinct improvement in sensitivity over microarray data analysis, with the DTSP-GeneExpression approach showing the highest sensitivity among the three methods. Specificity was high with all three methods, implying that if a gene was found to be associated with the drug-induced

toxicity using either methods, there was 3-4% chance that this would be a 'false positive' finding, with the DTSP algorithms having marginally improved specificity.

In terms of overall accuracy, the DTSP_Edgecent algorithm marginally outperformed the other two methods, with 95% accuracy followed by DTSP_GeneExp algorithm (94% accuracy) and Microarray data analysis with 92% accuracy.

Set 3 used a bigger 'positive' set, which improved the proportion of 'true positives' detected with all three methods at the expense of higher 'false Negatives'. As a much higher number of 'true positive' genes was identified inside the CTD benchmark (total 4004 genes) compared to the number of genes identified as toxicity-related from microarray data analysis (top 500, ranked by Signal-to-Noise ratio statistic), there were a higher proportion of false negatives with set 3.

However even with set 3, the two DTSP algorithms outperformed microarray data analysis in terms of 'true positives' detected (46% for microarray data analysis, 74% for DTSP_Edgecent and 71% for DTSP_GeneExp), Specificity (93% for microarray data analysis, 98% for DTSP_GeneExp and 99% for DTSP_Edgecent), Positive Predictive Value (46% - Microarray, 71% - DTSP_GeneExp and 74% - DTSP_Edgecent) and Overall accuracy (36% - Microarray, 65% - DTSP_Edgecent and 66% - DTSP_GeneExp).

The advantages of using the DTSP approach over standard microarray analysis are apparent from the above analysis. However, there is only a marginal difference in measures for the two DTSP algorithms and on some measures (e.g. %True positives, PPV and Specificity) the edge centrality-based DTSP algorithm marginally outperformed the gene expression-based approach. Does this necessarily imply that the edge centrality-based approach is superior to the gene expression-based approach? Further analysis is recommended before one DTSP approach can be preferred over the other. The following aspects of the comparative evaluation and the two DTSP algorithms need further consideration -

1. The gene expression measure emphasizes network dynamics (temporal changes in gene expression after drug administration) when identifying toxicity-related proteins whereas the edge centrality approach emphasizes topological connectivity. Understanding and incorporating network dynamics into a large network model may be inherently more complicated than computing edge centrality with the assumption that connectivity determines 'interestingness' irrespective of temporal changes that may happen to the network after drug administration. Given the limited amount of available evidence from comparative evaluation that demonstrates marginal superiority for the edge centrality approach, the choice of one DTSP algorithm over the other remains unclear. Further investigation is required, both in terms of using additional gene expression data as well as integrating additional data types, in order to refine the gene expression based approach. Edge centrality seems to provide an interesting alternative view on drug-toxicity signaling path but whether choice of a biomarker panel can be based solely on network connectivity remains to be seen. A measure that combines the strengths of the two DTSP algorithms may perhaps lead to an improved specificity and sensitivity over using the two distinct measures.

2. The choice and availability of an appropriate benchmark for evaluating such algorithms needs further investigation. CTD provides a high-quality resource that catalogs relationships between genes, diseases and chemicals from a variety of sources. However, for algorithms like the ones developed in this thesis and those that are implemented on large integrated networks, some of the same underlying sources would be used (e.g. co-occurrence between genes and diseases inside published literature) to populate both the STITCH/STRING databases and the CTD/benchmark. There is a need to identify benchmark databases that capture gene-chemical-disease relationship from independent experimental/direct evidence sources. Also, CTD provides many types of gene-disease-drug relationships, only of some of which can be

expected to be detected using the DTSP approach, at least partly due to the algorithm limitations with using only two data types.

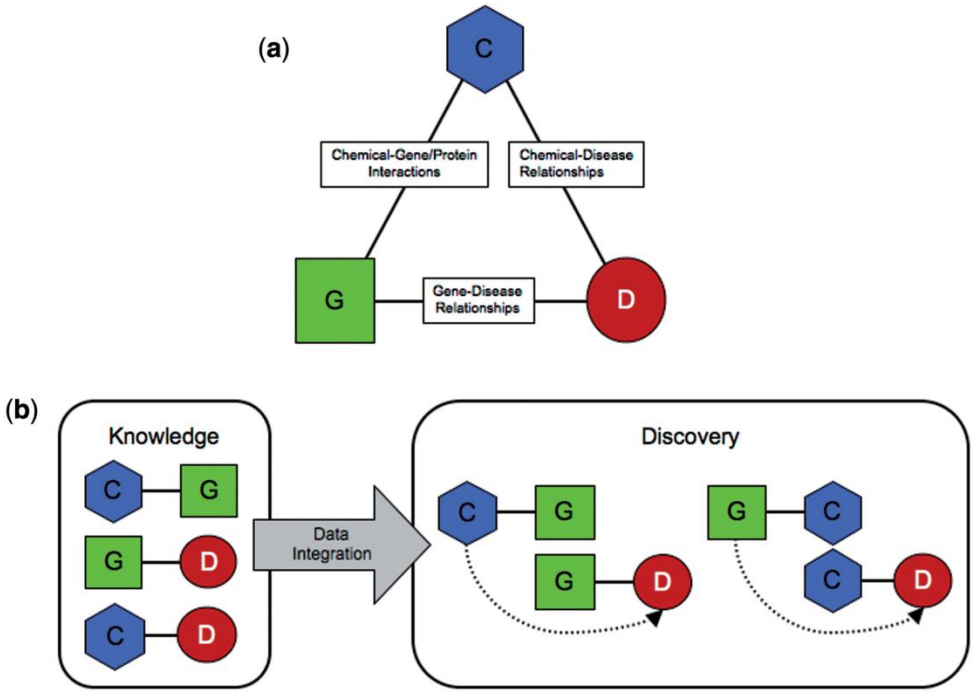

Figure 14. The CTD curation and integration paradigm (reproduced from Davis et al. 2011[Davis et al. 2009]. Evidence for a direct gene-chemical relationship and a direct chemical-disease relationship forms the basis for an inferred gene-disease relationship.

Moreover, as shown in Figure 15, inferred relationships between genes and disease in the CTD are based on the paradigm that if there is direct evidence in published literature associating a gene with a particular chemical and if there is also direct evidence that the same chemical is associated with a toxicity, then it is inferred that the gene and the disease may also be related. This paradigm contradicts underlying assumptions made with the two DTSP algorithms that primary or secondary drug targets (proteins) are only relevant if they enable topologically connectivity with proteins known to be associated with the drug-induced toxicity (i.e. DTSP_Edgecent algorithm) and are involved in highly differentially regulated drug-toxicity signaling paths (ie. DTSP_GeneExp algorithm). The 'positive' set identified with the benchmark, therefore requires further refinement in order to distil the specific genes that can be definitively linked to drug-induced non-immune neutropenia and not neutropenia in general

(includes congenital and many other subtypes of the disease). The apparent lack of sensitivity with all three methods can be at least partly attributed to the above limitations with using a comparison benchmark.

On the other hand, interpretation of results from signaling path detection approaches like the DTSP algorithms requires further refinement in path protein subsets. Further analysis of discovered paths may lead to identification of protein subsets that are unique to the drug and its mechanism of action and therefore, have a stronger association with the drug than to drug-induced toxicity across a set of drugs.

CHAPTER 6: CONCLUSIONS AND FUTURE WORK

The ability to apply signaling path detection to *in silico* toxicity evaluation offers exciting opportunities. This thesis demonstrates how background knowledge about drugs and their effects can be utilized in conjunction with a network-based model to discover toxicity-related path proteins. Integration of protein interaction and gene expression information yields insights into the dynamic changes that occur at a systems level after drug administration. Drug-Toxicity signaling paths and discovery of common characteristics of path proteins across a set of drugs provides opportunities for understanding toxic mechanisms and possible biomarker panels that can, in the future be used for screening new drug candidates. The DTSP algorithm represents one of the first steps towards driving an area of research in systems toxicology that can lead to increasingly accurate and comprehensive *in silico* toxicity evaluation models.

6.1 Thesis Contributions

This thesis has made the following important contributions:

First, the drug-toxicity signaling path detection algorithm is a unique approach for discovering signaling paths inside protein interaction networks that are directly relevant and downstream from drug target proteins and upstream from toxicity-related proteins inside a large protein interaction network. The DTSP algorithm is the first to apply signaling path detection to *in silico* toxicity evaluation.

Second, both local and global properties of biological networks have recently been studied and applied towards prediction of protein function. However, the definition and implementation of an edge centrality measure for detection of drug-induced toxicity hotspots inside protein interaction networks is a unique contribution of this thesis.

Third, algorithms aimed at re-engineering biomolecular networks from gene expression data and integration of genomic-proteomic data have been proposed recently. However, integration of gene expression and protein interaction data for detection of toxicity hotspots is a unique contribution of this thesis.

Four, yet another unique contribution of this thesis is a biomarker panel associated with drug-induced non-immune neutropenia, discovered using the DTSP algorithm.

The following section summarizes the next steps and areas in which the DTSP algorithm can be extended in the future.

6.2 Recommendations for Future Work

6.2.1 Experimental Confirmation

Validation of the DTSP algorithm is a necessary pre-requisite for the approach to be considered as a toxicity screen in pharmaceutical R&D. In order to confirm that the biomarker panel proposed in chapter 3 is predictive of drug-induced non-immune neutropenia, the results will need to be replicated on other drugs that are known to cause the toxicity. Laboratory-based experiments on identified path proteins may provide experimental evidence to support the preliminary hypothesis.

6.2.2 Applicability to other drug-induced toxicities

If the algorithm can be implemented and evaluated for toxicities with high-quality benchmark data that supports drug-toxicity association and protein-toxicity relationships, this will provide greater confidence in its ability to detect toxicity-specific protein hotspots inside large biological networks. Discovery of toxicity hotspots using the DTSP algorithm requires availability of *in vitro* gene expression data. In addition to this, some of the additional data types outlined in section 6.2.3 may also improve screening of candidate paths to reveal those that are directly related to the toxicity of interest.

6.2.3 Incorporation of Additional data types

Systems analysis for detection of toxicity hotspots can potentially utilize many other data types. The metric used for weighting edge of the network of reliable paths in the DTSP algorithm, can incorporate additional background knowledge on network proteins, such as miRNA (microRNA) annotations and protein localization information. For example, when considering events like hepatotoxicity, paths consisting of proteins known to be localized in the liver may be considered more relevant. Incorporation of miRNA annotations for path proteins may reveal common regulators of proteins involved in drug-toxicity signaling paths. Improved mechanistic understanding of pathway regulation may lead to improved biomarkers for prediction of drug-induced toxicities. The algorithm may also benefit from additional data types in the form of chemical properties of the drugs, pharmacokinetic properties and pharmacodynamic changes at various time points after drug administration as well as higher level canonical pathway annotations of network nodes. Finally, drug-toxicity signaling paths that are 'active' only in the presence of more than one drug, may be associated with adverse effects that result from administration of multiple drugs [Tatonetti et al. 2012]. Understanding drug-drug interactions through the lens of automatic signaling path detection may be an important benefit from our approach.

6.2.4 System-wide temporal and stochastic simulation

Protein interaction network analysis described in this thesis represents a non-temporal, non-dynamic, semi-stochastic , 'top down' approach starting at drugs and their clinical effects, using a constructed protein interaction network and analyzing it to understand the role of proteins involved on drug-toxicity signaling paths. An alternative approach would be to utilize a 'bottom up' approach, where individual components of the network are first studied in greater detail. The bottom up approach starts with identifying regulation functions for each network element and gradually building the holistic picture of the system, at the protein,

pathways, tissue, organ and organism leve1. However, analysis of smaller metabolic networks has been undertaken by many other researchers to leverage the ‘bottom up’ approach. One of the limitations with adopting a thorough quantitative approach at the level of a large network is the scarcity of temporal data on protein interactions and all levels above it (pathways, cell, tissue and organ) for the human biological system. This limitation is rapidly being replaced by the limitation in our ability to mine these networks to reveal the ‘truth’ underlying patho-physiological phenomena.

6.2.5 **Network Medicine**

Drugs of the future may be able to target sub-networks instead of proteins, thereby giving rise to a completely new approach to drug discovery, recently termed as ‘network medicine’. Pawson et al. suggest two different strategies to drive network medicine, one involves a synthetic biology approach that aims at rewiring (adding new interactions) of the network using small molecules or novel synthetic modular proteins. This strategy exploits the modular nature of signaling proteins to change the topology and wiring of the network by adding or deleting interactions. The second approach relies on changing the information flow inside the network by targeting the phosphorylation states of key proteins.

As systems biology is a relatively nascent field, many more approaches for targeting protein hotspots or modules directly may emerge over the next decade. Our ability to understand the dynamics of biological networks would be the key to discovering safe and efficacious therapeutic interventions.

Bibliography

[Lamb 2007] Lamb. The Connectivity Map: a new tool for biomedical research. Nat Rev Cancer 2007;7:54-60

[Sharan et al. 2007] Sharan et al. Network-based prediction of protein function. Mol Syst Biol 2007;3:

[Barnett 2005] Barnett. Culture and the Structure of the International Hyperlink Network. Journal of Computer-Mediated Communication 2005;11:217-238

[Shaikh et al. 2007] Shaikh et al. Graph Structural Mining in Terrorist Networks. 2007

[Owen-Smith and Powell 2004] Owen-Smith et al. Knowledge Networks as Channels and Conduits: The Effects of Spillovers in the Boston Biotechnology Community. ORGANIZATION SCIENCE 2004;15:5-21

[Chen et al. 2008] Chen et al. The thematic and citation landscape of Data and Knowledge Engineering (1985–2007). Data & Knowledge Engineering 2008;67:234-259

[Alfarano et al. 2005] Alfarano et al. The Biomolecular Interaction Network Database and related tools 2005 update. Nucl. Acids Res. 2005;33:D418-424

[Quackenbush 2007] Quackenbush. Extracting biology from high-dimensional biological data. J Exp Biol 2007;210:1507-1517

[Oprea et al. 2007] Oprea et al. Systems chemical biology. Nat Chem Biol 2007;3:447-450

[Albert 2007] Albert. Network Inference, Analysis, and Modeling in Systems Biology. Plant Cell 2007;19:3327-3338

[Hopkins 2008] Hopkins. Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 2008;4:682-690

[Hopkins 2007] Hopkins. Network pharmacology. Nat Biotech 2007;25:1110-1111

[Edwards and Aronson 2000] Edwards et al. Adverse drug reactions: definitions, diagnosis, and management. The Lancet 2000;356:1255-1259

[Ganter et al. 2006] Ganter et al. Toxicogenomics in drug discovery and development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix® database. Pharmacogenomics 2006;7:1025-1044

[Alm and Arkin 2003] Alm et al. Biological networks. Current Opinion in Structural Biology 2003;13:193-202

[Barabasi and L. 1999] Barabasi et al. Emergence of Scaling in Random Networks. Science 1999;286:509-512

[Jeong et al. 2000] Jeong et al. The large-scale organization of metabolic networks. Nature 2000;407:651-4

[Luscombe et al. 2002] Luscombe et al. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. Genome Biology 2002;3:research0040.1 - research0040.7

[Jeong et al. 2001] Jeong et al. Lethality and centrality in protein networks. Nature 2001;411:41-42

[Fraser et al. 2002] Fraser et al. Evolutionary Rate in the Protein Interaction Network. Science 2002;296:750-752

[Yu et al. 2007] Yu et al. The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. PLoS Comput Biol 2007;3:e59

[Girvan and Newman 2002] Girvan et al. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America 2002;99:7821-7826

[Radicchi et al. 2004] Radicchi et al. Defining and identifying communities in networks. Proceedings of the National Academy of Sciences of the United States of America 2004;101:2658-2663

[Spirin and Mirny 2003] Spirin et al. Protein complexes and functional modules in molecular networks. Proceedings of the National Academy of Sciences of the United States of America 2003;100:12123-12128

[Dietmann et al. 2006] Dietmann et al. Resources and Tools for Investigating Biomolecular Networks in Mammals. Current Pharmaceutical Design 2006;12:3723-3734

[Steffen et al. 2002] Steffen et al. Automated modelling of signal transduction networks. BMC Bioinformatics 2002;3:34

[Bugrim et al. 2004] Bugrim et al. Early prediction of drug metabolism and toxicity: systems biology approach and modeling. Drug Discovery Today 2004;9:127-135

[Gomase and Tagore 2008] Gomase et al. Toxicogenomics. Curr Drug Metab 2008;9:

[Kitano 2002] Kitano. Systems Biology: A Brief Overview. Science 2002;295:1662-1664

[Ekins 2006] Ekins. Systems-ADME/Tox: Resources and network approaches. Journal of Pharmacological and Toxicological Methods 2006;53:38-66

[Dearden 2003] Dearden. In silico prediction of drug toxicity. Journal of Computer-Aided Molecular Design 2003;17:119-127

[Fliri et al. 2007] Fliri et al. Analysis of System Structure-Function Relationships. ChemMedChem 2007;2:1774-1782

[Paolini et al. 2006] Paolini et al. Global mapping of pharmacological space. Nat Biotech 2006;24:805-815

[Waring et al. 2001] Waring et al. Clustering of Hepatotoxins Based on Mechanism of Toxicity Using Gene Expression Profiles. Toxicology and Applied Pharmacology 2001;175:28-42

[Lamb et al. 2006] Lamb et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. Science 2006;313:1929-1935

[Aravind 2000] Aravind. Guilt by Association: Contextual Information in Genome Analysis. Genome Research 2000;10:1074-1077

[Huynen et al. 2000] Huynen et al. Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences. Genome Research 2000;10:1204-1210

[Dittrich et al. 2008] Dittrich et al. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics 2008;24:i223-231

[Chua et al. 2006] Chua et al. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics 2006;22:1623-1630

[Letovsky and Kasif 2003] Letovsky et al. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 2003;19:i197-204

[Hu et al. 2007] Hu et al. A Novel Approach for Mining and Fuzzy Simulation of Subnetworks From Large Biomolecular Networks. Fuzzy Systems, IEEE Transactions on 2007;15:1219-1229

[Schwikowski et al. 2000] Schwikowski et al. A network of protein-protein interactions in yeast. Nat Biotech 2000;18:1257-1261

[Hishigaki et al. 2001] Hishigaki et al. Assessment of prediction accuracy of protein function from protein--protein interaction data. Yeast 2001;18:523-31

[Vazquez et al. 2003] Vazquez et al. Global protein function prediction from protein-protein interaction networks. Nat Biotechnol 2003;21:697-700

[Karaoz et al. 2004] Karaoz et al. Whole-genome annotation by using evidence integration in functional-linkage networks. Proceedings of the National Academy of Sciences of the United States of America 2004;101:2888-2893

[Deng et al. 2003] Deng et al. Prediction of protein function using protein-protein interaction data. J Comput Biol 2003;10:947-60

[Bader 2003] Bader. Greedily building protein networks with confidence. Bioinformatics 2003;19:1869-1874

[Arnau et al. 2005] Arnau et al. Iterative Cluster Analysis of Protein Interaction Data. Bioinformatics 2005;21:364-378

[King et al. 2004] King et al. Protein complex prediction via cost-based clustering. Bioinformatics 2004;20:3013-3020

[Dunn et al. 2005] Dunn et al. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. BMC Bioinformatics 2005;6:39

[Hu and Wu 2007] Hu et al. Data mining and predictive modeling of biomolecular network from biomedical literature databases. IEEE/ACM Trans Comput Biol Bioinform 2007;4:251-63

[Yildirim et al. 2007] Yildirim et al. Drug-target network. Nat Biotech 2007;25:1119-1126

[Goh et al. 2007] Goh et al. The human disease network. Proceedings of the National Academy of Sciences 2007;104:8685-8690

[Xu and Li 2006] Xu et al. Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics 2006;22:2800-2805

[Ekins et al. 2005] Ekins et al. Techniques: Application of systems biology to absorption, distribution, metabolism, excretion and toxicity. Trends in Pharmacological Sciences 2005;26:202-209

[Ideker et al. 2001] Ideker et al. Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. Science 2001;292:929-934

[Hüffner et al. 2008] Hüffner et al. Algorithm Engineering for Color-Coding with Applications to Signaling Pathway Detection. Algorithmica 2008;52:114-132

[Kelley et al. 2003] Kelley et al. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proceedings of the National Academy of Sciences of the United States of America 2003;100:11394-11399

[Alon et al. 1995] Alon et al. Color-coding. J. ACM 1995;42:844-856

[Hüffner 2007] Hüffner. Algorithms and Experiments for Parameterized Approaches to Hard Graph Problems. Dissertation 2007;

[Scott et al. 2006] Scott et al. Efficient algorithms for detecting signaling pathways in protein interaction networks. J Comput Biol 2006;13:133-44

[Ideker and Sharan 2008] Ideker et al. Protein networks in disease. Genome Res. 2008;18:644-652

[Valentin and Hammond 2008] Valentin et al. Safety and secondary pharmacology: Successes, threats, challenges and opportunities. Journal of Pharmacological and Toxicological Methods 2008;58:77-87

[Noga et al. 1995] Noga et al. Color-coding. J. ACM 1995;42:844-856

[Hüffner et al. 2007] Hüffner et al. FASPAD: fast signaling pathway detection. Bioinformatics 2007;23:1708-1709

[Lindfors et al. 2009] Lindfors et al. Detection of Molecular Paths Associated with Insulitis and Type 1 Diabetes in Non-Obese Diabetic Mouse. PLoS ONE 2009;4:e7323

[Subramanian et al. 2005] 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 2005;102:15545-15550

[Multiple 2005] Multiple. Physician's desk reference. 2005

[Mattingly et al. 2006] Mattingly et al. The Comparative Toxicogenomics Database: A Cross-Species Resource for Building Chemical-Gene Interaction Networks. Toxicol. Sci. 2006;92:587-595

[Kuhn et al. 2008] Kuhn et al. STITCH: interaction networks of chemicals and proteins. Nucl. Acids Res. 2008;36:D684-688

[Kaushansky 2006] Kaushansky. Lineage-Specific Hematopoietic Growth Factors. N Engl J Med 2006;354:2034-2045

[von Vietinghoff and Ley 2008] Von Vietinghoff et al. Homeostatic Regulation of Blood Neutrophil Counts. J Immunol 2008;181:5183-5188

[Daniel et al. 2009] Daniel et al. Idiosyncratic drug-induced agranulocytosis: Possible mechanisms and management. American Journal of Hematology 2009;84:428-434

[Pessina et al. 2005] Pessina et al. Hematotoxicity Testing by Cell Clonogenic Assay in Drug Development and Preclinical Trials. Current Pharmaceutical Design 2005;11:1055-1065

[DrugBank] Drugbank.

[Berman et al. 2000] Berman et al. The Protein Data Bank. Nucleic Acids Res 2000;28:235-42

[Jensen et al. 2009] Jensen et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucl. Acids Res. 2009;37:D412-416

[Dandekar et al.] Dandekar et al.

[Valencia and Pazos 2003] Valencia et al. Prediction of protein-protein interactions from evolutionary information. Methods Biochem Anal 2003;44:411-26

[Jensen et al. 2004] Jensen et al. ArrayProspector: a web resource of functional associations inferred from microarray expression data. Nucleic Acids Research 2004;32:W445-W448

[Zhang et al. 2005] Zhang et al. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucl. Acids Res. 2005;33:W741-748

[Athanasiou and GERSON 2006] Athanasiou et al. Genetic Markers in the CSF2RB gene associated with an adverse hematological response to drugs. 2006

[Barreda et al. 2004] Barreda et al. Regulation of myeloid development and function by colony stimulating factors. Dev Comp Immunol 2004;28:509-54

[Welte et al. 2003] Welte et al. STAT3 deletion during hematopoiesis causes Crohn's disease-like pathogenesis and lethality: a critical role of STAT3 in innate immunity. Proc Natl Acad Sci U S A 2003;100:1879-84

[Touw and Bontenbal 2007] Touw et al. Granulocyte Colony-Stimulating Factor: Key (F)actor or Innocent Bystander in the Development of Secondary Myeloid Malignancy? Journal of the National Cancer Institute 2007;99:183-186

[Simon 2003] Simon. Neutrophil apoptosis pathways and their modifications in inflammation. Immunological Reviews 2003;193:101-110

[Geest and Coffer 2009] Geest et al. MAPK signaling pathways in the regulation of hematopoiesis. Journal of Leukocyte Biology 2009;

[Drutskaya et al. 2005] Drutskaya et al. Inhibitory effects of tumor necrosis factor on hematopoiesis seen in vitro are translated to increased numbers of both committed and multipotent progenitors in TNF-deficient mice. Exp Hematol 2005;33:1348-56

[Flanagan and Dunk 2008] Flanagan et al. Haematological toxicity of drugs used in psychiatry. Human Psychopharmacology: Clinical and Experimental 2008;23:

[Crawford et al. 2004] Crawford et al. Chemotherapy-induced neutropenia. Cancer 2004;100:228-237

[Desai et al. 2011] Desai et al. A systems biology approach for detecting toxicity-related hotspots inside protein interaction networks. IEEE Health Informatics and Systems biology 2011;

[Quackenbush 2003] Quackenbush. Genomics. Microarrays--guilt by association. Science 2003;302:240-1

[Walker et al. 1999] Walker et al. Pharmaceutical target discovery using Guilt-by-Association: schizophrenia and Parkinson's disease genes. Proc Int Conf Intell Syst Mol Biol 1999;282-6

[Zhang and Gant 2008] Zhang et al. A simple and robust method for connecting small-molecule drugs using gene-expression signatures. BMC Bioinformatics 2008;9:258

[Reich et al. 2006] Reich et al. GenePattern 2.0. Nat Genet 2006;38:500-501

[Flicek et al. 2008] Flicek et al. Ensembl 2008. Nucleic Acids Research 2008;36:D707-D714

[Ma et al. 2008] Ma et al. Sequence analysis of the SRGN, AP3B1, ARF6, and SH2D1A genes in familial hemophagocytic lymphohistiocytosis. Pediatric Blood & Cancer 2008;50:1067-1069

[Niemann et al. 2004] Niemann et al. Localization of serglycin in human neutrophil granulocytes and their precursors. J Leukoc Biol 2004;76:406-15

[Xia et al. 1995] Xia et al. Opposing Effects of ERK and JNK-p38 MAP Kinases on Apoptosis. Science 1995;270:1326-1331

[Boldt et al. 2002] Boldt et al. The role of MAPK pathways in the action of chemotherapeutic drugs. Carcinogenesis 2002;23:1831-1838

[Roth et al. 2011] Roth et al. Gene expression-based in vivo and in vitro prediction of liver toxicity allows compound selection at an early stage of drug development. J Biochem Mol Toxicol 2011;25:183-94

[Mongan and Hamadeh 2011] Mongan et al. Studying Organ-Specific Toxicity Using Gene-Expression Profiling. 2011

[Afshari et al. 2011] Afshari et al. The evolution of bioinformatics in toxicology: advancing toxicogenomics. Toxicol Sci 2011;120 Suppl 1:S225-37

[Hamadeh et al. 2002] Hamadeh et al. An overview of toxicogenomics. Curr Issues Mol Biol 2002;4:45-56

[Davis et al. 2011] Davis et al. The Comparative Toxicogenomics Database: update 2011. Nucleic Acids Research 2011;39:D1067-D1072

[Salimi and Vita 2006] Salimi et al. The Biocurator: Connecting and Enhancing Scientific Data. PLoS Comput Biol 2006;2:e125

[Davis et al. 2009] Davis et al. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks. Nucleic Acids Research 2009;37:D786-D792

[Tatonetti et al. 2012] Tatonetti et al. Data-Driven Prediction of Drug Effects and Interactions. Science Translational Medicine 2012;4:125ra31

[Diestel 2005] Diestel. Graph Theory. 2005

Appendix A: Graph Theoretic Definitions

This thesis adopts the graph-theoretic definitions proposed by Diestel [Diestel 2005]. A **graph** is a pair G = (*V*, E) of sets such that E Є [V^2] and V Ω E = ϕ . The elements of *V* are vertices (or nodes) of the graph G, the elements of E are its edges (or lines). The usual way to picture a graph is by drawing a dot for each vertex and joining two of these dots by a line if the corresponding two vertices form an edge.

The **degree** (or valency) $d_G(v) = d(v)$ of a vertex v is the number |E(v)| of edges at v ; by our definition of graph, this is equal to the number of neighbors of v. A vertex of degree 0 is isolated.

A **path** is a non-empty graph *P* = (*V*,E) of the form

$V = \{ x_0, x_1, x_2, \ldots\ldots\ldots x_k \}$ $\qquad\qquad$ $E = \{ x_0x_1, x_1x_2, x_2x_3 \ldots\ldots\ldots X_{k-1}X_k \}$

where the x_i are all distinct. The vertices x_0 and x_k are linked by *P* and are called its ends; the vertices $x_1, x_2, \ldots\ldots\ldots x_{k-1}$ are the inner vertices of *P*. The number of edges of a path is its length, and the path of length k is denoted as P^k. k is allowed to be zero. A path is often referred to by the natural sequence of its vertices, say $P = x_0x_1x_2 \ldots\ldots\ldots x_k$ and calling *P* a path from x_0 to x_k .

A **subgraph** of a graph G is a graph whose vertex and edge sets are subsets of those of G. A supergraph of a graph G is a graph that contains G as a subgraph. The graph is **directed** if its edges are directed (pointing toward either one of the ends) and undirected otherwise. A graph is complete (or called a clique) if every node has a connecting edge to every other node.

The **complete graph** on n vertices is often denoted by K_n where K_n would have n(n-1)/2 vertices.

The ***clustering coefficient*** is used to quantify the extent to which a node is a cluster member. For example, in a network, if a node is directly connected to five

neighbors, the clustering coefficient calculates the ratio of the number of direct links observed among the five neighbors over the number of possible direct links among the five neighbors. If all five neighbors are fully interconnected, the node is said to have a high clustering coefficient (value of 1) and vice versa (value of zero). Formally, the clustering coefficient is defined as

$$C_i = 2n / k_i(k_i-1)'$$

where *n* denotes the number of direct links connecting the *ki* nearest neighbors of node *i*. The clustering coefficient is 1 for a node at the center of a fully inter-linked cluster, while it is zero for a node that is part of a loosely connected group. A global measurement related to C_i is the average clustering coefficient *C* over all nodes in the network, characterizing the overall tendency of nodes to form clusters or groups.

VITA
Kaushal Desai

Education

2005-2012: PhD (c) – Info. Studies, Drexel University.
Thesis title (proposed): *Systems toxicology: discovering signaling paths to enable network medicine*

1998-2001: M.S – Biomedical Engineering, Drexel University, 1998-2001

Thesis title: An ultrasound image database for Breast cancer research

1998-2001: M.S – Information Systems, Drexel University, 1998-2001

1992-1996: B.E – Biomedical Engineering, Uni. of Mumbai, 1992-96

Teaching/Training Experience

- Member of a global team that trained ~1200 physicians and statisticians in clinical program design and interpretation at AstraZeneca Pharmaceuticals globally during 2011-2012.
- Guest Lecture: **Gene Expression Data Mining**, Drexel University, 2004.
- Guest Lecture: **Microarray data analysis**, University of Delaware, 2006.

Publications/ Presentations

- **Kaushal Desai** *et al.*, "Mining protein interactions and gene expression data to gain insights into drug-induced toxicity mechanisms," Proceedings of *IEEE Biomedicine and BioInformatics,* 2011.
- **Kaushal Desai** *et al.*, "A systems biology approach for detecting toxicity-related hotspots inside protein interaction networks," *Journal of Bioinformatics and Computational Biology,* 2011.
- **Best Paper Award:** IEEE-HISB 2011. **Kaushal Desai** et al., "Drug-Toxicity Signaling Path (DTSP) detection", IEEE-HISB conference, IBM Research Center, California, 2011.
- **Kaushal Desai**, "EHR/EMR Data Mining: From Therapeutic Targets to Effective Medicines", Panel Speaker at BioIT World Expo, Boston 2011.
- **Kaushal Desai**, Popov B, "Semantic annotation of clinical studies for knowledge-driven drug development", European Semantic Technologies Conference, Vienna, 2008.
- Yan Sun, Zhaohui Cai, **Kaushal Desai** et al., "Classification of rheumatoid arthritis status with candidate gene and genome-wide single-nucleotide polymorphisms using random forests", BMC Proceedings 2007.

- Xia L., Beaudoin J., Bui Y., **Kaushal Desai**. "Exploring characteristics of social classification", Proceedings of the 17th SIG Classification Research Workshop, 2006.