INTEGRATING GENOMIC DATA TO BUILD NETWORKS FOR PROTEINS AND SMALL MOLECULES

ANA BELL

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF
COMPUTER SCIENCE

ADVISOR: DR. OLGA G. TROYANSKAYA

NOVEMBER 2013

ABSTRACT

Advances in high-throughput genome-wide sequencing technologies have generated a massive amount of genomic data. Coupled with the ever-increasing performance of computing technologies, there is potential for a revolution in our knowledge of biology; hence the emergence of computational biology. A key goal of computational biology is to understand and model how biological processes work and to apply this knowledge to resolve complex human diseases. To that end, this thesis represents the work on two separate advances in network-based analysis of the large compendium of genomic data. We apply our knowledge of algorithms to the genomic data available in order to (1) build tissue and development specific gene interaction networks and (2) understand drug action on the molecular level.

The difficulties inherent in sequencing and functionally analyzing biologically and economically significant organisms have recently been overcome. *Arabidopsis thaliana*, a versatile model organism, represents an opportunity to evaluate the predictive power of biological network inference for plant functional genomics. Functional relationship networks are powerful tools that enable rapid investigation of uncharacterized genes. We provide a compendium of functional relationship networks for *A. thaliana*, leveraging data integration based on microarray, physical and genetic interaction, and literature curation datasets. To our knowledge this is the first work that includes tissue, biological process, and development stage specific networks, each

predicting relationships specific to an individual biological context. These networks summarize a large collection of *A. thaliana* data for biological examination. We found validation in the literature for many of our predicted interactions.

Functional networks and network-level pathway models thus represent an accurate and sensitive summary of the processes happening in the cell. In the second part of this thesis, we use these models to understand drug action. We integrate large amounts of heterogeneous data and build pathway-level networks that present interactions between compounds and proteins. We test our methodology in *Saccharomyces cerevisiae* (yeast). Our two step integration process, where we first predict protein-protein interaction networks for various protein-protein interaction types and then use these networks to predict protein-compound interaction networks, provide detailed insight into how pathway level knowledge can be leveraged to predict compound-level interactions.

ACKNOWLEDGEMENTS

I would, first and foremost, like to thank my advisor, Dr. Olga Troyanskaya for being a mentor in all stages of my graduate student career. Her advice and support was invaluable during highs and lows alike. I thank her for offering many helpful critiques of my thesis work and for encouraging me to pursue my passion for teaching through extracurricular lecture experiences. The Troyanskaya lab members, past and present, have also been an integral part of my grad school career, and I thank them for the insightful conversations, specifically Curtis and Arjun for their very helpful discussions and inputs. I want to thank Andrea LaPaugh, Mona Singh, Kai Li, and Thomas Funkhouser for agreeing to be on my thesis committee and for their insightful comments on my work.

I would not have made it this far without the support of my parents, Ana and Emil. They stressed the importance of higher education and are my perpetual cheerleaders, along with my sister, Cristina. Their presence in my life is immeasurable. Like the sun dries up the pavement seemingly instantly on a hot summer day after an east coast rainstorm, they brighten up my life instantly.

I am incredibly lucky to have met my best friend, companion, and husband, CJ. We instantly shared a connection and soon realized we shared many similar values. It is now hard to imagine being apart and I know that our adventures are just beginning.

I would like to end by acknowledging my friends, who have made my time at Princeton incredibly enjoyable. Aravindan, Aditya, Rajsekar, thanks for the late-night FIFA games. DJ, Noah, Jebro, Jeff, Jess, June, Justin, Rodolfo, Sam, and Seth, thanks (most notably) for kidnap the bride, Formal Thursday, and many other quirky outings. Aaron, Chris, and Vicky, thanks for the tennis matches and the entertaining (and sometimes bizarre) conversations. Ernie and Herb, thanks for trusting us with your home. Members of the CS softball team and captain Scott, thanks for the fun games. Last but not least, I thank Melissa, who has been a superb graduate coordinator from my first visit day to my FPO.

This work was supported by NSF CAREER award DBI-0546275; NIH grants R01 GM071966 and T32 HG003284; and NIGMS Center of Excellence grant P50 GM071508.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION AND BACKGROUND

Big data is a hot topic that is gaining popularity in many fields such as marketing, economics, web searches, as well as computational biology. The large amount of data available for many different organisms being studied is overwhelming. Making sense of it all in a genomics setting is an intimidating and sometimes daunting task. This manuscript will discuss two contributions to the field of functional genomics leveraging large datasets.

1.1 Motivation

Living organisms are made up of DNA, which is the genetic code used to encode proteins. The mechanisms and the details of this process and gene functions are still not adequately understood. In the past few years the explosion of data related to this problem has helped expand our knowledge of biological networks. Biologists are at the forefront of generating the data to narrow the gap between our knowledge and the biological truth. Yet there are several limitations they face; the brute force method of performing experiments is slow, despite the advent of large-scale genome sequencing technologies; each experiment is costly, and even though gathering, storing, and analyzing data is now cheap, we still need better analytical methods to understand the complexity of biological networks from the overwhelming amount of data. To that end, we are pursuing data-driven biology. Biologists perform experiments that cost them time and money, while computer scientists use computational techniques to

analyze large quantities of data; with the help of computer scientists, the search space of biological hypotheses for biologists can be quickly narrowed from essentially infinite to something more tangible and attainable. The interdisciplinary collaboration between these two fields, working towards a common goal of understanding functional protein networks and human gene function, is necessary and invaluable.

The cyclic process of computational biology seems ideal: biological experiments yield gold standard (known biological truth) data, which in turn is used by computer scientists to generate hypotheses that are then relayed back to biologists to be verified, and true predictions are appended to the gold standard. Yet, there exists a disconnect between this theoretical view and what is done in practice, where predictions are not followed up on. Other issues arise as well. Gold standards may be incorrect to begin with, because after all, they are annotated by humans. Further, some experimental procedures may produce false positives or false negatives, which would further harm the gold standards, the very thing that we rely on as the ground truth. Nevertheless, we are beginning to realize the importance of careful annotations, and are making an effort, now more than ever, to bridge the gap between these two worlds.

The sheer amount of genomic data that has become readily available over the past few years is increasing. Early high-throughput technologies began in the 1990s, and allowed the process of sequencing genes to be parallelized. These next-generation sequencing technologies spurred a revolution in the sequencing world. For example, the amount of genetic sequencing data that has been stored at the one of the largest

bioinformatics institutes since then, European Bioinformatics Institute (EBI), has grown exponentially, according to their 2012 annual report; the number of nucleotide sequences grew from a few million in 2008 to 200 trillion in 2012 in part because the number of genome sequenced grew from 50 in 2008 to 450 in 2012 [1]. We observe a similar trend in the amount of experimental data available from the Gene Expression Omnibus (GEO) database [2].

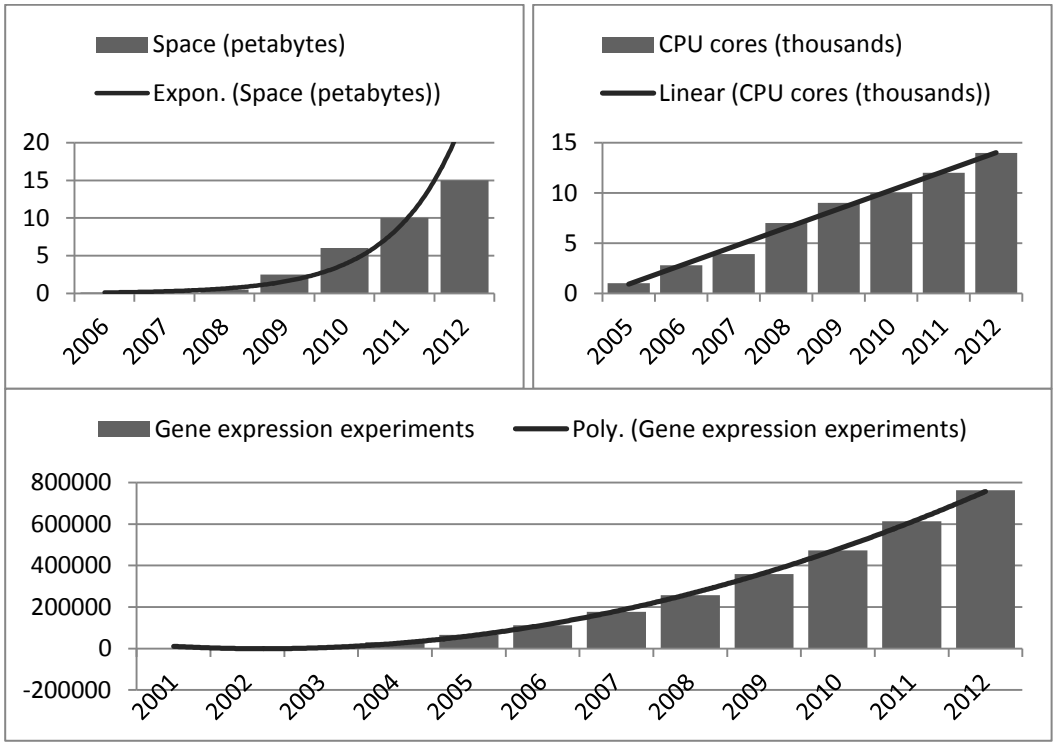

Figure 1. EMBL-EBI space, CPU cores, and GEO experiments
The amount of space devoted to genomic data at EMBL-EBI has grown exponentially, while the number of CPU cores allocated to analyzing said data has grown roughly linearly. The number of microarray experiments in GEO has also grown exponentially.

The diversity of the genomic data being produced is equally notable. There is, first and foremost, the raw biological data (sequences) that comes from sequencing the

genomes of various organisms. Biological techniques like mass-spectrometry and computational techniques like gene prediction algorithms can divide an organism's sequence into masses of molecules, which we call genes. With this basic information and careful biological experimentation, our genomic data knowledge grows; we can annotate types of interactions that happen between pairs of proteins; we can observe how small mutations in sequences affect the function of the gene, the phenotype of an organism, or the impact it has on a specific disease. Further still, if we introduce more complex algorithmic tools, our data set diversifies and expands even more; we can calculate, on a large scale, the similarity between sequences within or between organisms to build phylogenetic trees that show evolutionary distances between organisms; we can even predict interactions between genes that form complexes to achieve a certain function in the cell. In this thesis, we are interested in predicting these functional interactions and presenting them in the form of a fully-connected gene-gene interaction network.

1.2 Genomic Data and Gold Standards

Our computational methods use experimental data under a supervised learning framework to predict gene function and drug targets. Our experimental data consists of microarrays that provide measurements of expression levels of genes simultaneously [3], protein domains that are functional subunits [4], and physical interactions where individual genes bind together to work towards a common goal in the cell [5] [6] [7].

1.2.1 Ontologies

Gold standards (known biological truths) for protein functions come from the Gene Ontology (GO) [8]. This ontology is organized as a directed acyclic graph, which aims to provide a vocabulary of terms and to curate functional interactions between genes into three main categories: biological process, molecular function, and cellular component. In this project we look at the biological process branch, as it contains GO terms related to the function of cells, tissues, and organs. To define our gold standard, we choose a subset of GO terms that are specific enough to have biological meaning but broad enough to have sufficient genes annotated to them. For our gold standard we consider genes that are annotated to the same GO term as functionally related; based on concrete experimental evidence, we believe functionally related genes work in concert to perform a certain task in the cell.

1.2.2 Experimental Data

The experimental data used in our studies consists of microarrays, protein domains, sequence similarity, and co-membership in protein families.

A microarray is the result of a series of experiments done to determine the expression levels of genes. We begin with a physical platform on which known sequences of DNA, called probes, are placed. There can be as many at 1 million probes on one slide. Biologists prepare complementary DNA (cDNA) from the sample, which they tag with a fluorescent dye. When the slide is washed with the sample cDNA, certain

bindings to associated probes will occur, depending on the level of expression in the sample. Those that successfully bind are measured by a scanner reading the fluorescence intensity and normalized. This intensity may be compared to some control intensity that was measured under specific conditions. The difference between the sample and the control measurements represents how much each gene is expressed under the sample condition.

Protein domains are portions of a protein believed to be important enough that they are often conserved across different species. They are generally independently stable, meaning that they can fold to create a stable structure that can then be rearranged or swapped with other domains. Proteins that share the same domains are likely to share the same function since domains are often the part of the protein that interacts. Protein families were derived from the Pfam database [9] and are often constructed based on domains they share. These families represent functionally conserved regions of a genome sequence. This database is divided into Pfam-A (curated) and Pfam-B (lower quality since families are automatically generated using an algorithm).

Because similar genetic sequences equate to similar protein sequences, sequence similarity is another indicator for functional similarity. Sequence similarity is a measurement between pairs of genes. Basic Local Alignment Search Tool (BLAST) [10] [11] allows us to measure sequence identity by scoring how well a pair of proteins aligns to each other. Only pairs that meet a threshold limit are believed to be sufficiently similar.

Not all of this data is reliable. Microarrays can be inconsistent, especially if experimental conditions are not very well controlled, leading to batch effects and technical noise. Protein domains are believed to be evolutionarily conserved, but are not guaranteed to have the same function or take part in the same biological process. Similarly, although sequence similarity is important for function prediction, genome-scale experiments may forgo specificity (the true negative rate of predictions) or sensitivity (true positive rate of predictions). Because of possibly noisy datasets, each dataset, on its own, is crippled from providing accurate gene function predictions. Data integration improves the accuracy of predictions while keeping a good balance between a reasonable sensitivity and specificity. As biologists generate more and more high throughput datasets, even though they designed experiments with a specific hypothesis in mind, there is more biological information in the dataset that we can exploit. We discuss some techniques in the next section.

1.3 Functional Networks

Although a great deal of genome-scale data exists, functional annotations for genes are largely incomplete. In this work, we use functional networks to predict gene function. A functional network is a fully connected network defined via a graph data structure; nodes will represent genes and edges represent functional relationships between genes, which are inferred via computational techniques, such as Bayesian integration. Nodes are connected by an edge, which is weighted with the probability that the genes perform the same function or work toward the same goal in the cell. Performing the

same function can mean that they physically interact to form a larger entity or that they have a similar structure necessary for function. This network structure lends itself to the notion of guilt-by-association, which states that genes with unknown functions can be inferred by surrounding genes with known function.

1.3.1 Prior Work

Much progress has been done in this field. Building functional networks relies on identifying patterns in the abundant genomic data; this is most efficiently realized with machine learning and data mining techniques. There is no one technique that is guaranteed to work, and all have merits and drawbacks; Bayesian networks [12] [13] [14] [15] [16] and support vector machines (SVMs) [17] [18] [19] are two of the most popular techniques, and we will discuss them both in this thesis.

One of the first papers to do probabilistic data integration on a large scale and take advantage of heterogeneous data sources was published in 2003. Troyanskaya et. al. presented a framework called MAGIC (Multisource Association of Genes by Integration of Clusters) [12]. They used Bayesian networks to increase accuracy of protein function predictions by combining abundant microarray data with pairwise data from non-microarray sources such as known protein-protein interactions, genetic interactions, and sequence similarity. Their work was extended in 2007, where the notion of context-specific functional networks was introduced as bioPIXIE (biological Process Interface from eXperimental Interaction Evidence) [20]. In this work, instead

of a single global network that leverages heterogeneous genomic data to predict functional interactions, several networks were used to predict functional interactions in diverse biological contexts. These networks led to new insights into biological function; large datasets contain information relevant to multiple biological contexts and given any specific context only a little information from the dataset relevant. In a specific context, discarding some dataset information means decreasing the number of negative examples, which in turn, decreases the pool of potential false positives. This work allowed biologists to view how genes interacted in several biological processes, and so now functional interactions that were once diluted in a global network became more apparent when analyzed in various contexts. In Chapter 2 of this thesis, we further extend this model and the notion of biological contexts to include not only biological processes, but also organism tissues and developmental stages.

Drawing upon the notion of context-specific functional interactions, Barutcuoglu et. al. incorporated the hierarchical structure of biological process contexts to predict genes annotated to specific biological processes [19]. They trained SVMs on multiple data types and combined their predictions in a Bayesian framework; the framework took advantage of the fact that the biological contexts they were interested in predicting for had a hierarchical structure, which was modeled as a Bayesian network. Their framework was the basis for Park et. al. [21], who also used the hierarchical method of improving predictions to create mechanistic networks; these networks used interaction types as contexts. In Chapter 3 of this thesis, we extend these networks by adding another type of dataset to the already rich dataset group and we use these

mechanistic networks to provide insight into how proteins interact with small molecules (compounds).

1.4 Drug Target Prediction

One of the more challenging problems in the field of bioinformatics is understanding how diseases work and how the introduction of a drug alters the function of its target protein and possibly an entire biological pathway (a group of proteins that perform a series of actions to achieve a certain task in the cell). Drugs are typically small molecules that are carefully manufactured such that they bind to interact, given a high enough affinity, with specific protein targets. This drug-protein interaction aims to adjust the behavior of proteins to address and possibly correct a malicious action in the cell. The problem of an extensive drug-protein pair search space slows the progress of biological experimentation, so machine learning techniques are necessary in the discovery of drug-protein interactions and, ultimately, in drug discovery.

1.4.1 Prior Work

Drug-protein interactions have been both experimentally validated and computationally predicted. Several techniques to solve this problem have been implemented in three broad ways: data mining for chemical properties, observing phenotypic effects of drugs on targets, and visualizing the available data in some intuitive way. The pharmacological space is comprised of chemical and genomic information such as the genomic sequence, 3D structure, and numerical traits like

molecular weight and size. These traits have been utilized in the context of bipartite graphs to create classes of drug-target interaction networks [22] and to distinguish between etiological drugs (targeting the cause of disease) and palliative drugs (alleviating pain without curing the disease) in humans [23]. Sequences and 3D structures have been aligned to determine similar small molecules and infer similar binding partners [24] [25]. Even similarity of target proteins to drugs has been incorporated [26] [27] for predicting targets by leveraging ligands annotated to the drug targets [28]. Bayesian networks, as a probabilistic approach, have been applied on specific small molecules such as kinase inhibitors [29] and larger compound sets, while still using these chemical properties [30]. Biochemical data was used in conjunction with a support vector machine classifier to predict interaction probabilities for an input drug-protein pair [31] [32]. Phenotypic effects after drug administration have been analyzed by gene expression profiles and have been shown to be useful in identifying and relating genes to drugs [33] [34] [35]. In yeast, drug side-effect profiles and protein-binding profiles were simultaneously analyzed using sparse canonical correlation to show a relationship between side effects and drug-targeted proteins [36]. Using protein-protein interaction networks, disease-related networks, and literature mining, another group built a drug-protein connectivity map and showed that molecular signatures differed between different drug classes and diseases [37]. More generally, a few tools aim to visualize and present a global network of the chemical-protein data gathered from numerous databases [38] [39] [30] [23].

In these prior investigations into the problem of drug-protein interaction prediction, the focus is generally on the various properties of drugs, but there is little incorporation of functional information and no attempt to integrate a large amount of known heterogeneous data relating to how proteins are functionally related to other proteins. Leveraging this interaction data will be useful for determining whether a drug interacts with a particular protein and is especially valuable when determining the effect that a drug will have on an entire pathway. The challenge is to construct drug-protein networks that give insight into the types of protein-protein interactions that contribute to the drug-protein interactions. In Chapter 3, we discuss how we leveraged high-throughput heterogeneous data, sequence data, and 3D structural data to first construct protein-protein mechanistic networks in a genome-wide setting, and then how we used these networks to infer interactions between drugs (and other compounds) and proteins.

1.5 Biological Questions in this Dissertation

We extend previous machine learning techniques in order to help biologists answer two pressing biological questions by providing informed insights to narrow down their experimental search space. This manuscript will focus on the following two questions:

1. How do proteins behave in different biological processes, tissues, and developmental stages in the plant *Arabidopsis thaliana*?

Finding protein function is an important problem in bioinformatics. High-throughput technologies have helped tackle the initial problem of determining what the sequences of a particular genome are. However, the sequence of an organism is invariant and does not incorporate different biological contexts nor does it tell us about the function of genes. Discovering gene function will help us make headway into fields such as personalized medicine and disease treatment. Despite vast resources being spent on this important problem, it has proven difficult to completely pin down. First, relying on biological experimentation can only go so far because experiments are costly and take too much time; computational methods alleviate this burden. Second, there is a lot of data with varying quality and good machine learning techniques are necessary to analyze them quickly and efficiently. Myers et. al. have shown that in *Saccharomyces cerevisiae* the same group of proteins will interact differently in different biological processes [20] (for example, response to stress vs. inflammatory response). In Chapter 2, we look at the model organism for plants and show that context sensitive networks improve gene function prediction over a globally predicted network. We extend the model to include other biological contexts such as different tissues of the plant (for example, roots vs. leaves) and developmental stages of the plant (for example, shoot emergence vs. rosette growth). The first section of Chapter 2 describes the Bayesian integration method we used and how we incorporated the plant gold standards and data in the integration pipeline. The second section of Chapter 2 is a discussion of our results and we show that (a) context-specific networks are able to predict functional interactions better than a global network, (b) that several predictions we make in context-specific networks are validated in recent literature, and (c) that our method for

up-weighting or down-weighting certain datasets leads to different contributions of these datasets to particular functional interactions. By incorporating tissue and developmental information in functional predictions we have shown that proteins function changes not only in different biological processes but in different parts of the organism and at different times in the organism's life cycle.

2. How do small molecules and drugs interact with and target different proteins in *Saccharomyces cerevisiae*?

With a better understanding of protein function, we can leverage our predictions to investigate more complex biological problems. Understanding disease is a particularly important one, as we are moving towards the idea of personalized medicine. While humans have 99% of our DNA in common with each other, the remaining 1%, along with environmental factors and lifestyle choices, still leaves much variability and is likely to be the key to understanding disease. To that end, we must first understand the mechanism and effects of protein-drug interactions. Because humans are such complex multicellular organisms, we start with a simpler model organism like yeast, which has been heavily studied. This translates to a larger gold standard pool and, ultimately, a better understanding of the organism. In Chapter 3, we discuss our methodology to find drug targets in yeast. Future work can be to expand our methodology further and find drug targets in more complex organisms like humans. Furthermore, while our primary goal is to investigate the effects drugs have on proteins, we also include other small molecules (compounds) such as ethanols and

sugars to yield a richer compendium of compound-protein interactions. Previous work in this field has often focused on using properties of the compounds in order to arrive at conclusions about how other compounds interact with proteins. We continue this trend but, in addition, leverage data that we create in the form of protein-protein interaction predictions for various interaction types. These mechanistic networks provide further insight into the cellular wiring that may be shared among proteins interacting with the same compound. In the first section of Chapter 3, we describe the methodology that we use to arrive at these compound-protein interaction networks, beginning with how we construct our mechanistic networks and ending with how we incorporate that data into our compound-protein interaction predictions. In the second section of Chapter 3, we discuss our findings and interesting observations arising from the analysis of compound-protein interaction networks.

2 INTEGRATED FUNCTIONAL NETWORKS OF PROCESS, TISSUE, AND DEVELOPMENTAL STAGE SPECIFIC INTERACTIONS IN ARABIDOPSIS THALIANA

Recent years have seen an explosion in plant genomics, as the difficulties inherent in sequencing and functionally analyzing these biologically and economically significant organisms have been overcome. *Arabidopsis thaliana*, a versatile model organism, represents an opportunity to evaluate the predictive power of biological network inference for plant functional genomics.

Plants are complex and diverse organisms and have adapted evolutionarily to almost every ecological niche on the planet. They have surpassed many evolutionary challenges so that they can populate different areas of the earth; they are immobile so moving pollen is a big problem for them, and they have to derive their own food. Despite, or because, of these challenges, they have become well-studied in an agriculture setting. Plants have several key uses. First, they are studied for the use of pharmaceuticals. A quarter of all medicines contain ingredients derived from plants. More recently, biotechnology has allowed researchers to modify plants to have specific therapeutic proteins. Second, plants are well-known natural air filters. Through photosynthesis, they absorb carbon dioxide and release oxygen to the

environment. Within the plant, the carbon reacts with water to form formaldehyde. This is condensed and combined with vitamins to form sugar and starches. Photosynthesis captures less than 2% of light energy so if this act is manipulated to be more efficient, we can generate more yield for biorefineries. Agricultural and pharmaceutical applications of plant genomics have focused on understanding the metabolic and biochemical potential of specific plant tissues and environmental responses [40]. *Arabidopsis thaliana*, also known as thale cress or mouse-ear cress, is the most common model organism for plants, with a short life cycle (6 weeks to go from germination to maturity so about 8 generations can be studied within a year), relatively few genes (about 28,000 genes, compared with ~6,000 genes for yeast and ~25,000 genes for human), and a fully sequenced genome [41]. It is a multi-cellular organism with multiple tissue types and developmental stages, and much of its tissue-specific and stage-specific molecular biology has yet to be determined.

Many *A. thaliana* gene products are functional only in a specific tissue or during a specific developmental period [42] [43]. The ability to predict tissue- or development-stage-specific function from genomic data would aid in appropriately targeting experimental work; doing experiments on every plant structure at each of its development stages individually would be tedious and costly. Additionally, it would be challenging to summarize the resulting genomic data efficiently, since the combinatorics of 30 developmental stages [44] by over 50 plant structures [45] makes a large compendium of predictions unwieldy as raw data. With this as motivation, we

have created probabilistic networks providing a data-driven view of protein functional relationships in *A. thaliana*.

In this chapter, we provide a compendium of functional relationship networks for *A. thaliana* leveraging data integration based on over 60 microarray, physical and genetic interaction, and literature curation datasets. These include tissue, biological process, and development stage specific networks, each predicting relationships specific to an individual biological context. These biological networks enable the rapid investigation of uncharacterized genes in specific tissues and developmental stages of interest and summarize a very large collection of *A. thaliana* data for biological examination. We found validation in the literature for many of our predicted networks, including those involved in disease resistance, root hair patterning and auxin homeostasis.

These context-specific networks demonstrate that highly specific biological hypotheses can be generated for a diversity of individual processes, developmental stages, and plant tissues in *A. thaliana*. All predicted functional networks are available online at http://function.princeton.edu/arathGraphle.

The work presented in this chapter is published in [46] and includes contributions from Curtis Huttenhower, Anjali Iyer-Pascuzzi, Philip N Benfey, and Olga Troyanskaya. Philip and Anjali performed the laboratory experiments, Curtis and Olga conceived the study, and Olga supervised the project.

2.1 Methods

In addition to producing global functional networks summarizing the general interactions occurring among *A. thaliana* genes, we performed additional integrations re-weighting the data to emphasize various cellular, developmental, and tissue-specific processes. Each integration is defined by one or more curated gold standards [47], each listing genes whose products are known to be active in the areas of interest (e.g. the photosynthesis pathway, dry seed developmental stage, or leaf tissue). By learning how informative each dataset is with respect to each gold standard, we re-weighted the datasets and combined them to infer a single genome-wide functional network in each context of interest. Our methods generated three types of gene functional networks:

- GLOBAL-PROCESS network represents functional interactions between genes on a global scale in the context of all biological processes
- GLOBAL-DEVEL network represents functional interactions between genes on a global scale in the context of all developmental stages
- PROCESS-DEVEL networks represent functional interactions between genes in contexts characterized by the intersection of biological processes and development stages
- TISSUE-DEVEL networks represent functional interactions between genes in contexts characterized by the intersection of tissues and development stages

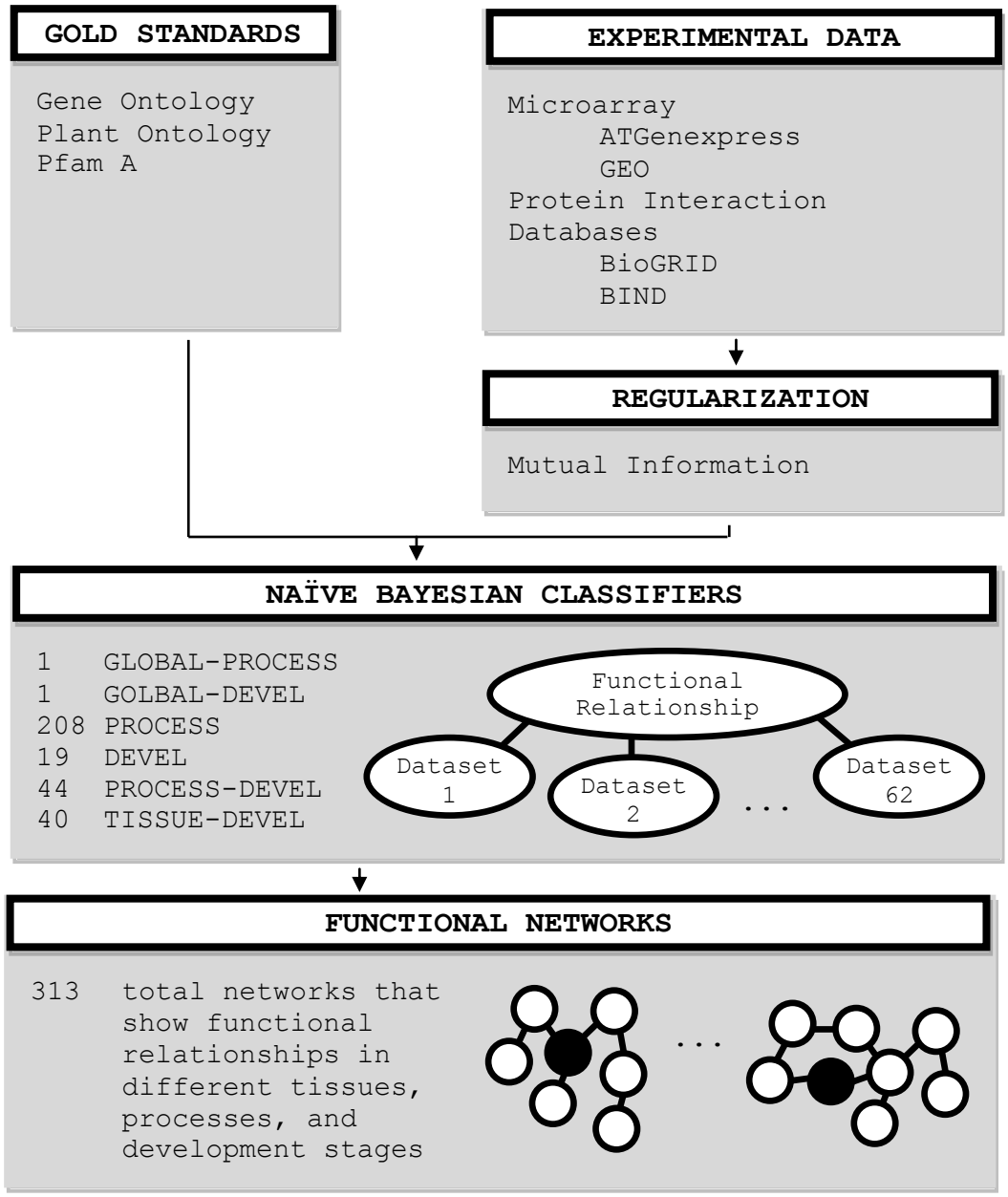

Figure 2. Schematic of the process, tissue, and developmental stage specific genomic data integration pipeline.
We used regularized Bayesian classifiers [48] to integrate genome-scale data for *A. thaliana* including 55 expression datasets from GEO [49] and 5 physical and genetic interaction datasets from BIND [50] and bioGRID [51]. Using curated biological knowledge from the Gene Ontology [8], Plant Ontology [45], and Pfam [9], we reweighted these datasets to infer genome-wide biological networks focused on individual biological processes, developmental stages, and plant tissues.

2.1.1 Gold Standard Generation

We created three gold standards, each containing subsets of positive (related) and negative (unrelated) protein pairs. For the GLOBAL-PROCESS standard, we selected a set of interesting terms from the Gene Ontology as described by [47]. Briefly, they used expert curation to choose an evaluation standard from the GO hierarchy that had sufficient gene annotations to each term, yet was specific enough to have biological significance. Gene pairs co-annotated to one of these terms were considered to be related, and pairs containing genes annotated to some term (but not co-annotated) were considered to be unrelated. The exact method is described in [13]; assuming that the number of annotations to a term approximately corresponds to the term's biological specificity, a gene pair is selected as a positive if both genes are co-annotated to a GO term with less than 300 annotations. Negative examples were chosen randomly as in [15], to yield ten times as many negatives pairs than positive ones. This resulted in 188,343 positive and 1,183,813 negative pairs in the GLOBAL-PROCESS standard.

The GLOBAL-DEVEL standard was created similarly, save that genes were required to be co-annotated to a development stage in the Plant Ontology (PO). These gold standards were decomposed into subsets for the PROCESS and DEVEL compendia by limiting positive pairs to individual processes and development stages, respectively, and randomly sub-sampling ten times as many negatives. The PROCESS-DEVEL and TISSUE-DEVEL standards intersected these process- and developmental-stage-

specific gold standards with an identically generated tissue-specific standard using 43 PO tissue terms.

2.1.2 Bayesian Data Integration

A naïve Bayesian classifier is a graphical model with a simple structure, as outlined below. The top node is the classifier node – what we want to predict. Assuming independence between our datasets, given the classification, we can then represent each dataset as a node that is independently influenced by the top node.

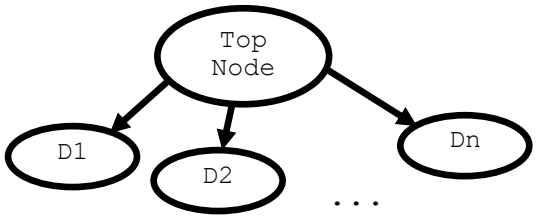

Figure 3. Naive Bayesian Classifier Diagram
Each node represents an event. The top node influences the observed data nodes and the probability of any of the observed data events D1…Dn happening is only dependent on the TopNode.

The formula at the heart of Bayesian networks is $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$, which calculates the posterior probability of A given B as the probability of observing B given A multiplied by the prior probability of A and divided by the probability that B occurs. We can put this formula in the context of the network above as

$$P(TopNode|D1, D2, \dots Dn) = \frac{P(D1|TopNode)P(D2|TopNode)\dots P(Dn|TopNode)P(TopNode)}{P(D1,D2\dots Dn)}$$

since we assume independence of the datasets. Each dataset has a conditional

probability table (CPT) associated with it. In our work, the table is discretized into bins depending on what data we are looking at: binary data has two bins while microarray expression gets seven bins corresponding to various levels of expression levels. The CPT is initially populated by keeping a count of how many of the dataset records occur in the $TopNode$ records. Knowing the total number of records for each dataset, we can find out the probability distribution of our dataset given $TopNode$, namely $P(D1|TopNode)$, $P(D2|TopNode)$, …, $P(Dn|TopNode)$. We also know $P(TopNode)$ as the prior (total number of positives divided by total number of negatives) and $P(D1, D2 \dots Dn)$ as the probability that a record (specific interaction pair) is in those datasets. This method of calculating the posterior probability has been shown to work well and is fast because it uses simple counting as opposed to being an iterative model.

The Bayesian network described was applied to the earliest networks [12] by using GO annotations as the gold standard, yielding a binary score for proteins known to interact. The datasets were reduced to a pairwise score as well; microarrays gene expression data pairs, for example, were obtained by Pearson correlation measures for $X = (x_1, x_2, \dots x_N)$ (set of data points) for one gene and $Y = (y_1, y_2, \dots y_N)$ for another gene $\rho = \frac{\sum z_x z_y}{N}$ where, N is the total number of pairs, and z the normalized score for each pair (x_i, y_i). Scores across different datasets were normalized using a z-score where $z = \frac{x-\mu}{\sigma}$ where μ is the mean and σ the standard deviation.

To extend the framework to work in context-specific networks, a separate Bayesian classifier was constructed for each context [20]. Between these classifiers, the gold standard was the one thing that changed; instead of using the entire GO as the positive and negative set, subsets of GO were used, each subset representing the relevant positive and negative interactions in that particular context.

Each functional relationship network was predicted by a corresponding Bayesian classifier trained as detailed in [13] and [48]. A naive classifier was constructed for each gold standard as described above: one each for GLOBAL-PROCESS and GLOBAL-DEVEL, 208 PROCESS terms from the Gene Ontology, 19 DEVEL terms from the Plant Ontology, and 40 PROCESS-DEVEL intersections and 44 TISSUE-DEVEL intersections (each containing at least 10 genes).

Each classifier integrated the same data, broadly comprising of co-expression data, protein sequence families, and physical and genetic protein-protein interactions 58 microarray datasets that were gathered from AtGenExpress [52] [53] [54] [55] [56] and GEO [49]. This data was converted into pairwise scores by Pearson correlation, z-transformation to obtain a normal distribution $Z = \frac{1}{2} log \frac{1+p}{1-p}$, and z-scoring to distribute this with mean 0, standard deviation 1 for each dataset. These co-expression scores were discretized into 7 bins from $-\infty$ to -1.5, -1.5 to -0.5, -0.5 to 0.5, 0.5 to 1.5, 1.5 to 2.5, 2.5 to 3.5, 3.5 to ∞. Protein families were drawn from the automatically generated PFam B [9], and protein interactions were taken from BIND [50], BioGRID

[51], computational predictions and enzyme assays used for functional annotations [57], and annotations extracted from literature in TAIR (The Arabidopsis Information Resource) [58]; all were quantified as binary variables to indicate the presence or absence of an interaction. This resulted in 60 total datasets integrated in each classifier.

2.1.3 Regularization Using Mutual Information

Naive Bayesian classifiers assume that all datasets are independent, which becomes increasingly less true as large amounts of biologically similar data are integrated. As detailed in [48], this leads to overconfident and less accurate predictions, which we resolve without loss of efficiency by regularizing the naive classifiers. This process mixes in a uniform prior with weight exponentially proportional to the amount of information shared by each dataset, thus down-weighting datasets with less unique information. Mutual information was calculated between each pair of datasets $I(D_k, D_i)$ using the discretization described above and, for each dataset pair, converted to a fraction by dividing by the total amount of possible shared information

$$I'(D_k \cdot D_i) = \frac{I(D_k \cdot D_i)}{\min(H(D_k), H(D_i))}$$

These fractions were summed for each dataset, $U_k = \sum_{i \neq k} I'(D_k \cdot D_i)$ and exponentially weighted as $\alpha_k = 2^{U_k + 1} - 1$. In combination with Laplace smoothing tunable with parameter $\beta_k = 2$, this yields a regularized classification probability between genes g_i and g_j:

$$P_{i,j}(FR) \propto \prod_{k=1}^{n} \frac{\beta_k \left| D_k = d_k\left(g_i, g_j\right)\right| + \alpha_k}{\beta_k |D_k| + \alpha_k |d_k|}$$

2.1.4 Computational Performance Evaluation using Cross Validation

Cross-validation is a statistical technique used to evaluate how well the parameters fit for a particular model perform on a separate held-out evaluation set. In this thesis, we use *k*-fold cross validation. The original data is divided into *k* equal-size partitions. *k*-1 partitions represent the training set given to the model and the remaining set is used to evaluate the model predictions. This procedure is repeated by taking each of the single partitions as the evaluation set and averaging the evaluation results across all *k* runs.

When evaluating how well an evaluation set does, we use $precision = \frac{TP}{TP+FP}$ (how many positives are returned at some cutoff), $recall = sensitivity = \frac{TP}{TP+FN}$ (how many positives are returned from all possible positives), and $specificity = \frac{FP}{FP+TN}$ (false positive rate), where $TP$is the true positives, FP is the false negatives TN is the true negatives, and FN is the false negatives. The results from a prediction are typically ranked high to low. When we predict on the evaluation set, we iterate a cutoff starting from the top prediction and move down the returned list, at each cutoff calculating the above values. A precision recall plot will typically show the tradeoff as we move the cutoff lower; at low recall we expect high precision and at high recall we expect low precision. We can also come up with an overall score called the area under the receiver operator characteristic (AUC). This score, from 0 to 1, is often used as a

summary statistic and it is the area under the specificity-sensitivity curve. An AUC of 0.5 means that the accuracy of the model is the same as an arbitrarily guess. In practice, good models that are biologically informative will have AUC values greater than 0.65.

We randomly withheld 20% of genes from the positive pairs and 20% from the negative pairs in our gold standard set, using any gene pair including at least one of these genes as a test set excluded during training. All performance evaluations were performed exclusively on test sets selected this way using 5-fold cross validation.

2.2 Results

Here, we investigate over 300 resulting global and context-specific functional networks generated for *A. thaliana* biological processes, tissues, and developmental stages. We analyzed the resulting networks as detailed below to generate novel biological hypotheses. We evaluated these networks computationally to determine the accuracy of their predictions, and we found that genomic datasets are differentially informative across varied contexts. Gene products' predicted roles and interactions also varied, and we found validation in the literature for specific interactions for many proteins. We highlight several of these interactions for a diversity of developmental and physiological processes, including those for PHOSPHOENYL PYRUVATE/ PHOSPHATE TRANSPORTER 2 (AtPPT2) during leaf and root developmental stages, the disease resistance proteins RESISTANCE TO PSEUDOMONAS 1 and 2

(RPS1 and RP2), the root epidermal patterning protein WEREWOLF (WER), and the auxin hormone receptor TRANSPORT INHIBITOR RESPONSE 1 (TIR1). Finally, we provide an intuitive, interactive representation of these results online at http://function.princeton.edu/arathGraphle.

2.2.1 Overview of Integrated Functional Networks Inferred for *A. thaliana* Pathways, Tissues, and Developmental Stages

We generated a range of networks (Table 1) addressing questions of increasing specificity regarding *A. thaliana* gene pair relationships. First, this includes two global functional networks representing overall relationships occurring within the *A. thaliana* genome independent of a specific tissue or developmental context. The first, GLOBAL-PROCESS, links genes with high probability if the integrated genomic data indicate that they are employed by the organism in similar biological roles; that is, if they participate in the same cellular processes. The second network, GLOBAL-DEVEL, links genes if they are expected to be co-active during the same developmental stage(s). We additionally inferred two compendia of context-specific networks, each describing functional relationships between genes predicted to occur only during a specific biological process or developmental stage. Creating biological process-specific networks (i.e. context-specificity) has been explored for the yeast and human genomes [59] [20] and provides a more specific view of genes and their functional interactions tailored to individual biological areas of interest. Here, we

expand context-specific inference to include developmental stages and plant tissues in addition to biological processes and pathways.

Table 1 Global and context-specific functional relationship networks.

COMPENDIUM TYPE	COMPENDIUM DESCRIPTION	# OF NETWORKS	EVALUATION (AUC RANGE)
GLOBAL-PROCESS	Global functional network linking genes active in similar biological pathways and processes	1	0.54
GLOBAL-DEVEL	Global functional network linking genes active in the same developmental stage(s)	1	0.63
PROCESS	Networks linking genes active in similar pathways only within the context of each specific biological process	208	0.46 – 0.79
DEVEL	Networks linking genes active in similar developmental stages only within the context of each specific developmental stage	19	0.43 – 0.74
PROCESS-DEVEL	Networks linking genes active in the same pathways during the same developmental stage	40	0.46 – 0.82
TISSUE-DEVEL	Networks linking genes active in the same plant tissues during the same developmental stage	44	0.5 – 0.78

As described in Table 1, this resulted in the PROCESS and DEVEL compendia of networks. Each PROCESS network represents the functional relationships predicted to occur during a specific biological process (e.g. “autophagy”, the “cell cycle”, “photosynthesis”, and so forth), and genes linked with high probability are expected to co-participate in this process. Each DEVEL network represents a plant developmental

stage (“germination”, “senescence”, etc.), and genes linked with high probability are expected to be co-active in that stage.

Finally, in order to investigate the interactions among biological processes, temporal developmental stages, and spatial locality in tissues, we generated two additional network compendia. The first, PROCESS-DEVEL, includes 40 networks each specific to a process/developmental stage pair (e.g. photosynthesis during leaf senescence). Only 40 of the ~4,000 possible pairs were analyzed due to a lack of curated training data for the remaining process/stage combinations. Similarly, the TISSUE-DEVEL compendium includes 44 networks, each predicting gene pairs expected to be co-active in a specific tissue location and at a specific time during development. All networks in these compendia were inferred using probabilistic Bayesian reweighting of 60 genomic datasets, and the results are analyzed in detail below.

2.2.2 Context-Specific Data Integration Improves Predictive Accuracy

We evaluated our genome-wide functional network predictions using gold standards based on the Gene Ontology [8], Plant Ontology [45], and Pfam A [9]. This let us determine how accurate each network was in assigning high probability to known functional interactions (i.e. gene pairs co-annotated in GO, PO, etc.) As seen in Figure 4, both the GLOBAL-PROCESS and GLOBAL-DEVEL networks were particularly accurate in the low recall, high precision area of greatest biological interest.

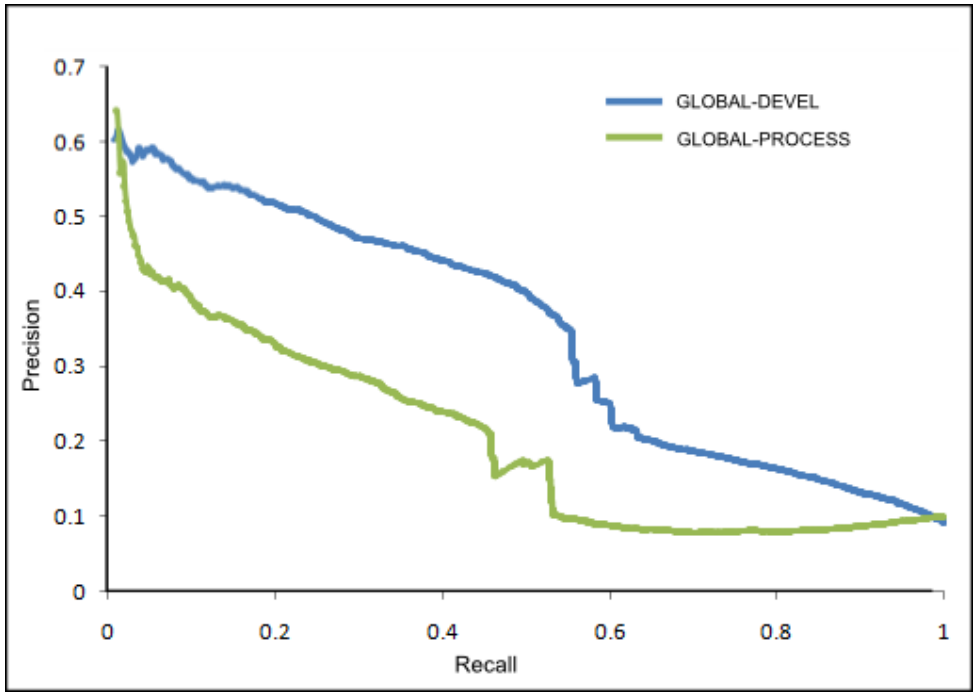

Figure 4. Performance of the GLOBAL-PROCESS and GLOBAL-DEVEL networks.
The two global networks were evaluated using 5-fold cross-validation with a 20% holdout gene set to test their ability to accurately recover functional and developmental-stage-specific protein interactions. The higher precision of the GLOBAL-DEVEL network suggests that co-functionality during developmental stages can be more accurately inferred from high-throughput data than can more general functional relationships, although both networks are predicted with significant accuracy.

Additionally, GLOBAL-DEVEL slightly outperforms GLOBAL-PROCESS, suggesting that gene pairs co-active during the same developmental stages are easier to predict from integrated genomic data than are gene pairs participating in the same biological processes. This is supported intuitively by the fact that developmental expression programs are, in many cases, more sharply defined than are biological

pathways and processes, and quantitatively by the fact that several of the integrated datasets explicitly incorporate developmental-stage-specific experiments. We further found that the context-specific networks usually performed better than the global networks (Figure 5).

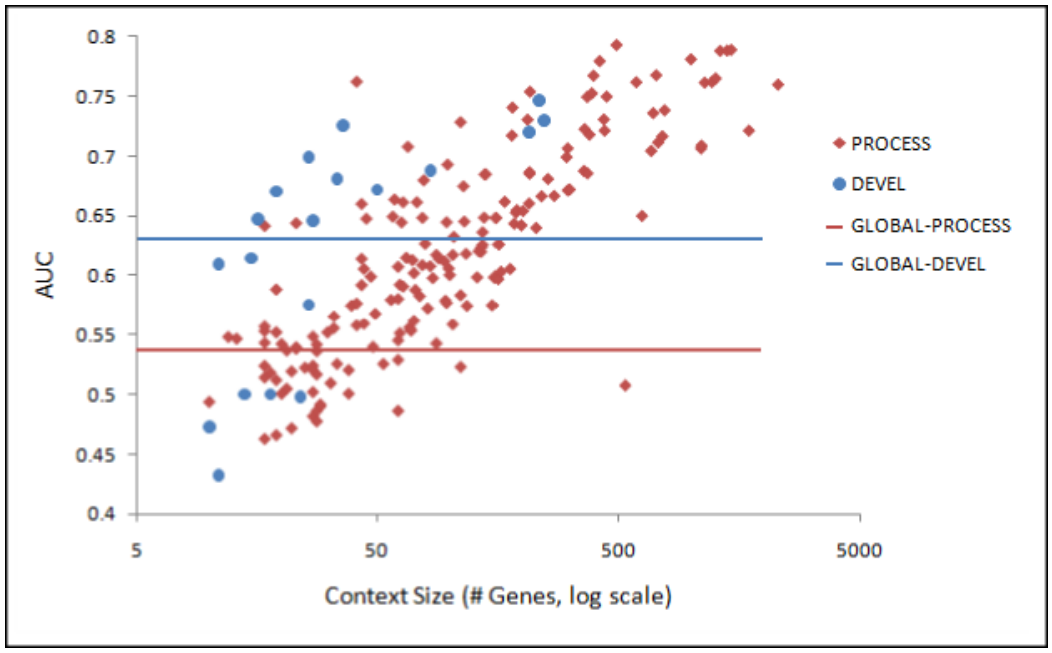

Figure 5. Context-specific functional networks are often more accurate than global networks.
AUC values for 208 biological process contexts (PROCESS networks) and 19 development contexts (DEVEL networks). The lines indicate the GLOBAL-PROCESS and the GLOBAL-DEVEL networks' performance.

As the network generation process is data-driven, the accuracy of each integration depends on (a) whether relevant biological signals are present in the data and (b) the availability of a sufficiently comprehensive gold standard. Contexts with very limited prior knowledge or a small number of genes annotated to them sometimes perform

marginally. We determine the performance using an AUC (area under the receiver-operator curve) value, which measures the probability that our classifier ranks a functional relationship better than a random classifier. For example, the floral organ development stage context with 34 genes has an AUC of 0.51. Overall more than half (55%) of developmental-stage specific integrations had AUCs over 0.63, that of the GLOBAL-DEVEL network. Many (74%) of the biological process specific integrations had AUCs over 0.54, that of the GLOBAL-PROCESS network. In addition to providing increased predictive power, these context-specific networks focus a very large collection of *A. thaliana* genomic data into individual areas of interest, enabling rapid and directed biological hypothesis generation.

Table 2 details the combinations of developmental stages and tissues/biological processes in the TISSUE-DEVEL and PROCESS-DEVEL compendia for which adequate gold standards were available for evaluation. Networks in plant structures such as embryo and carpel were generally predicted with higher accuracy than those in structures such as leaf and root. AUCs were particularly high in all development contexts and the leaf tissue and were particularly low in all tissues/biological processes for the germination development stage.

Table 2. Development stages and tissues/biological processes of interest
These nine tissue/process contexts had sufficient overlapping curated information to evaluate our accuracy in predicting functional relationships occurring during a specific developmental stage within one tissue. For example, the meristem activates gene programs to differentiate into shoot and root tissues during the D bilateral stage [20], and we accurately recover these predicted interactions.

DEVELOPMENT STAGE	TISSUE/BIOLOGICAL PROCESS	AUC	LEVEL
C globular stage	meristem leaf seed	0.822 0.818 0.754	Strong interaction with development
D bilateral stage embryo dev stages flower dev stages	Meristem	0.816 0.8 0.79	
0 germination flora organ dev stages flower dev stages	Carpel	0.66 0.73 0.71	Weak interaction with development

The globular stage and meristem combination network has the highest AUC in the TISSUE-DEVEL compendium, and the globular stage is indeed when primary meristems produce new cells that will ultimately differentiate and patterning of the shoot and root apical meristems begins [60]. The "globular stage" also has a high AUC with other tissues (leaf, root, and seed) and biological processes (the "organismal physiological process", the "reproductive physiological process", and "transcription"), suggesting that meristem activity in these tissues is prominent and significant. Other predictions for the meristem [61] are also informative: in the "bilateral stage", the meristems become distinguished as shoot and root meristems; in the "embryo development stages", the embryo develops radial patterning and primary shoot meristems are formed; and in the "flower development stage", floral meristem

genes help the transition from shoot to floral meristem [62]. All of these TISSUE-DEVEL networks achieve high AUCs. In contrast, a specialized tissue like the carpel has both low and high predictive powers across development stages. Since the stigma, not carpel, is the receptive tissue where germination happens [63], accuracy is low in the germination development stage but higher in the flower development stage and floral organ development stages.

2.2.3 Bayesian Integration Highlights Experimental Datasets Informative in Specific Biological Contexts of Interest

We summarize the "weight" given to each dataset during Bayesian integration by calculating its overall influence on the posterior probability of functional relationship. This provides a measure of how informative each dataset is within each context of interest (Figure 6).

Figure 6. Weights automatically determined for each dataset contributing to predictions in each context.

Weights are calculated as the influence of each dataset on the posterior probability in the process or development network's Bayesian classifier, where a higher number indicates a greater influence.

Highly specific datasets such as physical interactions tend to be informative in many process and developmental contexts. The GLOBAL-PROCESS network, which is the most diffuse and difficult to predict, is not strongly influenced by most datasets and focuses on those that are particularly large and/or diverse. The GLOBAL-DEVEL network, unsurprisingly, is highly influenced by expression datasets incorporating developmental-stage-specific exposures (e.g. hormone treatments and the *A. thaliana* expression atlas [58]). The heterogeneity of dataset contributions increases as context size shrinks, until the smallest contexts are heavily influenced by particularly relevant data (e.g. chemical treatments of seedlings is highly informative in the dry seed stage).

2.2.4 Regularization of Bayesian Network Parameters Using Dataset Mutual Information Efficiently Increases Prediction Accuracy

Naïve Bayesian models assume independence between all input datasets, which can artificially inflate predicted probabilities when this assumption is violated (e.g. when multiple very similar datasets are integrated). Conversely, a full Bayesian model accounting for naturally-occurring dependencies (similar experimental conditions, platform and lab effects, etc.) would be inefficient to learn and evaluate using dozens of whole-genome datasets. Our solution to this issue was to regularize the Bayesian learning process using mutual information between datasets as a prior to up-weight or down-weight the total possible contribution of each dataset. This mixes a uniform prior with each dataset's predictions, weighted relative to the amount of information it shares with other datasets, and does so as a preprocessing stage without diminishing

the efficiency of naive Bayesian learning and inference. Figure 7 shows normalized pairwise mutual information scores between all datasets.

As expected, physical interaction datasets (labeled in Figure 7 on the vertical axis by "int pfam", "int myristoilation", "int bind", "int ppi", "int biogrid") cluster together and are quite different from the main body of microarray expression data. Microarray data falls into several large classes: abiotic stresses, biotic stresses, chemical treatments, hormone treatments, and physical protein-protein interactions. "Abiotic treatments" are the most similar (and thus down-weighted), since they evoke strong transcriptional responses that are easy to detect during the integration process [64] [65] [66]. Similarly, other abiotic treatments – "different temperature treatments of seeds" and "hormone treatment – basic hormone treatment of seeds" are similar and share more data than most dataset pairs. These datasets are unique in that they stress *A. thaliana* seeds as opposed to seedlings, and their up-weighting (Figure 7) may indicate that the response to these stresses is easier to detect in seeds than in other experimental conditions.

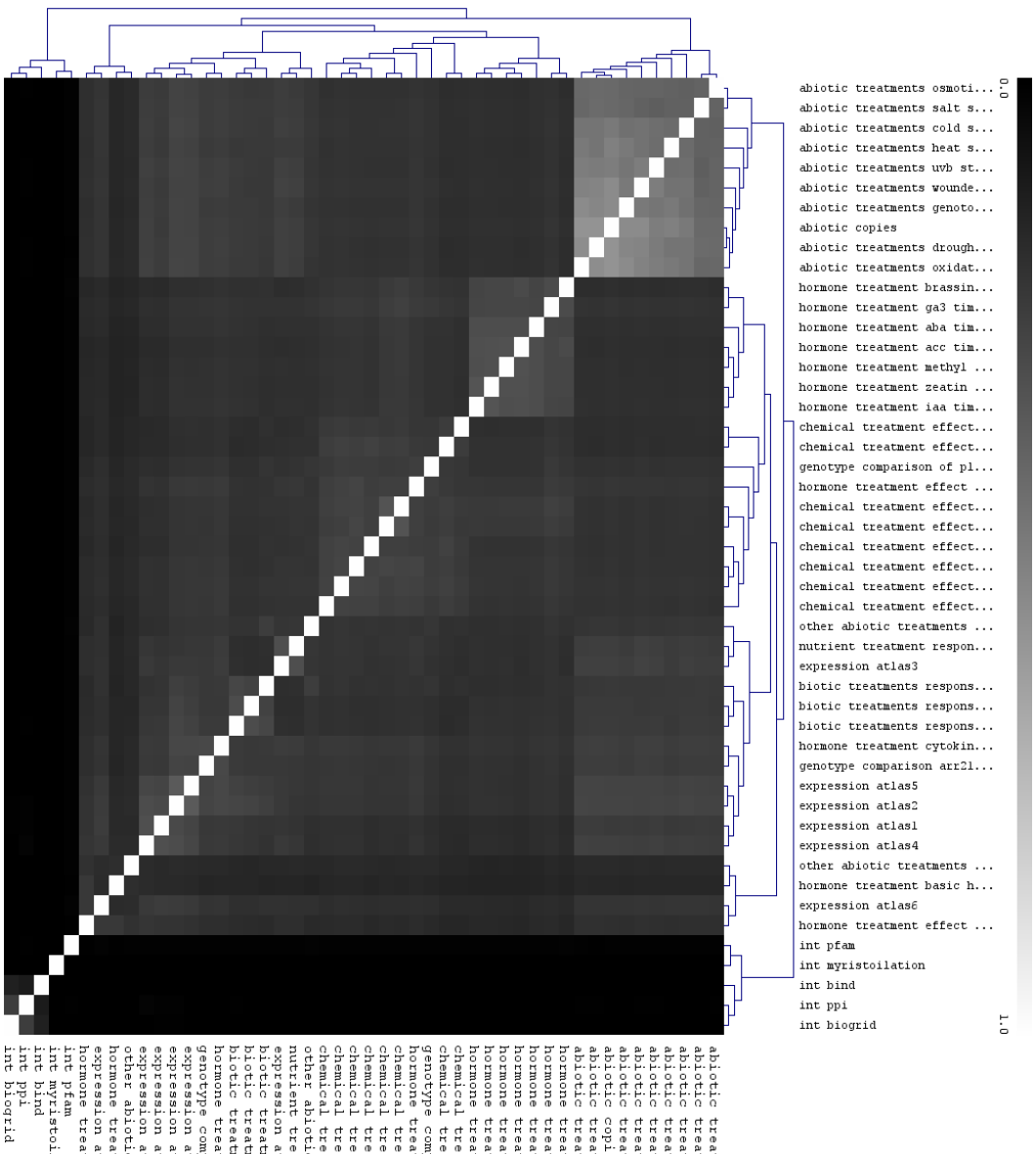
abiotic treatments osmoti...
abiotic treatments salt s...
abiotic treatments cold s...
abiotic treatments heat s...
abiotic treatments uvb st...
abiotic treatments wounde...
abiotic treatments genoto...
abiotic copies
abiotic treatments drough...
abiotic treatments oxidat...
hormone treatment brassin...
hormone treatment ga3 tim...
hormone treatment aba tim...
hormone treatment acc tim...
hormone treatment methyl ...
hormone treatment zeatin ...
hormone treatment iaa tim...
chemical treatment effect...
chemical treatment effect...
genotype comparison of pl...
hormone treatment effect ...
chemical treatment effect...
chemical treatment effect...
chemical treatment effect...
chemical treatment effect...
chemical treatment effect...
chemical treatment effect...
other abiotic treatments ...
nutrient treatment respon...
expression atlas3
biotic treatments respons...
biotic treatments respons...
biotic treatments respons...
hormone treatment cytokin...
genotype comparison arr2l...
expression atlas5
expression atlas2
expression atlas1
expression atlas4
other abiotic treatments ...
hormone treatment basic h...
expression atlas6
hormone treatment effect ...
int pfam
int myristoilation
int bind
int ppi
int biogrid
0.0
1.0
abiotic treatments osmotic stress time coursel
abiotic treatments salt stress time coursel
abiotic treatments cold stress time coursel
abiotic treatments heat stress time coursel
abiotic treatments uvb stress time coursel
abiotic treatments wounded stress time coursel
abiotic treatments genotoxic stress time coursel
abiotic copies
abiotic treatments drought stress time coursel
abiotic treatments oxidative stress time coursel
hormone treatment brassinolide time course in wildtype seedlings
hormone treatment ga3 time course in wildtype seedlings
hormone treatment aba time course in wildtype seedlings
hormone treatment acc time course in wildtype seedlings
hormone treatment methyl jasmonate time course in wildtype seedlings
hormone treatment zeatin time course in wildtype seedlings
hormone treatment iaa time course in wildtype seedlings
chemical treatment effect of cycloheximide inhibitors on seedlings
chemical treatment effect of proteasome inhibitors on seedlings
genotype comparison of plant hormone related mutants
hormone treatment effect of brassinosteroids on seedlings seedlings
chemical treatment effect of photosystem inhibitors on seedlings
chemical treatment effect of brassinosteroid inhibitors on seedlings
chemical treatment effect of 1 sa d on seedlings
chemical treatment effect of gibberellic acid inhibitors on seedlings
chemical treatment effect of auxin inhibitors on seedlings
chemical treatment effect of ethylene inhibitors on seedlings
other abiotic treatments light treatments
nutrient treatment response to sulfate limitation
expression atlas3
biotic treatments response to secretion system deficient bacteria
biotic treatments response to phytophthora infestans
biotic treatments response to bacterial comycete derived elicitors
hormone treatment cytokinin hormone treatment treatment of seedlings
genotype comparison arr1c overexpression
expression atlas5
expression atlas2
expression atlas1
expression atlas4
other abiotic treatments different temperature treatments of seeds
hormone treatment basic hormone treatment of seeds
expression atlas6
hormone treatment effect of aba during seed imbition
int pfam
int myristoilation
int bind
int ppi
int biogrid

Figure 7. Normalized pairwise mutual information scores between all datasets. To regularize the Bayesian classifiers used in this study, we calculated the mutual information between each pair of datasets. These values were normalized as fractions of the total possible shared information and used to exponentially down-weight datasets containing a large fraction of redundant information. The raw mutual information values are shown here and serve to group datasets that are related for technical (e.g. similar microarray platform) or biological (e.g. similar experimental treatment) reasons.

2.2.5 Development-Specific Networks Enable Biological Hypothesis Generation

As an example of biological hypothesis generation using the DEVEL networks, we investigated the most confident interactions predicted for a specific protein, *AtPPT2* (*AT3G01550*) within two development stages. *AtPPT2* encodes a PHOSPHOENOLPYRUVATE (PEP)/PHOSPHATE TRANSLOCATOR (PPT) [67] that mediates cytosol-plastid PEP transport [68]. It is highly associated with several genes in the "leaf development stage", but it lacks the same activity in the "root development stage". Given this difference, we investigated its top 5 predicted interaction partners in each tissue context. We also investigated the contribution that various datasets have on the predicted interactions. For *AtPPT2,* for example, we found that datasets containing experiments done on the root contributed over 2 times more information (based on posterior probability, Figure 8) in root development than the same experiments done on the shoots.

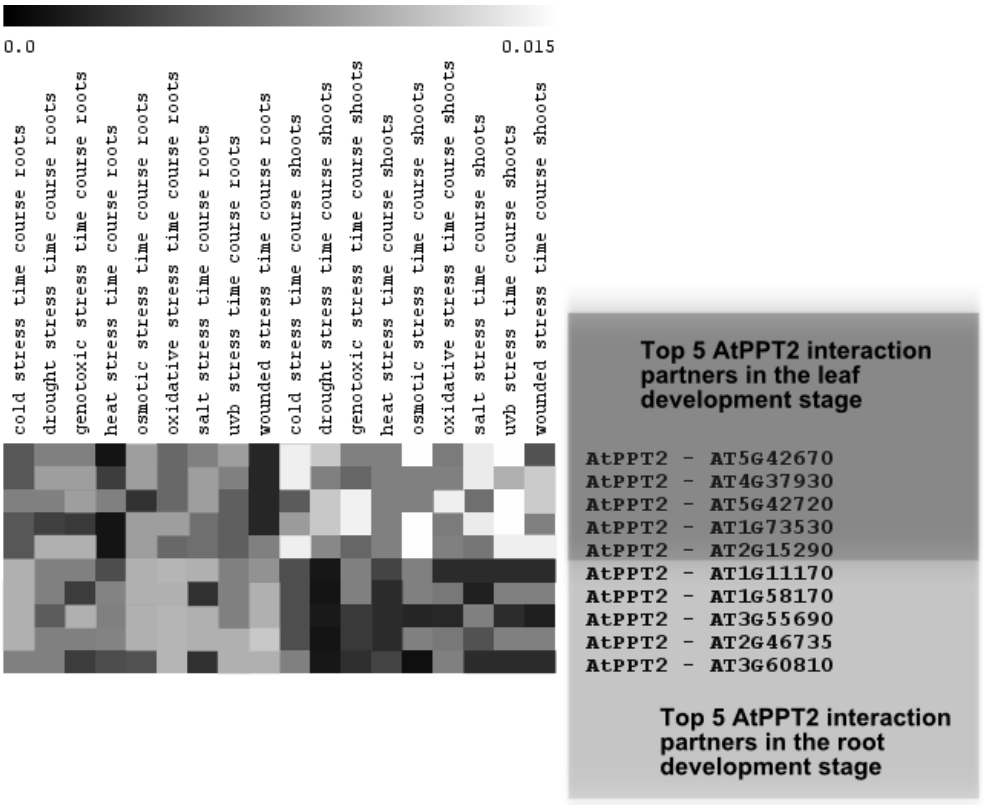

Figure 8. Information contributed by root and shoot experiments in the leaf and root development contexts.
Predicted interaction partners for *AtPPT2* in the leaf and root development stages. In the former case, experiments in shoots are approximately twice as informative as those in roots; the reverse is true in the latter case. This suggests that our network inference process can correctly learn which datasets are most informative in specific contexts.

The opposite effect was observed in the leaf context, with experiments on roots down-weighted and leaf experiments up-weighted. For both root and leaf development, the protein-protein interaction datasets did not have much influence at all compared to the microarray datasets on any of the pairs.

An interesting case study is the predicted functional relationship between genes *AT4G37930* and *AtPPT2* in the "leaf development stage", which is most influenced by the following datasets: a) a study of drought stress in shoots [58], b) salt stress in shoots [58], c) UVB stress in shoots [58], d) osmotic stress in shoots [58], and e) cold stress in shoots [58]. A clear hypothesis implied by this prediction is thus that *AT4G37930* and *AtPPT2* both play a role in the cellular response to stress in shoots. Additional experiments not included in our input data [68] show that *AtPPT2* is highly expressed only in "leaf development stages: and not in the "root development stages".

2.2.6 Predicted Interactions in Several Networks are Literature-Validated

RPM1 INTERACTING PROTEIN 4 (RIN4), RESISTANCE TO PSEUDOMONAS SYRINGAE pv. MACULICOLA 1 (RPM1) and RESISTANCE TO PSEUDOMONAS SYRINGAE 2 (RPS2) were predicted to be co-active in the GLOBAL-PROCESS network and in the "vegetative growth stages". RIN4 has been shown to physically interact with RPM1 and RPS2, and the three proteins are part of the plant's defense response to the bacterium *P. syringae* [69] [70]. In the vegetative stage, RIN4 is also predicted to be co-active with NDR1, which physically interacts with RIN4 *in vivo* [71]. Further, in the GLOBAL-DEVEL network, RIN4 is predicted to be co-active with NPR1-like protein 4 (NPR4). Mutations in NPR4 result in susceptibility to *P. syringae*, and although NPR4 has not previously been shown to associate with RIN4, our predicted network suggests these proteins may interact.

Our GLOBAL-DEVEL network predicts an interaction between the root hair patterning regulator WEREWOLF (WER) and additional proteins in the root hair development pathway, including CAPRICE (CPC), GLABRA3 (GL3), and ENHANCER OF GLABRA3 (EGL3). In addition, this network predicts that GL3 and EGL3 interact, and that CPC is interacts with EGL3 and GL3. WER is known to regulate expression of CPC [72], and both WER and CPC regulate expression of EGL3 and GL3 [73]. Further, GL3 and EGL3 physically interact [74]. We also found that the transcription factors (TFs) MAGPIE (MGP), NUTCRACKER (NUC) and JACKDAW (JKD) are co-active in the "seedling growth stage", while MGP and NUC are co-active in the "root development stages". These three proteins are part of a network involved in ground tissue patterning in the root [75] [76]. MGP and NUC are downstream direct targets of the ground tissue patterning regulator SHORTROOT (SHR) [75]. JKD and MGP physically interact both with each other and with SHR and another key ground tissue patterning transcription factor (TF), SCARECROW (SCR) [76]. *MGP* transcription depends on SHR and SCR, while *JKD* transcription in embryogenesis is independent of SHR and SCR, but becomes dependent on these TFs at later stages [33]. Though *mgp* mutants do not have a phenotype, *jkd* mutants show a small reduction in root length compared to wild type plants. Additionally, reducing *MGP* expression in the *jkd* mutant showed that these proteins have opposing effects on SHR and SCR in the ground tissue [33].

A third predicted network involves the plant hormone auxin. *TRANSPORT INHIBITOR RESPONSE 1* (*TIR1*), encodes an auxin receptor that regulates auxin-

mediated transcription [77] [78]. TIR1 has been shown to interact with ASK1, ASK2, AtCUL1, and AUX/IAA proteins [79] [80], all of which are predicted to be co-active in the GLOBAL-DEVEL network. Our network further predicts that TIR1 interacts with proteins not known to associate with the receptor, such as AT3G23640, a heteroglycan glucosidase involved in carbohydrate metabolism, and AT2G36720, an uncharacterized transcription factor, suggesting that these proteins may be involved in auxin related processes.

Together, these results show that our networks can accurately predict interactions in different plant developmental stages in a wide array of physiological processes.

2.3 Conclusions

Here, we present an ensemble of genome-wide functional relationship networks predicted for *A. thaliana* using Bayesian integration of 60 experimental datasets. We infer six classes of networks: one GLOBAL-PROCESS network predicting genes participating in related biological roles; one GLOBAL-DEVEL network predicting genes co-active in the same developmental stage(s); a compendium of PROCESS networks, each containing relationships specific to one biological process or pathway; a compendium of DEVEL networks, each predicting co-activity within an individual developmental stage; and the PROCESS-DEVEL and TISSUE-DEVEL compendia calling out processes and tissue-specific activity occurring during individual developmental stages. Each network reweights the genomic data compendium to yield

predictions tailored to an individual biological context of interest. The leaf- and root-specific networks predicted that the *AtPPT2* protein functions during leaf development but not root development, which has since been confirmed experimentally [68]. We further identified several literature-validated interactions among our predicted interactions.

We anticipate that these context-specific predictions of *A. thaliana* functional relationships will be useful to drive future hypotheses generation regarding protein function and interactions as they change among *A. thaliana* tissues and developmental stages. With these networks, biologists can pose extremely specific questions regarding individual genes' interactions within isolated plant tissues and at only one (or more) time(s) during development, allowing them to discover novel gene functions more rapidly. A web interface to our predictions, available at http://function.princeton.edu/arathGraphle, provides these networks in a convenient interface accessible to the wider biological and bioinformatics communities.

3 INTERACTIONS BETWEEN PROTEINS AND SMALL MOLECULES

Drug discovery and development are cornerstones of biomedical research, and enormous efforts have been channeled towards rational drug design and high-throughput drug screening. Yet, there is a dearth of novel, specific, single-target drugs [23]. Compounding this problem are the facts that most diseases are complex multi-gene/multi-process dysfunctions and most drugs participate in 'off-target' interactions [81] [82]. Drug discovery research in the past few years is therefore increasingly adopting functional genomics – observing, modeling and analyzing genome-wide gene/protein read-outs – to grasp systems-level gene deregulation in disease and understand potential drug action [23] [83] [84] [85]. A culmination of these recent efforts is *network pharmacology* – an approach combining network biology with drug discovery to tackle complex interactions between genes/proteins and many-to-many drug-target associations [86]. While such studies have been attempted in human, several biological and practical difficulties still loom large. The human genome is large with nearly 25,000 genes, many of which are functionally redundant, diverse and uncharacterized [87] [88]. The human organism is remarkably multicellular, composed of more than 200 cell-types. Also, the amount of human functional-genomics data, although abundant, does not yet match-up to the scale of the biological complexity.

We address the problem of drug discovery in humans by tackling drug-protein interactions in a simpler organism first, namely *Saccharomyces cerevisiae*. *S. cerevisiae* is the most commonly studied type of yeast, with about 6,000 genes. It is single celled and has a simple life cycle, doubling in less than two hours. Its cells can be manipulated outside their natural environment and grown under controlled conditions; as such, it is easily cultured and there is much experimental data available for it [89]. For example, an experimental undertaking replaced every yeast gene with a drug resistance marker to show that about 1000 of the ~6000 yeast genes are vital for yeast survival [90]. This task could not be so readily done with human genes. In a separate study, an experimental technique called haploinsufficiency profiling (HIP) was applied by decreasing the dosage of a drug-target gene from two copies to one copy and observing changes in drug sensitivity for one gene [91] or an entire collection [92]. Such systematic evaluations of drug targets in yeast are useful and can be translated to humans.

While yeast is a simple organism, its cell structure is complex enough to be comparable to plants and animals. Yeast and humans share similar mitochondria (cell structures that generate the ATP used as chemical energy and are involved in other tasks such as cell signaling, cell growth, and cell death) and several studies have shown that modeling diseases in yeast can translate well to the humans [93] [94] [95] [96]. Other basic mechanisms such as transcriptional regulation, trafficking, and proteasomal function is well conserved between yeast and humans, enabling studies in and modeling of neurodegenerative diseases in yeast [97]. Ultimately, yeast is not a

perfect model for human disease because it lacks certain features that multi-cellular organisms have, such as immune systems, organs, or tissues. However, conservation of basic cellular processes and metabolic pathways make it a great tool to understand basic mechanisms behind drug-gene interactions and drug discovery [98].

In the previous chapter, we presented functional networks that help biologists answer specific biological questions. Our networks were able to accurately predict functional gene interaction in various biological processes, tissues, and developmental stages. We now extend these functional networks to networks that accurately predict interactions for specific interaction types. With these functional networks, in this chapter we switch gears and engage in another computational biology problem: can these networks provide insight into the mechanism of chemical compound and drug interactions?

In this chapter, we address this question by broadening our interaction space to not only include drugs, but also other chemical compounds such as ethanols or sugars. There is much more information on these general compounds; including them in our interaction predictions will provide better insight into the groups of compounds that drugs belong to when interacting with proteins. We use machine learning techniques to leverage a bounty of heterogeneous data by integrating it with pathway-level protein-protein interaction networks in a chemical compound setting. This integration yields a compendium of compound-protein interactions, some compounds being drugs, that assigns interaction probabilities to a compound-protein pair based upon the

integrated data. Our compendium is made up of 702 chemical compounds and 5559 yeast proteins. The majority of these compounds are small molecules and chemicals but 13 are FDA approved drugs. This compendium will be a valuable tool for biologists to narrow their experimentation search space and, ultimately, to future drug discovery.

3.1 Methods

We present a machine-learning framework for predicting interaction between proteins and small molecules (chemical compounds and drugs) in *Saccharomyces cerevisiae* (yeast). The framework relies on a two-step data integration process that incorporates mechanistic protein interactions to learn compound-protein interaction predictions.

1. We first create "mechanistic gene-gene interaction networks" by leveraging microarray expression, protein domain, protein family, and protein structural data in various interaction-type contexts using a hierarchically-corrected integration of support vector machines, similar to a previous technique shown to perform well [21]. Pathway-level information from 30 different interaction types (for example, "regulatory interaction", "phosphorylation", "synthetic rescue") yields 30 different protein-protein interaction networks.

2. These 30 interaction-type networks are consolidated into one large protein-protein interaction network. Based on a gold standard of known compound-protein interactions, we apply a support vector machine to predict novel interactions between compounds and proteins. We present a “compound-protein network” with interaction probabilities for every compound-protein pair.

These steps are visualized in Figure 9 and we will describe them in greater detail in the sections that follow.

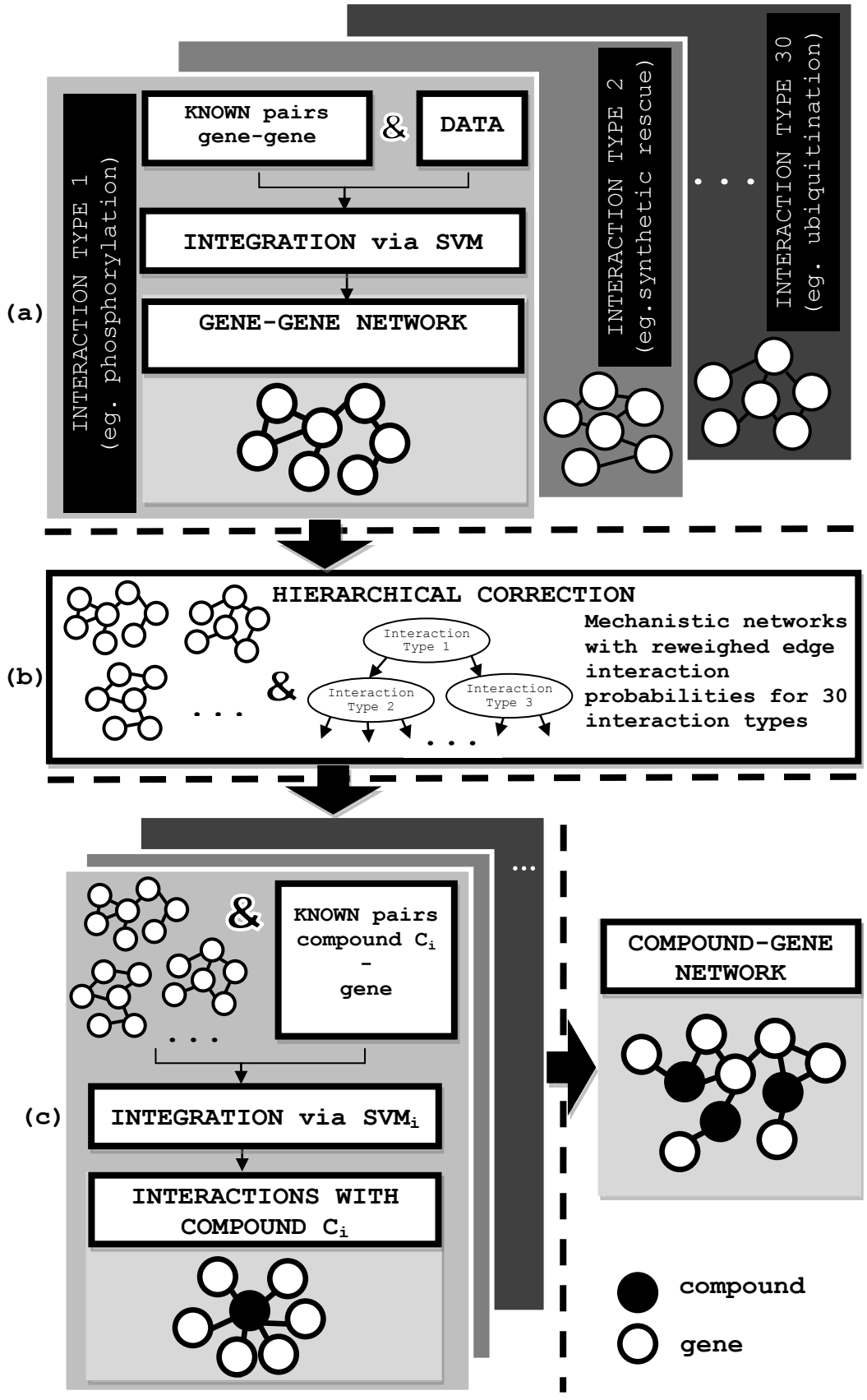
(a)
INTERACTION TYPE 1
(eg. phosphorylation)
KNOWN pairs
gene-gene
&
DATA
INTEGRATION via SVM
GENE-GENE NETWORK
INTERACTION TYPE 2
(eg.synthetic rescue)
. . .
INTERACTION TYPE 30
(eg. ubiquitination)
(b)
HIERARCHICAL CORRECTION
Interaction Type 1
Interaction Type 2
Interaction Type 3
Mechanistic networks with reweighed edge interaction probabilities for 30 interaction types
(c)
KNOWN pairs
compound C_i
-
gene
INTEGRATION via SVM_i
INTERACTIONS WITH
COMPOUND C_i
COMPOUND-GENE
NETWORK
compound
gene

Figure 9. Meta integration pipeline
Two step integration process. **Step 1**: In (a), we first gather large-scale data features (microarray experiments, protein sequence/structural information, etc.). Then, for each interaction type, we curate a gold standard of known gene-gene pairs, and use a support vector machine to leverages the features and the gold standard to generate a predicted genome-wide gene-gene interaction network. In (b) we use the hierarchical structure of interaction types to modify our previous interaction-type networks and generating mechanistic networks. **Step 2**: In (c) we consider each compound of interest separately. We use the mechanistic networks from Step 1 as new data features to an SVM; then, for each compound, based on a gold standard of known gene interactions with that compound, we predict interaction probabilities between that compound and all genes. This procedure results in a probabilistic interaction network between all compounds and genes.

3.1.1 Support vector machines

The support vector machine (SVM) was first introduced in 1998 [99] and has since been used in different settings with success in the areas of text categorization, face recognition, speech patterns, and as in our case, computational biology. Classification using SVMs aims to separate a set of two classes, positive and negative examples, by drawing a plane in such a way that the distance between the plane and some support vectors within a margin is maximal. The process of finding an optimal hyperplane separating the two classes (Figure 10) involves finding a line $\mathbf{w} \cdot \mathbf{x} - b = 0$ where $\mathbf{w}$ is a vector perpendicular to the plane and $\mathbf{x}$ is the vector of examples. The margin boundaries (the distance between the optimal hyperplane and the closest examples) are given by the equations $\mathbf{w} \cdot \mathbf{x} - b = 1$ and $\cdot \mathbf{x} - b = -1$. To determine the optimal hyperplane, we maximize the distance between the margins $2/||\mathbf{w}||$. In this thesis, we used a linear SVM [100].

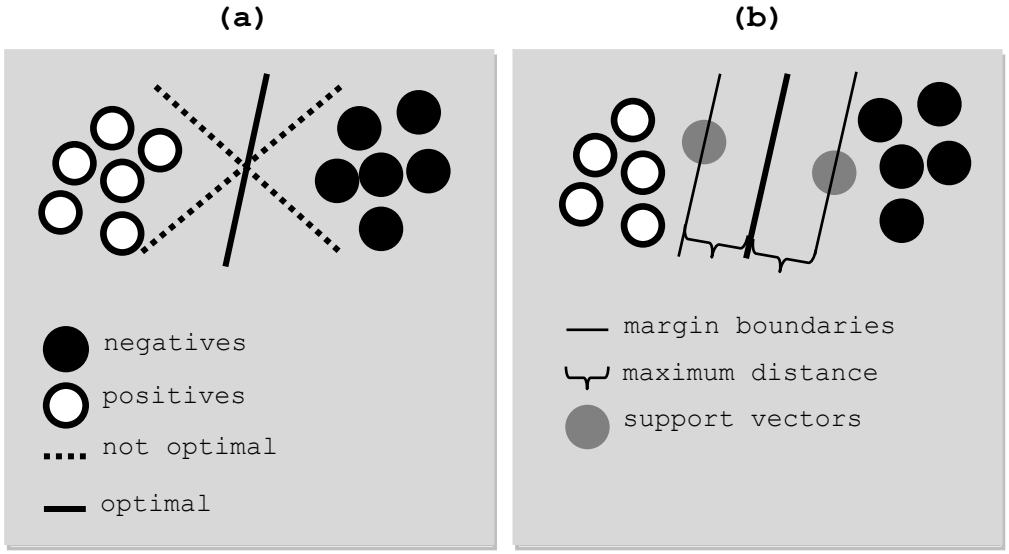

Figure 10. Support vector machine.
(a) An infinite number of planes can be drawn to separate the set of positive examples from the set of negative examples, but only one separates them in an optimal way. (b) The optimal plane is found by maximizing the distance between one or more support vectors on either side of the plane.

In the next sections, we will explain how we adapted the general SVM for our data.

3.1.2 Interaction-Type Functional Networks

In this section, we will describe our methods to arrive at interaction-type functional networks (Figure 9 (a)). Using multiple mechanistic networks for drug-protein prediction has not been in the drug discovery field, but we believe the pathway-level biological information inherent in our 30 interaction-type networks will provide useful insights into how compounds and proteins interact.

3.1.2.1 Gold Standards

We first obtained the gold standard as generated by [21]. This gold standard was assembled using a mix of data mining from relevant databases and manual curation for more specific interaction types. Positive examples were mined from the Saccharomyces Genome Database (SGD) [101] by matching their interaction labels to our interaction types, the Kyoto Encyclopedia of Genes and Genomes (KEGG) [102], and GO [8]. Negative examples were randomly chosen since protein interactions are sparse. Table 3 shows the 30 interaction types that were considered in this study.

Table 3. 30 interaction types
The table and descriptions are from the supplement of [21].

INTERACTION TYPES	DESCRIPTION
Complex	Any macromolecular complex composed of two or more protein subunits.
Covalent modification	Protein regulation by transfer or removal of a molecule or atom from a donor to an amino acid side chain that serves as the acceptor of the transferred molecule or (as in regulating an enzyme) by altering the amino acid sequence itself by proteolytic cleavage.
Functional group transfer	Transfer or removal of a functional group.
Functional relationship	Two proteins function in same biological process.
Interaction pathway	Two genes are associated at the pathway level, including post-transcriptional, transcriptional and post-translational regulation or the functional dependency such as synthetic interactions.
Isoenzyme	Enzymes that differ in amino acid sequence but catalyze the same chemical reaction.
Mediated by small molecule	Two proteins where a small molecule is involved as part of a protein modification.
Metabolic interaction	Functionally associated at the metabolic level.
Non covalent binding	Two proteins interact in a non-covalent nature.
Peptide transfer	Transfer peptide to protein.
Phenotypic aggravation	Mutation or overexpression of one gene results in suppression of any phenotype (other than lethality/growth defect) associated with mutation or overexpression of another gene.

Phenotypic alleviation	Mutation or overexpression of one gene results in enhancement of any phenotype (other than lethality/growth defect) associated with mutation or overexpression of another gene.
Phenotypic interaction	Mutation or overexpression of one gene results in alteration of any phenotype (other than lethality/growth defect) associated with mutation or overexpression of another gene.
Phosphate transfer	Addition/removal of a phosphate (PO4) group to/off a protein.
Phosphorylation	Addition of a phosphate (PO4) group to a protein.
Physical interaction	Two proteins physically interact.
Posttranscriptional regulation	Post-transcriptional regulation is the control of gene expression at the RNA level.
Posttranslational regulation	Post-translational regulation refers to the control of the levels of active proteins.
Regulatory interaction	A gene regulates a gene either at the RNA, protein or transcription level.
Same enzyme class	Two enzymes that share the same enzyme class.
Shared pathway	Two proteins are closely involved in a pathway
Synthetic aggravation	Mutation or deletion of one gene aggravates the effect of a strain mutated/deleted for another gene.
Synthetic alleviation	Mutation or deletion of one gene alleviates the effect of a strain mutated/deleted for another gene.
Synthetic growth defect	Interaction is inferred when mutations in separate genes, each of which alone causes a minimal phenotype, result in a significant growth defect under a given condition when combined in the same cell.
Synthetic interaction	Interaction in which a combination of mutations in two or more genes of a single strain results in a phenotype that is different in degree or nature from the phenotypes conferred by the individual mutations.
Synthetic lethal	Mutations or deletions in separate genes, each of which alone causes a minimal phenotype, result in lethality when combined in the same cell under a given condition.
Synthetic rescue	Mutation or deletion of one gene rescues the lethality or growth defect of a strain mutated/deleted for another gene.
Transcriptional regulation	Transcriptional regulation is the change in gene expression levels by altering transcription rates.
Ubiquitination	The post-translational modification of a protein by the covalent attachment of one or more ubiquitin monomers.
Ubiquitin transfer	The post-translational modification of a protein by the covalent attachment or removal of one or more ubiquitin monomers.

3.1.2.2 Experimental Data

Our data sources consisted of 3523 microarray experiments, protein domains, sequence similarity, co-localization, and transcription factor binding sites, as in [21].

While this data is diverse, there is a crucial missing piece especially in the context of drug-protein binding: protein structure information. A protein's secondary and tertiary structural conformation encodes potential binding properties and may change depending on the biological function that protein performs.

The two structural data-types we use are: binding site conservation and docking. Binding site conservation (BSC) was available from [103]. The study showed that binding sites are conserved among close homologs (proteins derived from a common "ancestor"), as expected, as well more remote structural neighbors whose evolutionary relationship is not as well defined. Their method, for example, is able to correctly determine pairs of proteins that are structurally related in three cases: when an original protein complex is related to another (a) by some mathematical translation/rotation, (b) by an overlap, or (c) by some local alignment of sequences. Briefly, they identify structural neighbors for a given query protein and the locations of interfacial residues of the neighbors that are part of a complex are "mapped" to residues in the query protein to generate a "contact map" associated with each structural neighbor. Interface conservation areas can be identified by summing individual contact maps and generating a contact frequency heat map. We used their z-scores, representing contact map overlap for structural neighbors of a protein, as experimental data.
Protein docking is the process of computationally modeling and determining the most likely conformation that two proteins will interact. Docking algorithms take two protein structures as input and use a mixture of computational algorithms and heuristics to execute rotations and transformations in a 3D space to determine the

minimum free energy required to force the pairs of structures in a specific conformation. We should make it clear that the results from docking algorithms are, in and of themselves, potential bindings that can occur between pairs of proteins based on the hypothesis that proteins orient themselves to minimize the free energy of the resulting complex [104]. Protein docking is a computationally intensive task requiring several orientations and conformational changes need to be considered for each pair of proteins. We used the HEX docking algorithm [105] because it is fast and has a command-line interface that enables automation of many docking jobs on a cluster.

As input to HEX, we retrieved the 3D structure of our proteins; out of 5559 proteins, only 1825 had a known structure in the Protein Database (PDB) in March 2011 [106] [107]. We ran the HEX algorithm on all the possible combinations of these proteins with known structure, about 1.7 million pairs, which took about 6 months on a cluster of 58 CPU cores. We used the resulting docking score from HEX to rank all our protein pairs.

3.1.2.3 Support Vector Machine Classification

With an arsenal of experimental data in hand, based on the type-specific gold-standards, we employed a separate SVM classifier to construct a fully-connected network for each of the 30 different interaction-types. We constructed the SVM input feature vectors for each protein pair as described in [21], where each feature is the score for a pair of proteins in one of the input datasets. Scores from different data-

types were calculated as follows: for microarray expression values, we subtracted expression values between two proteins in an experimental condition; for sequence similarity, we used the e-value from BLAST outputs; for other types of data, pairwise protein scores were already inherent in the data. We used a linear kernel SVM available in the Sleipnir library [108].

3.1.2.4 Feature Selection

Our microarray compendium consisted of 3523 different microarray conditions, contributing to a large percentage of our input data. In order to both manage input size and reduce input redundancy, we used feature selection on the large number of microarray datasets. Feature selection has been shown to improve performance [109] [110].We used the *gist-fselect* auxiliary program available under the software package Gist version 2.3 [111]. The inputs to a feature selection algorithm are evaluated using a quality metric and low-quality features are removed. We tried several metrics implemented in this tool, and determined that Welch's approximate t-test was the best for our data because our samples may not have equal variances. The t-statistic between two datasets is calculated by

$$t = \frac{|X_1 - X_2|}{\sqrt{\frac{s_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

where X_k is the sample mean for the kth dataset, s_k^2 is the variance for the kth dataset and N_k is the size of the kth dataset. Each feature selection cutoff is applied to the 30

interaction types to generate 30 networks. These 30 datapoints are visualized in a boxplot and we compare the feature selection cutoffs in a pairwise fashion using the paired Wilcoxon rank sum test (Figure 11). The Wilcoxon signed rank test between the feature selection cutoffs x and y tests that the null distribution of $x - y$ is symmetric about 0. Selecting about one third of the total features yields the best performance.

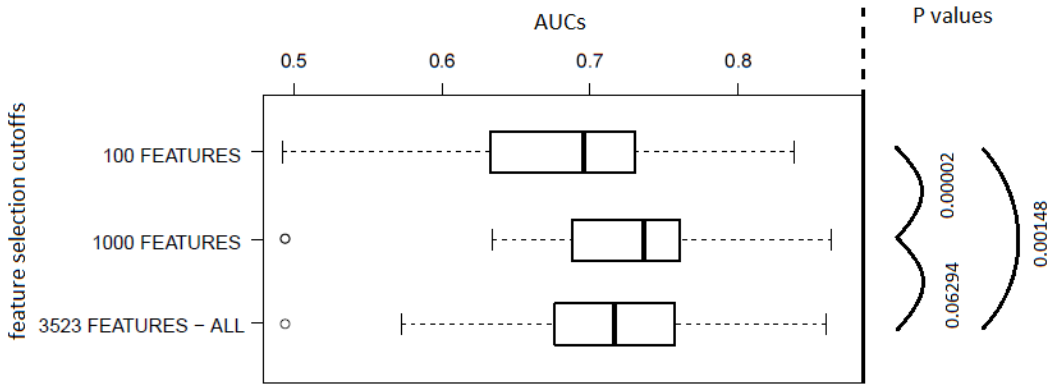

Figure 11. AUCs for three different feature selection cutoffs
We show a boxplot for three sets of data, representing the 30 interaction-type AUCs. A p-value is generated using the paired Wilcoxon test between each feature selection cutoff. The overall best cutoff is to use 1000 features.

3.1.3 Hierarchically-corrected Mechanistic Protein Networks

Park et. al. introduced a method for generating mechanistic interaction networks that incorporates the hierarchical relationships between interaction types encoded in an interaction ontology [19] [21]. The ontology of the 30 interaction types in Table 3 represents the cellular/molecular/epistemic organization of protein interactions [21] in the cell; for example, “peptide transfer” is a type of “physical interaction”, “transcriptional regulation” is a “regulatory interaction”, and “synthetic rescue” is a

"synthetic interaction" (Figure 12). We apply the hierarchical correction method [19] on our interaction-type networks learnt from the previous step.

We begin by superimposing a Bayesian network on the interaction-type ontology. Each original interaction type node i in Figure 12 now represents SVM outputs, (noisy observations Y_i) of the true interactions (latent event X_i) in the Bayesian network (Figure 13). Each true label X_i depends on its children from the ontology $X_{i1}, X_{i2}, \ldots, X_{ik}$ for each interaction type i. Each noisy observation Y_i depends on true labels X_i for each interaction type i.

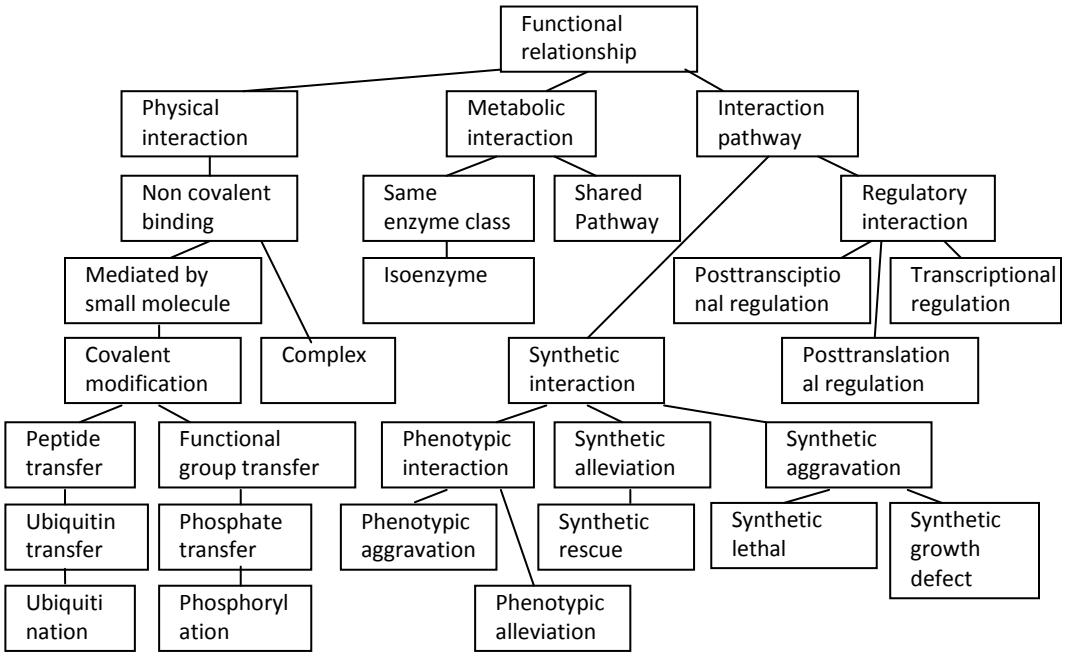

Figure 12. Interaction type ontology
Organized as a hierarchy, the various interaction-types relate to each other in levels of specificity and cellular organization. Leaf nodes are very specific interaction types, while higher nodes are more general.

With this model in hand, we can learn the structure and parameters of the Bayesian network from our data [112]. Maximum likelihood estimates for interactions convert the results of our SVMs into conditional probabilities for each pair. We obtained maximum likelihood estimates $P(Y_i|X_i)$ using expectation maximization for each interaction type i.

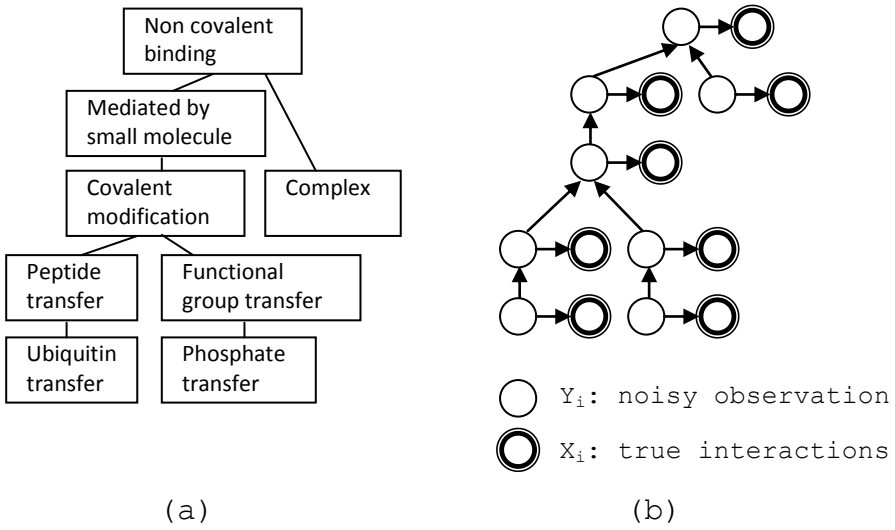

Figure 13. Bayesian superimposition over interaction type ontology
(a) A part of the interaction ontology. (b) Bayesian network corresponding to the part of the interaction-type ontology in (a). Noisy observations Y_i depend on true interactions X_i. Variables Y_i correspond to each node in the interaction-type ontology.

To incorporate the interaction-type hierarchy, we determined conditional probabilities for each interaction type and its children. For example, an interaction pair annotated to children "peptide transfer" and "phosphate transfer" is also an interaction pair in any of the parents of these two interaction types, like "covalent modification". In this example, we can write that:

$$1 = P(covalent\ modification = 1 \mid peptide\ transfer = 1)$$

$$1 = P(covalent\ modification = 1 \mid phosphate\ transfer = 1)$$

We can use maximum likelihood again to count the number of training labels to obtain other parameters: $P\left(covalent\ modification = 1 \left| \begin{matrix} peptide\ transfer = 0, \\ phosphate\ transfer = 0 \end{matrix} \right.\right)$

In the end, general functional relationships are not sufficient for predicting protein interactions; mechanistic networks can help us distinguish the interactions of different types, for example, "metabolic interaction", "physical interaction", or "regulatory interaction".

3.1.4 Protein-compound interaction networks

The technique for generating protein-compound interaction networks was inspired by Guan et. al [113]. They used SVMs to determine a set of genes that are associated with a certain phenotype (observable trait) in the laboratory mouse. In the following subsections, we describe a similar technique to infer a set of genes that are associated with a certain compound (small molecule) in yeast.

3.1.4.1 Gold Standards

Our main resource for compound-protein interaction gold standards is the Search Tool for Interactions of Chemicals (STITCH) [39]. STITCH aggregates compound and protein interactions annotated in a few different sources, including DrugBank [114],

PubMed [115], and PharmGKB [116]. They gathered information from over 25 different databases and used direct database interaction entries, text mining, and known interactions to produce likelihood scores for compound-protein pairs. STITCH data consists of a long list of pairs, which are compounds and proteins for many different organisms along with a score given based on database, experimental, and text mining evidences. To select the compounds to investigate we filtered their list based on two criteria:

(a) The interaction should be between a compound and a yeast protein

(b) The interaction should have a STITCH score greater than 0.7

Filtering the STITCH database using these criteria yielded 475 compounds each known to interact with at least 5 yeast proteins. Only two of these compounds are known and approved drugs.

Therefore, to annotate more compound-protein pairs, preferably drug-protein pairs, we used an existing method of transferring functional knowledge between organisms [117]. Unlike yeast, several human proteins have been studied for interaction with drugs. This knowledge could be transferred from human to yeast to infer potential drug-protein interactions. Identifying homologs with conserved functional roles for knowledge transfer between organisms improves coverage and was first used by Chikina et. al. [118] to transfer annotations using sequence similarity; it was extended by Park et. al. [117]to identify homologs with similar functional profiles by network-based methods. In this chapter, we leverage the functionally analogous human-yeast

gene pairs from these previous studies to enrich our original yeast compound-protein set to include more drug-protein pairs. We do this by additionally picking interactions between compounds and human proteins (in STITCH), and transferring the knowledge to infer to interactions between those compounds and yeast proteins. This procedure increased our compound set by 227 compounds, 11 of which are approved drugs. In total we have 702 different compounds.

Table 4. Compounds that are drugs
List of the 13 drugs in the compound set along with their descriptions from DrugBank [114].

DRUG NAME	DRUG DESCRIPTION
Adenocard	Treats irregular heartbeat (arrhythmias).
Bortezomib	Treats multiple myeloma and mantle cell lymphoma.
Colchicin	For treatment and relief of pain in attacks of acute gouty arthritis.
Daunorubicin	For remission induction in acute nonlymphocytic leukemia (myelogenous, monocytic, erythroid) of adults and for remission induction in acute lymphocytic leukemia of children and adults.
Doxorubicin	For the treatment of Koposi's sarcome connected to AIDS.
Estradiol	For the treatment of urogenital symptoms associated with post-menopausal atrophy of the vagina (such as dryness, burning, pruritus and dyspareunia) and/or the lower urinary tract (urinary urgency and dysuria).
Etoposide	For use in combination with other chemotherapeutic agents in the treatment of refractory testicular tumors and as first line treatment in patients with small cell lung cancer. Also used to treat other malignancies such as lymphoma, non-lymphocytic leukemia, and glioblastoma multiforme.
Famoxadone	Fungicide to protect agricultural products against various fungal diseases
Irinotecan	For the treatment of metastatic colorectal cancer
Lovastatin	For primary prevention of coronary heart disease and to slow progression of coronary atherosclerosis in patients with coronary heart disease.
Rapamycin	For the prophylaxis of organ rejection in patients receiving renal transplants.
Streptozocin	For the treatment of malignant neoplasms of pancreas (metastatic islet cell carcinoma).
Tamoxifen	For the treatment of breast cancer.

3.1.4.2 Data Compendium

Our data compendium for predicting interactions between compounds and proteins consisted of the 30 different interaction-type networks. As a reminder, each network was created by integrating known interactions of that type along with microarray, protein domain, protein family, localization, and structural data using an SVM classifier; the interaction probabilities were corrected by leveraging the hierarchical structure of the 30 interaction types. This data compendium is essentially a large interaction probability matrix where the rows are all the yeast genes (indexed i from 1 to N and the columns are the yeast genes (j) copied 30 times (k), to include the 30 interaction types, with each cell $(i, j + N(k-1))$. Containing the probability of gene i interacting with gene j in network k.

3.1.4.3 Support Vector Machine Classification

For building the final compound-gene networks (part (c) in Figure 9), we employed a separate SVM classifier for each of the 702 compounds; this way, each SVM predicts the interaction probabilities between that compound and all the yeast genes. The feature vector for a particular gene consists of all the protein interaction probabilities from our 30 mechanistic networks. The feature vector looks as follows:

$$\boldsymbol{G}_i = \{[P_{i11}, P_{i21}, \dots, P_{iN1}][P_{i12}, P_{i22}, \dots, P_{iN2}] \dots [P_{i1k}, P_{i2k}, \dots, P_{iNk}]\}$$

where $N = 5559$ yeast genes, i is the ith gene for $i \in [1, N]$, $k = 30$ interaction types. We used a linear kernel SVM available in the Sleipnir library [108].

3.1.5 Parametric Analysis of Gene Set Enrichment and Canonical Correlation Analysis

For each compound, we used parametric analysis of gene set enrichment (PAGE) [119] to determine if that compound interacts with any biologically coherent set of proteins with surprisingly high probabilities . PAGE for a compound calculates a z-score for each protein set by comparing the mean interaction probability of that set to the expected mean probability of a set of proteins of the same size:

$$z = \frac{(Sm - \mu)m^{1/2}}{\delta}$$

where μ and δ are the mean and standard deviation of distribution of interaction probabilities of that compound to all proteins, and Sm is the mean probability of interaction of the m proteins of interest, . We curated known biologically meaningful gene sets from GO (biological processes) and INTERPRO (protein sequence features) [100], and then used PAGE to calculate the association of these protein sets to each compound.

Next, to explore whether specific chemical compound groups associated with specific protein sets discovered in the previous analysis, we organized all the compounds into meaningful classes based on the Chemical Entities of Biological Interest (ChEBI), a curated hierarchy of compounds. We employed a simple method of determining a set of classes that were specific enough in the ontology to provide detailed information but general enough to have enough compounds from our compendium annotated to each class. For each compound, we compiled a list of all the pharmacological classes

we know it belongs to in the ChEBI hierarchy. To obtain a “fringe list” of chemical classes that is representative of all the compounds, we parsed the hierarchy bottom up, starting at each leaf, and kept the first terms that had at least 5 compounds from the total 702 compounds annotated to the hierarchy term. We manually re-curated the list of terms to eliminate very generic terms, yielding 40 classes.

To better realize a relationship between the biological processes and ChEBI compound classes, we additionally use regularized canonical correlation analysis by way of the R package CCA [120]. Canonical correlation analysis (CCA) aims to find relationships between the experimental units of two data sets. In our case, experimental units are drugs and the two data sets are biological processes and the protein families/domains. We use this method to visualize four things: (1) how drugs cluster and which biological process groups influence the drugs, (2) how drugs cluster and which protein domains/families influence the drugs, (3) how compound classes cluster and which biological process groups influence the classes, and (4) how compound classes cluster and which protein domains/families influence the classes.

CCA by itself assumes that the common units (in this case, drugs) between the experiments are larger than the number of experiments. In our case, this is not true; we have less drugs (13) than biological processes (227) or protein families/domains (526). For experimental vectors X and Y, the correlations that we want maximized are the linear combinations between

$$U^1 = Xa^1 = a_1^1 X^1 + a_2^1 X^2 + \cdots + a_p^1 X^p \text{ and}$$

$$V^1 = Ya^1 = a_1^1 Y^1 + a_2^1 Y^2 + \cdots + a_q^1 Y^q$$

We use the regularized CCA, which includes an extra step that performs leave-one-out cross validation to maximize the correlation

$$\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2) = arg \max_{\lambda_1, \lambda_2} CV(\lambda_1, \lambda_2)$$

where $CV(\lambda_1, \lambda_2) = cor(\{X_i a_\lambda^{(-i)}\}_{i=1}^n, \{Y_i b_\lambda^{(-i)}\}_{i=1}^n)$ where X_i and Y_i are particular experiments from each dataset that are left out. The graphical representation of CCA is typically in the form of a unit circle, where experiments are points on this circle. Radial lines from the center denote lines on which experiments are most similar and the farther they are from the center the more similar they are. A similar unit circle can be overlaid, where the common units are shown and the closest unit to an experiment represents the unit that is most representative of that experiment.

3.2 Results

To systematically assess the accuracy of our hierarchically integrated classifier that used Bayesian integration of independent SVMs to infer protein-compound interactions, we evaluate each step separately. We further analyze the resulting protein-compound networks to show how groups of compounds interact with groups of proteins. We remind the reader that our methods employ a two-step integrations

process where (a) we build mechanistic networks for 30 interaction types and (b) we leverage these networks to predict compound-protein pairs.

3.2.1 Adding Structure Data and Being Selective Improves Interaction-Type Predictions

To gain insights into how interaction-types play a role in gene-gene interaction predictions in yeast, we collected a large amount of diverse data. Using the method described in section 3.1.3, we integrated this data using SVMs and a Bayesian hierarchical correction method to provide protein-protein interaction networks in 30 different interaction types. We mentioned that including structural data in the form of docking and binding site conservation pairs has not been done in this setting. Having diverse data is always desirable as this provides more opportunity to explain biological phenomena from different perspectives. We now compare in Figure 14 the predictive power of each dataset alone: microarray expression experiments, sequence similarity, binding site conservation, and docking on the different interaction types. One thing worth mentioning is that the interaction types “posttranscriptional regulation”, “peptide transfer”, “ubiquitination”, and “ubiquitin transfer” have very few know protein interaction annotations. As such, they are harder interaction types to predict in because classification algorithms have fewer positive examples to extrapolate from.

Microarray data consist of a set of experimental conditions that show the gene expression level for a each experiment. Microarray data can be powerful tools despite

their sometimes noisy nature, and can often contain hidden patterns that may not have been intended to be studied in the original experiment. By themselves, they are generally able to predict interactions well, as we can see in Figure 14 (a). These microarray experiments seem to predict especially well the “complex”, the “non covalent binding”, and the “physical interaction” interaction types. As microarray data is so generic and encompasses a wide variety of interactions at work, it is able to predict these interaction terms that have the most gold standard gene pairs annotated to them really well.

Sequence similarity is hypothesized to occur due to similarity of function. This has not been proven, but several proteins stabilize their structures by common bonds between enzymes with similar catalytic residues. For certain protein domains, we can claim that if they have sufficient structural similarity then they have diverged from a common ancestor; we cannot claim that functional convergence leads to similarities in their sequences or structures [121]. In Figure 14 (b), we see that using sequence similarity data performs well for many of the interaction types, especially “complex”, “same enzyme class”, and “posttranslational regulation”. This implies that most of the sequence similarity data available shows that having structurally similar proteins will lead to interactions that relate to the proteins sharing an enzyme class forming a larger complex to achieve a goal in the cell. Having said that, having similar structures is not a prerequisite for two proteins to serve the same function or catalyze the same reaction. It has been shown, via a systematic search, that proteins with little sequence similarity catalyzed the same reactions [122]. Therefore, there is a need for other types

of data so that we may broaden our view of the interactions taking place and their causes. We will discuss two structural interaction types we included next.

Binding site conservation (BSC) data, as we mentioned before, aims to complement sequence similarity data such that it predicts complexes that form due to evolutionarily conserved sequences as well as sequences that have diverged. They use the structure of proteins and the many locations that they bind to each other to create binding site frequency contact maps; from these sites, they extrapolate how likely new protein pairs are to interact given their structures. Therefore, the BSC method does not exclusively rely on the sequence information to predict interactions. As such we see an increase in the predictive power of this data in Figure 14 (c). By far, BSC data predicts the best in the same enzyme class, and isoenzyme interaction types. We expected it to do well in the former because their analysis begins by classifying proteins in enzyme classes based on properties of their subunits. What is interesting is that they are able to predict interactions well in isoenzyme; this class contains pairs of proteins that differ in sequence but that catalyze the same reaction. By leveraging structural information and not just the protein sequences, they surpassed using sequence similarity in performance and are able to predict proteins pairs that catalyze the same chemical reaction despite them not having a similar sequence.

Docking interaction data provides plausible conformations of complexes between two proteins via a minimization between the energy required to place the proteins in those configurations. Thus, docking methods are based on chemical properties and physical

binding principles. In Figure 14 (d), predicting interactions using only docking does well in the “phenotypic interaction”, “synthetic aggravation”, “synthetic alleviation”, and “phenotypic aggravation” interaction types. The docking interaction data assumes that the two proteins bind together to form a complex; therefore a mutation, over-expression, or deletion of one of genes implies that the complex conformation will change; the original complex can no longer be formed under the same conformation and so the original phenotype or effect that the other gene was supposed to exhibit no longer holds true.

We typically look at the bar plots in Figure 14 to see how informative the interaction types are. It is possible to compare AUCs between different interaction types to see how well out method does, but the AUCs may be influenced by factors such as gold standard size (fewer gold standard protein pairs may lead to AUC overestimation) and gold standard quality (if a set of protein pairs were used to develop a particular algorithm, like docking, was similar to one of our sets then we would perform better than we otherwise would have).

We finally show in Figure 15 that when we predict and evaluate with all our data included (microarrays, protein domains, sequence similarity, localization, BSC, and docking), the AUC is higher than with any one of the datasets individually.

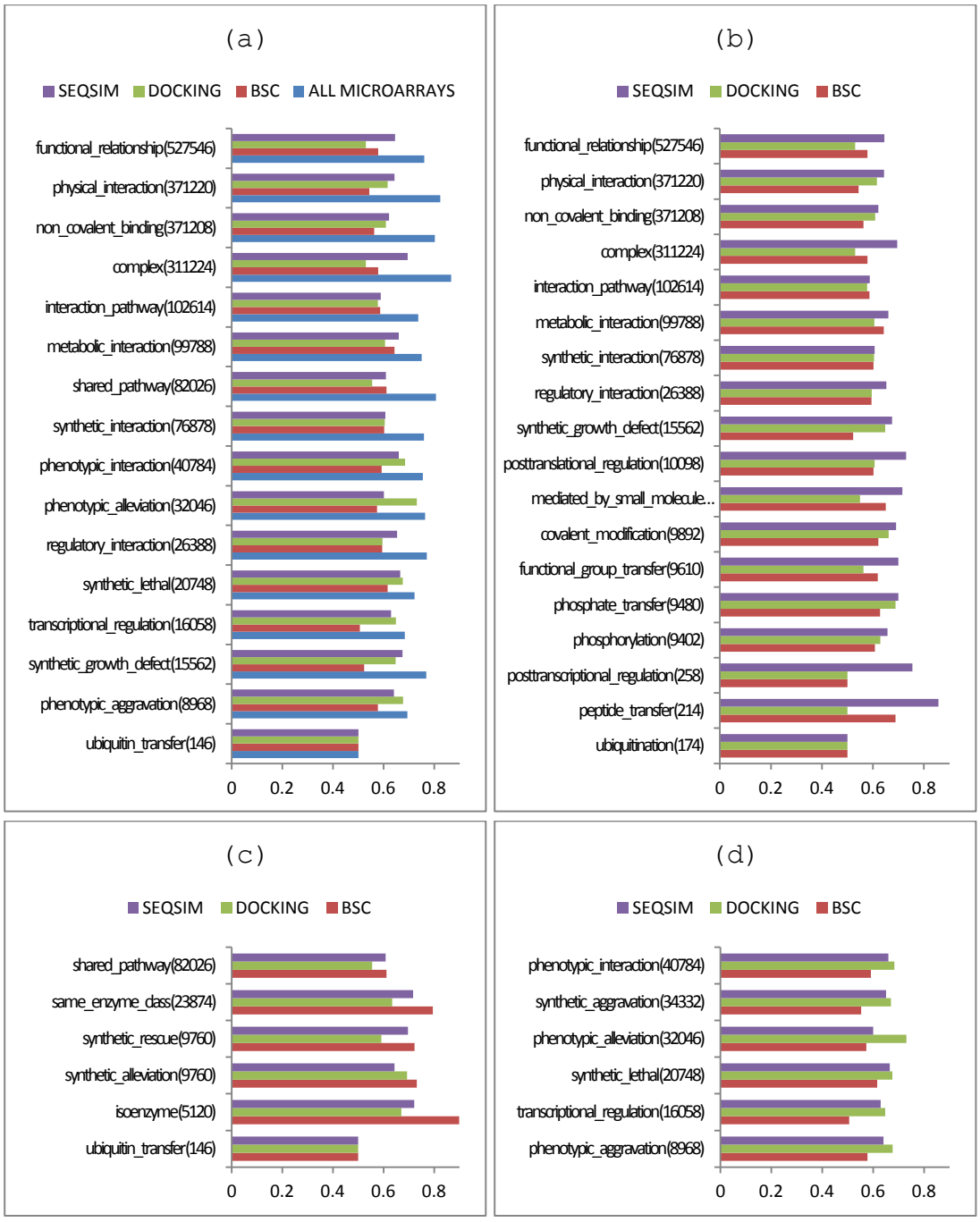

Figure 14. AUCs for different interaction types

AUCs when adding more data is increased for all interaction types. (a) shows that the interaction types where microarrays data performs better than binding site conservation, docking, or sequence similarity data. (b) shows the interaction types where sequence similarity performs better than binding site conservation or docking. (c) shows the interaction types where binding site conservation performs better than docking or sequence similarity. (d) shows the interaction types where docking performs better than sequence similarity or binding site conservation.

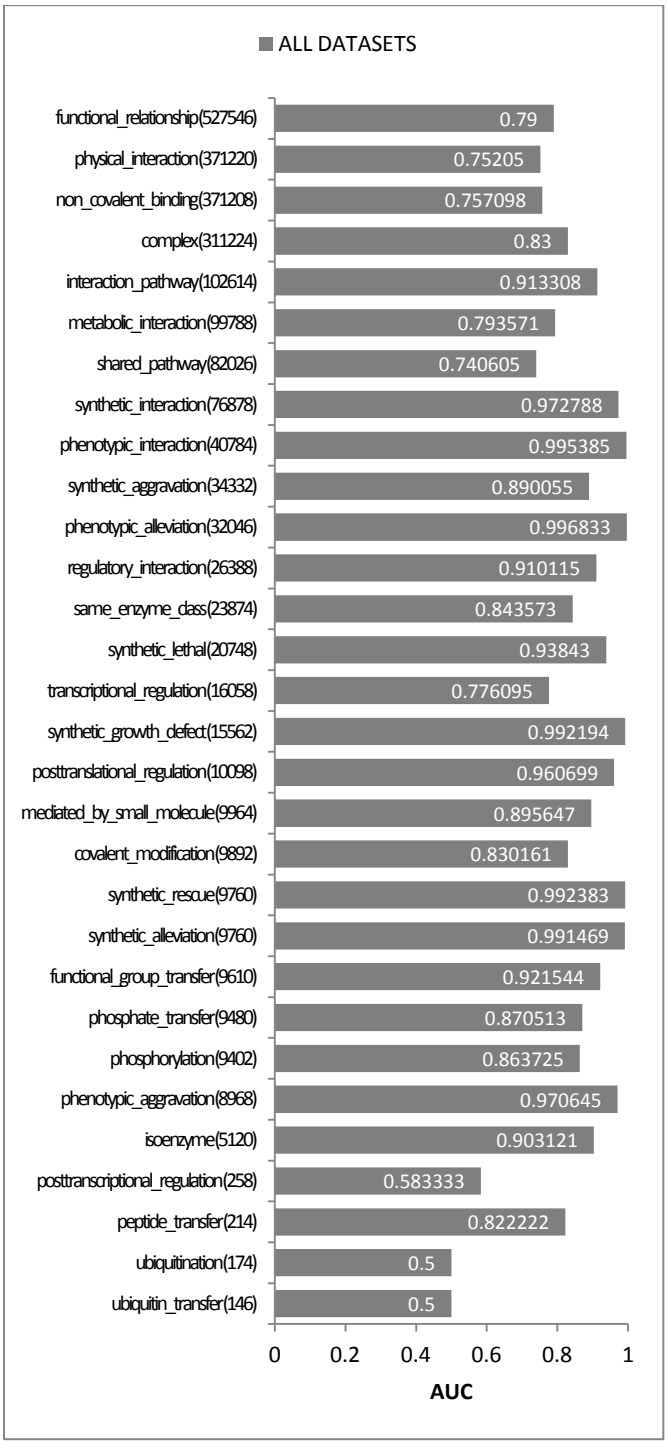

Figure 15. AUCs including all data
The overall evaluation of the 30 interaction-type networks when integrating all the dataset types (binding site conservation, docking, sequence similarity, microarrays) in the prediction method. Using all datasets evaluates better than using any one dataset individually.

3.2.2 Interaction Networks Predict Gene-Compound Interactions Well

In Figure 9 (a) and (b) we used Bayesian integration to hierarchically correct 30 independent SVMs and yield 30 interaction-type protein-protein interaction networks. Using the gold standard from the protein-compound interaction database STITCH and our integration pipeline from Figure 9 (c), we ask another question: how does our network perform across all the 702 compounds? To answer this, we replaced our protein-protein interaction network (Figure 9 (a) and (b)) with another input network. In Figure 16, we compare how four different networks evaluate across all compounds: a "global yeast network", our "interaction-type aggregation network", our "interaction-type aggregation network with gold standards", and STRING, each explained below.

The Search Tool for Retrieval of Interacting Genes/Proteins (STRING) [123] uses a mixture of three methods (gene fusion, genome neighborhoods contexts, and phylogenetic profiles by measuring mutual information) to predict functional annotations between gene pairs. They alter their final network predictions by iterating through their pairs and reweight those that are in their gold standard to have a probability of 1. Our "interaction-type aggregation network" is the network resulting from hierarchically correcting independent SVMs (that integrated a multitude of diverse data) using Bayesian integration. Our "interaction-type aggregation network with gold standards" is the network previously mentioned with pairs from our gold

standard modified to have a predicted interaction probability of 1, in a similar manner to STRING. The “global yeast network” predicts functional interactions via (a) a Bayesian integration of heterogeneous data and then (b) a transfer of homolog functional knowledge between organisms using an SVM trained on each GO term.

The bars plot in Figure 16 shows the number of compounds that fall into a specific AUC evaluation bin for each of the four networks. In general, the “global yeast network” performs with an average AUC of 0.71 across all compounds; while it encompasses much functional information, it does not capture certain pathway level interactions that seem to contribute to the improved prediction accuracy in some of the compounds. Our network, without modifying the gold standard pairs, performs better, with an average AUC of 0.75 across all compounds. In comparison to the global yeast network, including interaction-type information helped predict compound-interactions better. Since STRING converted their final predictions pairs from their gold standard to 1, we did the same to our network. We see that our network with gold standards performed better than STRING; our average AUC across all compounds was 0.80 versus 0.79 for STITCH and we were able to predict more compounds with higher AUCs than STRING.

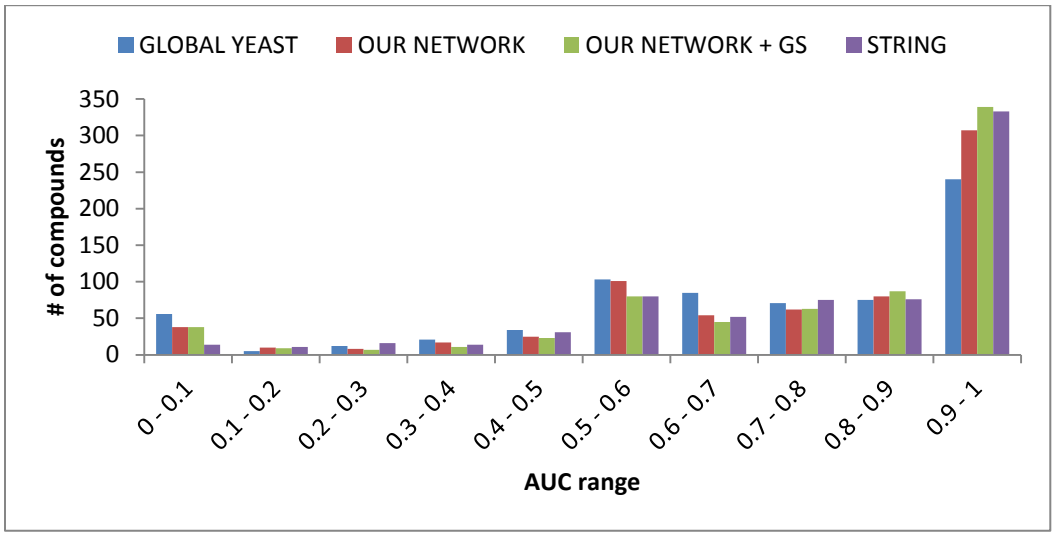

Figure 16. Network evaluation across all compounds
Here we plot the number of compounds that had AUCs falling in ten different AUC bins. There were some compounds (those with a small number of gene gold standard annotations) that had AUCs less than 0.5, meaning that we predicted them worse than random. For AUCs in ranges 0.5 to 0.8, the global yeast and our original network without gold standard conversion predict more compounds in those bins. For AUC ranges higher than 0.8, our network with gold standard conversion and STRING predict more compounds than the other two networks. The average AUC across all compounds is 0.71 for global yeast, 0.75 for our network, 0.8 for our network with gold standard conversion, and 0.79 for STRING.

3.2.3 Biological Evidence and Analysis of Biological Process and Protein Family Enrichment

We will investigate four drugs from our small molecule set:

- Bortezomib, a proteasome inhibitor that blocks the action of cellular complexes that break down proteins and is thus effective at anti-tumor activity
- Rapamycin, a drug for the prophylaxis of organ rejection in patients receiving renal transplants
- Doxorubicin, an antibiotic used in cancer chemotherapy, for the treatment of Koposi's sarcome connected to AIDS.
- Famoxadone, a fungicide to protect agricultural products against various fungal diseases

A commonly used method for determining sets of proteins that contribute to interactions is called gene-set enrichment, described in the Methods section. Briefly, gene-set enrichment tests whether seeing a group of proteins together is less likely than seeing a randomly chosen group of proteins together. We used PAGE, as described in the Methods section 3.1.5, to determine if the top predicted proteins interacting with a particular compound belong to a biological category with a higher likelihood than a randomly chosen group of proteins. The biological categories we are interested in are biological processes from GO [8] and protein domains and families from InterPro [124]. Next follows a discussion of the gene-set enrichment analysis.

Table 5. Gene-set enrichment of biological processes and protein families for four drugs

In this table, the columns represent 4 studied drugs from our compound set. The first 5 rows correspond to the top 5 biological processes that each drug was enriched for. The last 5 rows correspond to the top 5 protein families/domains that each drug was enriched for.

	BORTEZOMIB	RAPAMYCIN	DOXORUBICIN	FAMOXADONE
BIOLOGICAL PROCESS ENRICHMENT	proteasomal ubiquitin-independent protein catabolic process	chromatin modification	chromatin modification	cellular respiration
	proteasome assembly	ribosomal small subunit biogenesis	RNA splicing	glucose catabolic process
	ubiquitin-dependent protein catabolic process	ribosomal large subunit biogenesis	nucleosome organization	ATP biosynthetic process
	modification-dependent protein catabolic process	nucleosome organization	protein acetylation	nicotinamide nucleotide metabolic process
	proteasome regulatory particle assembly	ribonucleoprotein complex assembly	nucleotide-excision repair	pyruvate metabolic process
	proteasomal protein catabolic process	transcription elongation, DNA-dependent	DNA-dependent DNA replication	serine family amino acid biosynthetic process
	proteasomal ubiquitin-dependent protein catabolic process	chromatin assembly or disassembly	transcription initiation from RNA polymerase II promoter	fermentation
PROTEIN DOMANIN AND FAMILY ENRICHMENT	Family Proteasome, subunit alpha/beta	Domain Nucleotide-binding, alpha-beta plait	Domain Protein kinase, catalytic domain	Domain NAD(P)-binding domain
	Conserved_site Proteasome, beta-type subunit, conserved site	Domain RNA recognition motif domain	Domain Serine/ threonine-protein kinase-like domain	Domain Aldolase-type TIM barrel
	Domain Proteasome, alpha-subunit, conserved site	Domain Histone-fold	Active_site Serine/threonine-protein kinase, active site	Conserved_site Isocitrate/isopropylmalate dehydrogenase, conserved site
	Family Proteasome A-type subunit	Repeat WD40 repeat 2	Binding_site Protein kinase, ATP binding site	Domain Isopropylmalate dehydrogenase-like domain
	Family Proteasome B-type subunit	Domain WD40-repeat-containing	Domain Histone-fold	Domain Pyridoxal phosphate-dependent

		domain		transferase, major region, subdomain 1
	Family 26S proteasome subunit P45	Domain Helicase, C-terminal	Domain Protein kinase-like domain	Domain Pyridoxal phosphate-dependent transferase, major domain
	Domain Proteasome component (PCI) domain	Repeat WD40 repeat	Domain Serine/threonine-protein kinase domain	Domain Pyridine nucleotide-disulphide oxidoreductase

Bortezomib is a proteasome inhibitor, so it makes sense that many of the top biological processes and domains/families are proteasome related. Literature curation of rapamycin showed an experiment where rapamycin inhibited liver growth in rats by controlling ribosomal protein translation [125]. Targets of rapamycin have been shown to be WD40 in previous literature [126]. Doxorubicin treats many cancers. Literature curation showed that serine proteases mediates doxorubicin-induced apoptosis (cell death) [127] and that treatment of certain cells (U2OS) with doxorubicin induces phosphorylation at certain serine residues [128]. Famoxadone is a oxazolidinedione fungicide that inhibits mitochondrial respiration and decreases ATP production in the fungal pathogens [129]. Because of this, we are not surprised to see cellular respiration and the ATP biosynthetic process as the top enriched terms.

3.2.4 Clustering analysis

We extended the analysis in the previous subsection from just the top 5 enriched terms to all terms from 227 biological processes (from GO) and 526 protein

families/domains (from InterPro). We perform hierarchical clustering of these scores with respect to drugs and compound classes.

3.2.4.1 Drugs versus Biological Processes and Drugs versus Protein Families/Domains

Hierarchical clustering with a Pearson correlation coefficient was used to group drugs and biological processes or protein families/domains into similar clusters. This method was applied on the z-score values obtained from the PAGE analysis. We used the Pearson correlation metric to allow for distinction between vectors of PAGE z-score values that have a change of magnitude and direction; for example, if drug A's z-scores increase over the set of biological processes but drug B's z-scores decrease over the same set, then drugs A and B have a Pearson correlation coefficient of -1. Pearson correlation is calculated as follows for $X = \{x_1, x_2, \dots, x_N\}$ and $Y = \{y_1, y_2, \dots, y_N\}$ being z-scores for drugs all drug pairs X and Y:

$$\frac{\sum(x_i, y_i) - \frac{\sum x_i \sum y_i}{N}}{\sqrt{\left(\sum(x_i^2) - \frac{(\sum x_i)^2}{N}\right)\left(\sum(y_i^2) - \frac{(\sum y_i)^2}{N}\right)}}$$

Where $\sum(x_i, y_i) - \frac{\sum x_i \sum y_i}{N}$ is the sum of products, and $\left(\sum(x_i^2) - \frac{(\sum x_i)^2}{N}\right)$ and $\left(\sum(y_i^2) - \frac{(\sum y_i)^2}{N}\right)$ are sums of squares. Using these scores, clustering begins with the highest correlated pair of drugs. The drugs are merged into a new drug with z-scores the average of each (x_i, y_i) and Pearson correlation is found between this merged drug

and all the other drugs. The process is continued until there are no more drugs to merge.

We used this method to clustering z-scores belonging to 13 drugs and 227 biological processes from the PAGE analysis in the previous subsection into a hierarchical clustering heatmap called DvBP. Similarly, we obtained a heatmap for 13 drugs versus 526 protein families/domains that we call DvPFD. In this section, we will look at one of the resulting clusters from the DvPFD heatmap.

Figure 17. Drugs versus Protein Domains/Families (DvPFD) cluster
We show one of the clusters resulting from the PAGE analysis. Half of the drugs are clustered as interacting with certain proteins families and the other half are not. The drugs that cluster together have protein interaction predictions enriched for multidrug transporter proteins.

The darker boxes show that 6 drugs (tamoxifen, lovastatin, colchicin, estradiol, adenosine, and irinotecan) have high interaction probabilities with proteins that are a part of the families “Sugar/inositol transporter”, “General substrate transporter”, “Conserved site Sugar transporter”, and “Major facilitator superfamily/domain”. According to InterPro descriptions, these families include sugar transporters and multidrug transporters. Cells that have a high expression of multidrug transporter proteins and are treated with a particular drug will expel the drug out of them; those cells are now resistant to that particular drug [130]. Research into the literature finds studies that confirm this behavior with the drugs mentioned as interacting with proteins from these families. For example, tamoxifen is a breast cancer treatment drug and its metabolites has previously been shown to be involved with drug transporters [131]. Lovastatin is a drug used to prevent coronary disease, while colchocin is a drug used to relieve pain caused by gouty arthritis and both have been shown to interact with MDR1, a multidrug transporter [132] [133]. For the drugs not shown to interact with those drug transport families, studies typically show that drug targets are only activated under certain conditions. For example, rapamycin is not predicted to interact with those drug families and a study showed that rapamycin modulated multidrug target only if it was previously incubated to reach peak blood concentrations before exposing the drug to the cells [134].

3.2.4.2 Compound Classes versus Biological Processes and Compound Classes versus Protein Families/Domains

Hierarchical clustering with a Pearson correlation coefficient was also used to group compound classes and biological processes or protein families/domains into similar clusters. The compound classes were derived as explained in the Methods section 3.1.5. In this section, we will examine a particular cluster (Figure 18).

Compound classes such as “ester”, “organic phosphate”, and “phosphoric acid derivative” seem to have the same behavior and to be highly interactive with metabolic terms such as “glycogen metabolic process” and “protein dephosphorylation”. Those same compound classes lack the same interaction intensity with transport-related biological processes such as “protein import into nucleus” and various RNA processing terms. This makes intuitive sense; “ester”, “organic phosphate”, and “phosphoric acid derivative” are ATP-driven compound classes. To a large extent, the metabolic cluster is more ATP-dependent, while the transport cluster is not directly dependent.

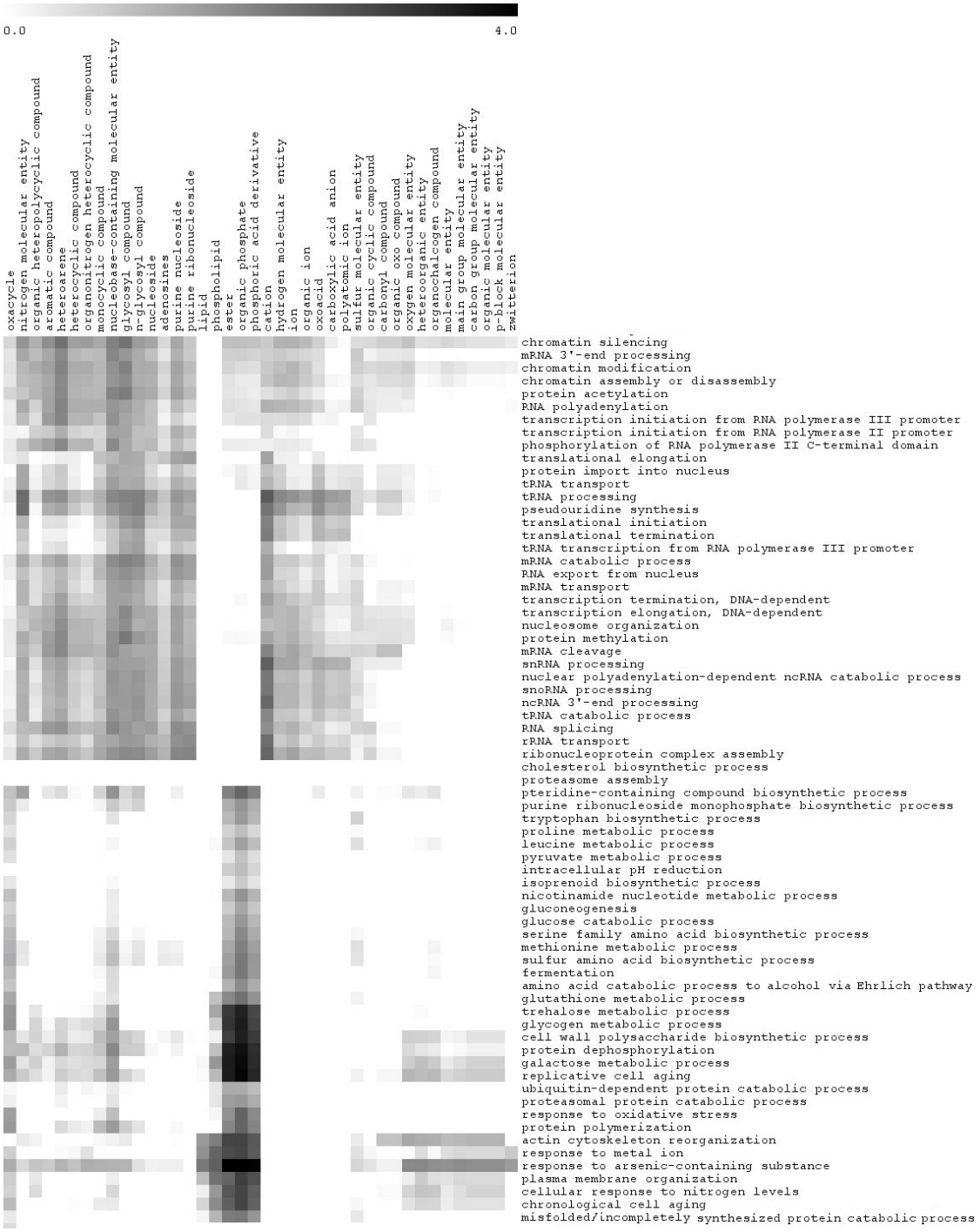

Figure 18. Compound classes versus Biological Processes cluster
This is one example of the clustering apparent from our PAGE analysis. We see a clear separation between interaction of the lipids classes and other classes with metabolism-related biological processes.

Using the regularized canonical correlation analysis (CCA) described in section 3.1.5, we found similar trends. Regularized CCA, as previously mentioned, organizes data points such that drugs behave similarly based on which side of the circle they fall on: left or right. Figure 19 shows that lovostatin, colchicine, tamoxifen, daunorubicin, and famoxadone fall on the right half of the circle. Additionally, the trend we observed in the DvPFD cluster in Figure 17 (these drugs are all related to transport biological processes and protein families/domains) is reinforced; drugs that clustered together in Figure 19 interact with genes related to transport proteins.

The left half of the circle in Figure 19 comprises of rapamycin, doxorubicin, etopside, and streptozotocin. The correlation with the "RNA processing" and "ribosome" clouds are validated in literature. For example, doxorubicin interacts with DNA by including a molecule between two other molecules and by inhibiting the creation of complex chemical products [135] [136] [137]. Etopside is a cancer drug acts by enabling DNA strands to break and inhibit the possibility of DNA molecules to re-ligate [138]. Rapamycin is known to have gene targets that regulate cell growth and nutrient sensing. An experimental procedure introduced rapamycin in yeast cells whose strands were damaged by UV. They found that the rate of DNA strand repair decreased [139], implying that the impact of the drug was at the DNA-level. Further, rapamycin has an impact on the TOR (target of rapamycin) signaling pathway; other studies found that rapamycin induces a rapid reduction in the ribosomal-protein mRNAs [140] [141]. Finally, streptozoticin is also known to add a methyl group molecule to certain DNA nucleotides, produce DNA strand breaks, changes in the structure of chromosomes,

and possibly cell death [142]. In contrast for example, the drug tamoxifen (not part of this cluster) has the potential for DNA damage properties but results are inconclusive [143].

Regularized CCA determined that the drugs on the extremes of these two halves (left-right) have very high clustering tendencies. Weaker clusters can be found between the remaining drugs, bortezomib, irinotecan, estradiol, and adenosine. For example, in Figure 19, bortezomib is functionally enriched in lipid-related protein families and stress-response biological processes. A few experiments attempted to prove this; in [144] they found that bortezomib triggers oxidative stress response and [145] found that treatment with bortezomib showed a reduction of lipid peroxidation (oxidative degradation of lipids), among other effects.

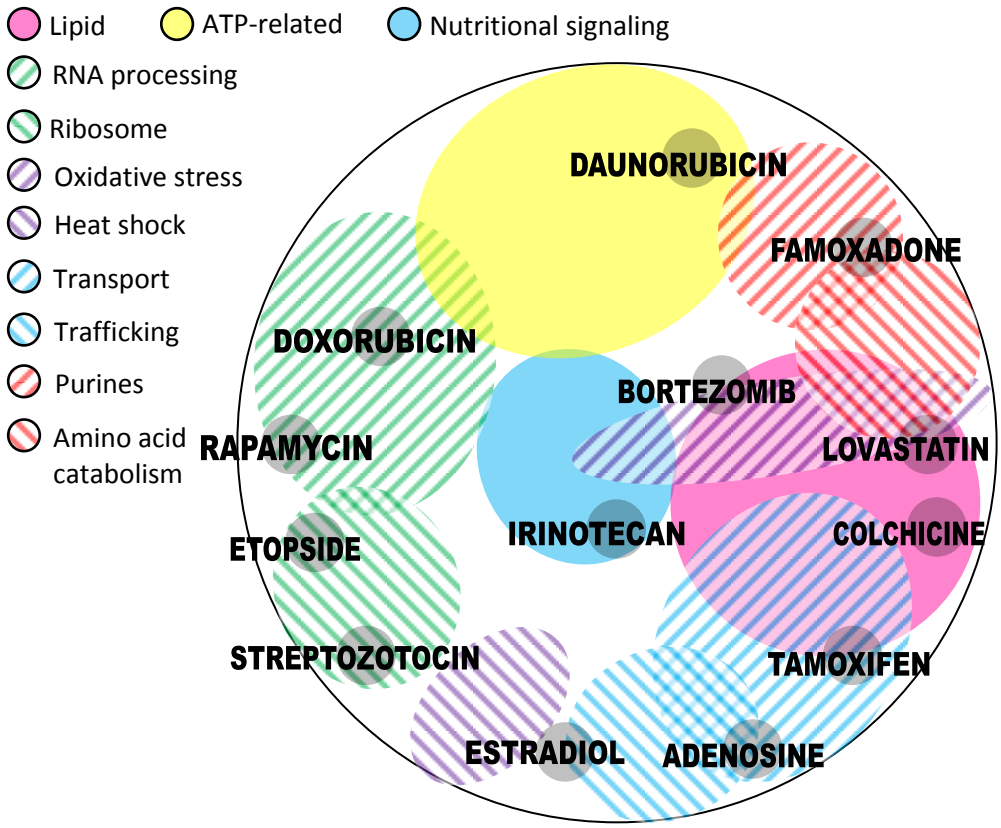

Figure 19. Regularized CCA on drugs
We show the common cluster clouds of biological processes and protein families/domains. Drugs are shown as grey dots and those farthest left are the strongest clustered together and similarly for drugs farthest right. The colored areas represent functionally enriched categories (biological processes and protein families/domains).

Finally, we performed a similar analysis on a broader scale. We looked at what chemical compound classes tended to correlate with functionally enriched categories of biological processes and protein families/domains, shown in Figure 20.

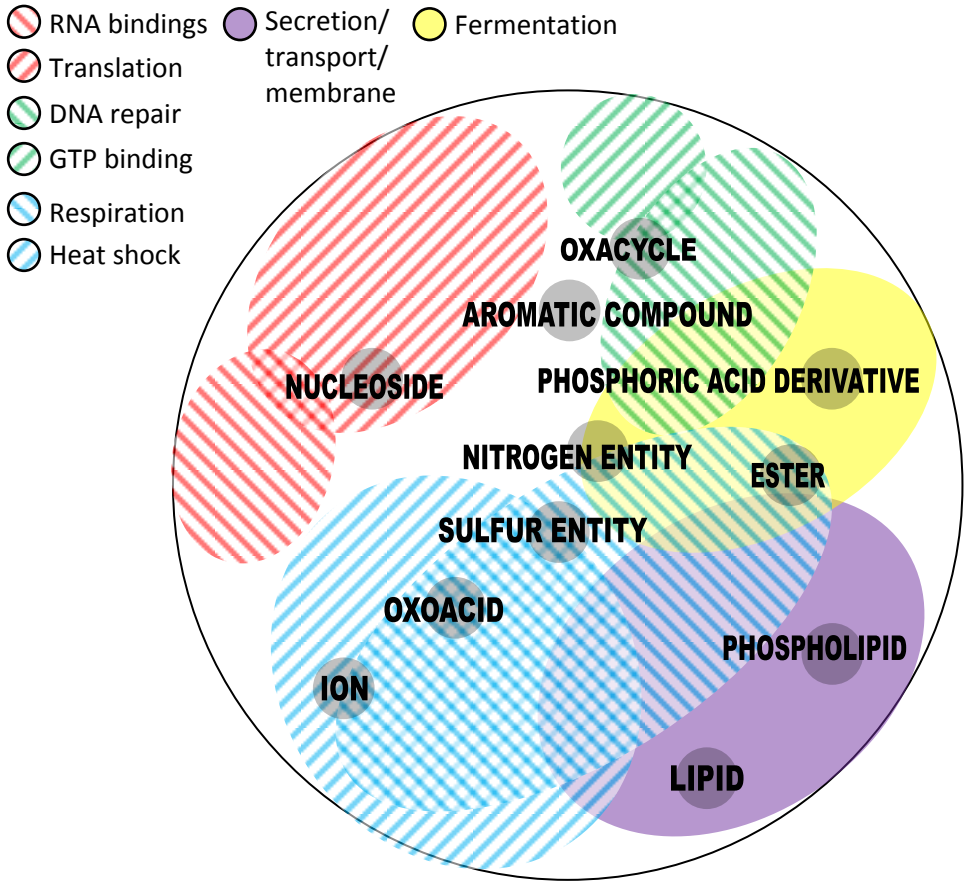

Figure 20. Regularized CCA on compound classes
We show the common cluster clouds of biological processes and protein families/domains. Compound classes are shown as grey dots and those farthest left are the strongest clustered together and similarly for compound classes farthest right. The colored areas represent functionally enriched categories (biological processes and protein families/domains).

In the analysis of Figure 20, the overlaps between the chemical classes and functional categories are straightforward. For example, we see that biological processes in the "secretion/ transport/ membrane" cloud such as "phospholipid catabolic process" and "lipid translocation" along with protein families/domains such as "Family Glycolipid", "Domain C2 calcium lipid binding domain", and "ATPase P-type phospholipid

translocating flippase" are correlated to the lipid and phospholipid classes of compounds. The "secretion/transport/membrane" cloud has, not surprisingly, known associations with lipids in literature [146].

Of particular interesting note are the high-level relationships that Figure 20 reveals between "ester" and "oxoacids". The combination of oxoacids and an alcohol (or phenol) reacts to produce esters. From the correlation metric, many stress response terms seem to span oxoacids and esters. Given the ubiquitous nature of esters (they are found in natural fats and oils, used in fragrances, certain explosives, and polyester plastics) they have many diverse and useful functional interactions, implying that they react to various stresses. For example, one of the highest scoring stress responses in esters was invasive growth response to glucose limitation. In [147] they found that a 3-isopropylmalate methyl ester signals yeast cells to switch their behavior to invasive growth when starved of amino acids. This invasive behavior causes cells to adhere to other cells and encode a glycoprotein attached to the cell wall [148]. Indeed, in the regularized CCA results, there are several glycoside-hydrolase-related domains and families predicted to correlate with the biological process invasive growth response to glucose limitation. This is related to fermentation, as glucose fermentation under various conditions showed different phenotypes in yeast [149]. Nitrogen, lipid, or glucose nutrient limitations prevented fermentation to occur in yeast [150].

The intuition behind the "respiration" cloud around the "ion" class in Figure 20 stems from the definition of aerobic respiration, where mitochondria are oxidized when

oxygen accepts electrons and starts a process that produces energy and then ATP. Without oxygen, the process of fermentation happens, which can be seen by the non-overlapping “fermentation” cloud and “respiration” cloud.

The last analysis in Figure 20 will be to look at the “oxacycle” cloud. Oxacycle as a term is too generic to describe small molecules. When we look at the compounds that were classified as an oxacycle, most are methionine variations, one is homocysteine, and one is nicotinamide. All these compounds have been linked to DNA damage and repair in literature. High doses of nicotamide inhibits an enzyme that detects DNA damage (PARP-1) from rejoining broken DNA strands *in vitro*, effectively halting DNA repair [151]. Homocysteine has been found to impair DNA repair [152]. Introducing selenium in the form of seleno-L-methionine to certain cell types has induced a DNA repair response [153].

3.2.5 Interaction Types between Protein-Protein Pairs Contribute to Compound-Protein Pairs

One final interesting analysis we will mention is to look at which interaction types contribute most to interactions between a particular compound and other proteins. In this section, we will discuss the drug bortezomib. The results of the SVM integration to predict compound-protein interactions (Figure 9 (c)) yield model weights for each input feature (as a reminder, our features were a long vector of protein interaction probabilities between all protein pairs for all 30 interaction types).

To study the interaction types that contributed most to bortezomib's interactions with various proteins, we averaged the model weights across all proteins for the 30 different interaction types for this drug. The highest average weights were attributed to "phosphorylation" (turns protein enzymes on or off) and "synthetic lethal" (combining mutations in separate genes in the same cell under the same conditions results in lethality) interaction types. Hence, phosphorylation is the most prevalent interaction type contributing to bortezomib's interactions with protein targets. This is validated in the inherent mechanisms behind bortezomib, as we explain next.

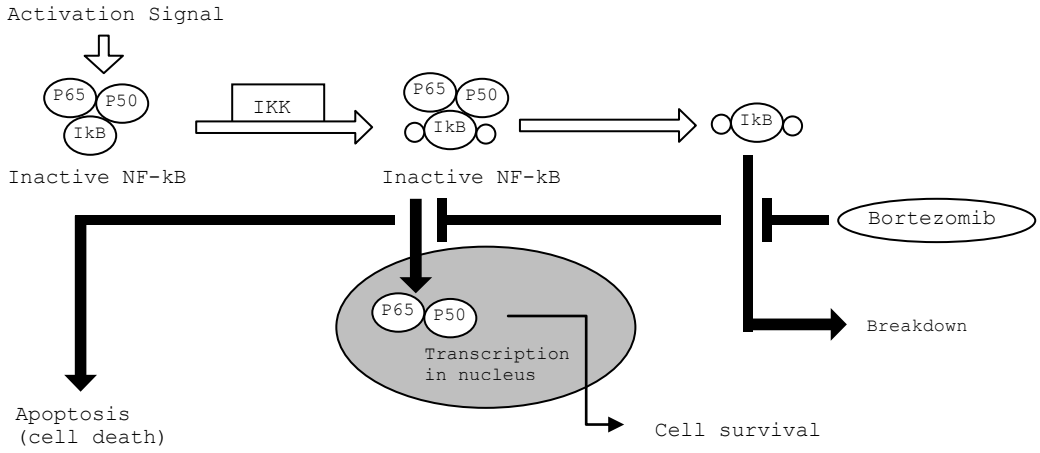

Figure 21. Bortezomib impact on the NF-kB pathway
Image replicated from [154]. An activation signal causes ikB to be phosphorylated by IKK. Once ikB is degraded by the proteasome, NF-kB enters the nucleus and enables genes to be transcribed to keep the cell alive. Introducing bortezomib inhibits the proteasome degradation mechanism and NF-kB cannot be activated; cells will now be vulnerable to death by other chemotherapeutic drugs.

First, we will discuss "phosphorylation". A cell responds to stress in various ways in order to survive. As shown in Figure 21, the normal process, without introducing

bortezomib, allows the protein complex NF-kB to enter the nucleus, where it begins the transcription of several genes that enable cell survival. Once protein IkB is phosphorylated by protein complex IKK, introducing bortezomib stops the protein complex NF-kB from entering the nucleus and initiating transcription [154]. This prohibits proteins from being activated to keep the cell alive, so it is now vulnerable to other chemotherapeutic drugs.

Second, we will discuss "synthetic lethality". Table 5 showed the biological processes and protein families/domains most enriched to bortezomib were those related to proteasomes (protein complexes regulate concentrations of other proteins by degrading the damaged ones). Another study set out to prove the common belief that bortezomib induces cell death by binding to the 20S cure subunit of proteasomes. They knocked down the 20S core in the presence of bortezomib and found that gave a synthetic lethal phenotype [155]. Their data indicated that knocking the 20S subunits made the cell more sensitive to the introduction of bortezomib. Lastly, they conclude that when they reduced the number of proteasome subunits active sites less bortezomib was needed by the cell to halt the proteasome function.

3.3 Conclusions

We used a hierarchically integrated classifier that used Bayesian integration of independent SVMs to infer protein interaction networks for 702 different compounds. These are by no means all the compounds available, but since our prediction

environment was yeast, the number of small molecules that had enough protein annotations to them was small. First, our compendium relied on protein-protein interactions in different interaction types as its data, so our compound-protein networks have underlying interaction-type associations for every compound-protein pair; we can tell which interaction types contribute to which compound-protein interactions, and we illustrated an example of how this information was useful in section 3.2.5. In addition to evaluating our networks using AUCs and hold-out sets in section 3.2.2, we validated some of the highly enriched gene-sets for a few of our compounds with literature curation in section 3.2.3 . We were interested in compounds that were approved drugs and determined that certain drugs clustered together in section 3.2.4. Further, certain drug clusters interacted with proteins that took part of certain biological processes and protein families/domains. We took clustering one step further and presented a heatmap that clustered certain compound classes with certain biological processes or protein families/domains. From this, we were able to infer some correlations between the drug types and the functions they tend to perform.

4 CONCLUSIONS

In this thesis, we shed light on two broad problems in computational biology. We built genomic networks with the purpose of determining (a) functional interactions between a pair of genes and (b) furthering our knowledge of how proteins interact with small molecules, specifically drugs. The large data compendium available for genomic data can often be daunting to use, and we showed that with careful methods, we were able to extract meaningful biological relationships.

In Chapter 2, we studied the broad problem of determining gene function in various biological contexts. Our compendium demonstrates the usefulness of data integration and includes networks that are "global" in the sense that they describe the overall set of functional interactions predicted to occur among *A. thaliana* proteins, independently of plant tissue, developmental stage, or environmental context. However, most networks in this compendium are context-specific: they describe only the functional relationships predicted to occur at a specific time or in a specific tissue. We integrated a compendium of *A. thaliana* genomic data (55 microarray and 5 interaction datasets) using a Bayesian framework to probabilistically weight each experimental dataset according to its relevance in diverse biological areas. The experimental framework for this study integrated gold standards from the Gene Ontology and Plant Ontology with *A. thaliana* data using regularized Bayesian classifiers; and the resulting predicted genome-wide functional networks were evaluated computationally and experimentally. These networks explain how proteins

in this organism behave in different tissues and development stages in addition to different biological processes. With our networks, biological researchers can determine whether a gene or genes of interest behave differently in various development stages or if they are active only in specific parts of the plant.

In Chapter 3, we switched focus to the problem of drug discovery. We first built mechanistic networks using a large number of diverse datasets, including structural docking and binding site conservation data; these networks relied on the structure of an interaction-type hierarchy to correct individual interaction-type protein-protein interaction predictions. The networks predicted functional interactions between proteins while at the same time having inherent pathway-level knowledge. In addition to just predicting interacting protein pairs in different interaction types, we went one step further and used these networks to predict interaction probabilities between proteins and small molecules (simple molecular compounds and drugs). These networks provided a compendium of interactions between 702 compounds (13 drugs) and 5559 yeast proteins. Because the datasets we used to predict compound-protein interactions were mechanistic networks, we were able to pinpoint what interaction types affected each interacting compound-protein pair the most. We showed, through both computational evaluation and literature curation that our compound-protein networks predict interactions well. For each compound, we found the gene-sets that were most enriched; we looked in depth at the enriched gene-sets of four drugs and found literature validation. Two cluster analyses on the compound-protein network provided an insight into (a) how certain biological processes interact with certain

compound classes, as well as how certain families of proteins interact with certain compound classes and (b) how certain biological processes interact with certain drugs, as well as how certain families of proteins interact with certain drugs. Ultimately, our compound-protein networks are a useful step towards drug discovery by revealing interesting underlying connections influencing interactions between drugs and proteins, which may otherwise go unnoticed.

5 REFERENCES

1. EMBL-EBI, *http://www.ebi.ac.uk/sites/ebi.ac.uk/files/shared/documents/EMBL_EBI_ASR_2012_lo-rez.pdf.* Annual Scientific Report, 2012.
2. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Res, 2002. **30**(1): p. 207-10.
3. Gresham, D., M.J. Dunham, and D. Botstein, *Comparing whole genomes using DNA microarrays.* Nat Rev Genet, 2008. **9**(4): p. 291-302.
4. Ponting, C.P. and R.R. Russell, *The natural history of protein domains.* Annu Rev Biophys Biomol Struct, 2002. **31**: p. 45-71.
5. Fields, S. and O. Song, *A novel genetic system to detect protein-protein interactions.* Nature, 1989. **340**(6230): p. 245-6.
6. Yates, J.R., 3rd, *Mass spectrometry. From genomics to proteomics.* Trends Genet, 2000. **16**(1): p. 5-8.
7. Szilagyi, A., et al., *Prediction of physical protein-protein interactions.* Phys Biol, 2005. **2**(2): p. S1-16.
8. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
9. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Res, 2010. **38**(Database issue): p. D211-22.

10. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

11. Mount, D.W., *Using the Basic Local Alignment Search Tool (BLAST).* CSH Protoc, 2007. **2007**: p. pdb top17.

12. Troyanskaya, O.G., et al., *A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae).* Proc Natl Acad Sci U S A, 2003. **100**(14): p. 8348-53.

13. Myers, C.L., et al., *Discovery of biological networks from diverse functional genomic data.* Genome Biol, 2005. **6**(13): p. R114.

14. Lee, I., et al., *A probabilistic functional network of yeast genes.* Science, 2004. **306**(5701): p. 1555-8.

15. Jansen, R., et al., *A Bayesian networks approach for predicting protein-protein interactions from genomic data.* Science, 2003. **302**(5644): p. 449-53.

16. Jaimovich, A., et al., *Towards an integrated protein-protein interaction network: a relational Markov network approach.* J Comput Biol, 2006. **13**(2): p. 145-64.

17. Pavlidis, P., et al., *Learning gene functional classifications from multiple data types.* J Comput Biol, 2002. **9**(2): p. 401-11.

18. Lanckriet, G.R., et al., *Kernel-based data fusion and its application to protein function prediction in yeast.* Pac Symp Biocomput, 2004: p. 300-11.

19. Barutcuoglu, Z., R.E. Schapire, and O.G. Troyanskaya, *Hierarchical multi-label prediction of gene function.* Bioinformatics, 2006. **22**(7): p. 830-6.

20. Myers, C.L. and O.G. Troyanskaya, *Context-sensitive data integration and prediction of biological networks.* Bioinformatics, 2007. **23**(17): p. 2322-30.

21. Park, C.Y., et al., *Simultaneous genome-wide inference of physical, genetic, regulatory, and functional pathway components.* PLoS Comput Biol, 2010. **6**(11): p. e1001009.

22. Yamanishi, Y., et al., *Prediction of drug-target interaction networks from the integration of chemical and genomic spaces.* Bioinformatics, 2008. **24**(13): p. i232-40.

23. Yildirim, M.A., et al., *Drug-target network.* Nat Biotechnol, 2007. **25**(10): p. 1119-26.

24. Schuffenhauer, A., V.J. Gillet, and P. Willett, *Similarity searching in files of three-dimensional chemical structures: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors.* J Chem Inf Comput Sci, 2000. **40**(2): p. 295-307.

25. Nettles, J.H., et al., *Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors.* J Med Chem, 2006. **49**(23): p. 6802-10.

26. Schuffenhauer, A., et al., *Similarity metrics for ligands reflecting the similarity of the target proteins.* J Chem Inf Comput Sci, 2003. **43**(2): p. 391-405.

27. Cheng, F., et al., *Prediction of chemical-protein interactions network with weighted network-based inference method.* PLoS One, 2012. **7**(7): p. e41064.

28. Keiser, M.J., et al., *Predicting new molecular targets for known drugs.* Nature, 2009. **462**(7270): p. 175-81.

29. Xia, X., et al., *Classification of kinase inhibitors using a Bayesian model.* J Med Chem, 2004. **47**(18): p. 4463-70.

30. Paolini, G.V., et al., *Global mapping of pharmacological space.* Nat Biotechnol, 2006. **24**(7): p. 805-15.

31. Sakakibara, Y., et al., *COPICAT: a software system for predicting interactions between proteins and chemical compounds.* Bioinformatics, 2012. **28**(5): p. 745-6.

32. Li, Q. and L. Lai, *Prediction of potential drug targets based on simple sequence properties.* BMC Bioinformatics, 2007. **8**: p. 353.

33. Waring, J.F., et al., *Clustering of hepatotoxins based on mechanism of toxicity using gene expression profiles.* Toxicol Appl Pharmacol, 2001. **175**(1): p. 28-42.

34. Lamb, J., et al., *The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.* Science, 2006. **313**(5795): p. 1929-35.

35. Brown, J.A., et al., *Global analysis of gene function in yeast by quantitative phenotypic profiling.* Mol Syst Biol, 2006. **2**: p. 2006 0001.

36. Mizutani, S., et al., *Relating drug-protein interaction network with drug side effects.* Bioinformatics, 2012. **28**(18): p. i522-i528.

37. Li, J., X. Zhu, and J.Y. Chen, *Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts.* PLoS Comput Biol, 2009. **5**(7): p. e1000450.

38. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.

39. Kuhn, M., et al., *STITCH: interaction networks of chemicals and proteins.* Nucleic Acids Res, 2008. **36**(Database issue): p. D684-8.

40. Meinke, D.W., et al., *Arabidopsis thaliana: a model plant for genome analysis.* Science, 1998. **282**(5389): p. 662, 679-82.

41. *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.* Nature, 2000. **408**(6814): p. 796-815.

42. Murphy, T.M., et al., *Requirement for abasic endonuclease gene homologues in Arabidopsis seed development.* PLoS One, 2009. **4**(1): p. e4297.

43. Drews, G.N., J.L. Bowman, and E.M. Meyerowitz, *Negative regulation of the Arabidopsis homeotic gene AGAMOUS by the APETALA2 product.* Cell, 1991. **65**(6): p. 991-1002.

44. Boyes, D.C., et al., *Growth stage-based phenotypic analysis of Arabidopsis: a model for high throughput functional genomics in plants.* Plant Cell, 2001. **13**(7): p. 1499-510.

45. Avraham, S., et al., *The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations.* Nucleic Acids Res, 2008. **36**(Database issue): p. D449-54.

46. Pop, A., et al., *Integrated functional networks of process, tissue, and developmental stage specific interactions in Arabidopsis thaliana.* BMC Syst Biol, 2010. **4**: p. 180.

47. Myers, C.L., et al., *Finding function: evaluation methods for functional genomic data.* BMC Genomics, 2006. **7**: p. 187.

48. Huttenhower, C., et al., *Exploring the human genome with functional maps.* Genome Res, 2009. **19**(6): p. 1093-106.

49. Barrett, T., et al., *NCBI GEO: archive for high-throughput functional genomic data.* Nucleic Acids Res, 2009. **37**(Database issue): p. D885-90.

50. Willis, R.C. and C.W. Hogue, *Searching, viewing, and visualizing data in the Biomolecular Interaction Network Database (BIND).* Curr Protoc Bioinformatics, 2006. **Chapter 8**: p. Unit 8 9.

51. Stark, C., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Res, 2006. **34**(Database issue): p. D535-9.

52. TAIR. *The Arabidopsis Information Resource (Drought stress time course)*. Available from: http://arabidopsis.org/servlets/TairObject?type=expression_set&id=1007966668.

53. TAIR. *The Arabidopsis Information Resource (Salt stress time course)*. Available from: http://arabidopsis.org/servlets/TairObject?type=expression_set&id=1007966888.

54. TAIR. *The Arabidopsis Information Resource (UV-B stress time course).* Available from: http://arabidopsis.org/servlets/TairObject?type=expression_set&id=1007966606.

55. TAIR. *The Arabidopsis Information Resource (Osmotic stress time course)*. Available from: http://arabidopsis.org/servlets/TairObject?type=expression_set&id=1007966835.

56. TAIR. *The Arabidopsis Information Resource (Cold stress time course)*. Available from: http://arabidopsis.org/servlets/TairObject?type=expression_set&id=1007966553.

57. Boisson, B., C. Giglione, and T. Meinnel, *Unexpected protein families including cell defense components feature in the N-myristoylome of a higher eukaryote.* J Biol Chem, 2003. **278**(44): p. 43418-29.

58. TAIR. *The Arabidopsis Information Resource*. May 2008; Available from: http://www.arabidopsis.org/portals/expression/microarray/ATGenExpress.jsp.

59. Huttenhower, C., et al., *A scalable method for integration and functional analysis of multiple microarray datasets.* Bioinformatics, 2006. **22**(23): p. 2890-7.

60. Malamy, J.E. and P.N. Benfey, *Organization and cell differentiation in lateral roots of Arabidopsis thaliana.* Development, 1997. **124**(1): p. 33-44.

61. Barlow, P., *McManus, M.T. and Veit, B.E. eds. Meristematic tissues in plant growth and development.* Ann Bot, 2002. **90**(4): p. 546-547.

62. Fletcher, J.C., *Shoot and floral meristem maintenance in arabidopsis.* Annu Rev Plant Biol, 2002. **53**: p. 45-66.

63. Dinneny, J.R. and M.F. Yanofsky, *Floral development: an ABC gene chips in downstream.* Curr Biol, 2004. **14**(19): p. R840-1.

64. Chu, L.Y., H.B. Shao, and M.Y. Li, *Molecular mechanisms of phytochrome signal transduction in higher plants.* Colloids Surf B Biointerfaces, 2005. **45**(3-4): p. 154-61.

65. Cho, S.K., et al., *Heterologous expression and molecular and cellular characterization of CaPUB1 encoding a hot pepper U-Box E3 ubiquitin ligase homolog.* Plant Physiol, 2006. **142**(4): p. 1664-82.

66. Gao, L. and C.B. Xiang, *The genetic locus At1g73660 encodes a putative MAPKKK and negatively regulates salt tolerance in Arabidopsis.* Plant Mol Biol, 2008. **67**(1-2): p. 125-34.

67. Weber, A.P., J. Schneidereit, and L.M. Voll, *Using mutants to probe the in vivo function of plastid envelope membrane metabolite transporters.* J Exp Bot, 2004. **55**(400): p. 1231-44.

68. Knappe, S., et al., *Characterization of two functional phosphoenolpyruvate/phosphate translocator (PPT) genes in Arabidopsis--AtPPT1 may be involved in the provision of signals for correct mesophyll development.* Plant J, 2003. **36**(3): p. 411-20.

69. Mackey, D., et al., *RIN4 interacts with Pseudomonas syringae type III effector molecules and is required for RPM1-mediated resistance in Arabidopsis.* Cell, 2002. **108**(6): p. 743-54.

70. Axtell, M.J. and B.J. Staskawicz, *Initiation of RPS2-specified disease resistance in Arabidopsis is coupled to the AvrRpt2-directed elimination of RIN4.* Cell, 2003. **112**(3): p. 369-77.

71. Day, B., D. Dahlbeck, and B.J. Staskawicz, *NDR1 interaction with RIN4 mediates the differential activation of multiple disease resistance pathways in Arabidopsis.* Plant Cell, 2006. **18**(10): p. 2782-91.

72. Ryu, K.H., et al., *The WEREWOLF MYB protein directly regulates CAPRICE transcription during cell fate specification in the Arabidopsis root epidermis.* Development, 2005. **132**(21): p. 4765-75.

73. Bernhardt, C., et al., *The bHLH genes GL3 and EGL3 participate in an intercellular regulatory circuit that controls cell patterning in the Arabidopsis root epidermis.* Development, 2005. **132**(2): p. 291-8.

74. Bernhardt, C., et al., *The bHLH genes GLABRA3 (GL3) and ENHANCER OF GLABRA3 (EGL3) specify epidermal cell fate in the Arabidopsis root.* Development, 2003. **130**(26): p. 6431-9.

75. Levesque, M.P., et al., *Whole-genome analysis of the SHORT-ROOT developmental pathway in Arabidopsis.* PLoS Biol, 2006. **4**(5): p. e143.

76. Welch, D., et al., *Arabidopsis JACKDAW and MAGPIE zinc finger proteins delimit asymmetric cell division and stabilize tissue boundaries by restricting SHORT-ROOT action.* Genes Dev, 2007. **21**(17): p. 2196-204.

77. Kepinski, S. and O. Leyser, *The Arabidopsis F-box protein TIR1 is an auxin receptor.* Nature, 2005. **435**(7041): p. 446-51.

78. Dharmasiri, N., S. Dharmasiri, and M. Estelle, *The F-box protein TIR1 is an auxin receptor.* Nature, 2005. **435**(7041): p. 441-5.

79. Chapman, E.J. and M. Estelle, *Mechanism of auxin-regulated gene expression in plants.* Annu Rev Genet, 2009. **43**: p. 265-85.

80. Mockaitis, K. and M. Estelle, *Auxin receptors and plant development: a new signaling paradigm.* Annu Rev Cell Dev Biol, 2008. **24**: p. 55-80.

81. Hellerstein, M.K., *Exploiting complexity and the robustness of network architecture for drug discovery.* J Pharmacol Exp Ther, 2008. **325**(1): p. 1-9.

82. Chang, R.L., et al., *Drug off-target effects predicted using structural analysis in the context of a metabolic network model.* PLoS Comput Biol, 2010. **6**(9): p. e1000938.

83. Cheng, F., et al., *Prediction of drug-target interactions and drug repositioning via network-based inference.* PLoS Comput Biol, 2012. **8**(5): p. e1002503.

84. Ho, C.H., et al., *Combining functional genomics and chemical biology to identify targets of bioactive compounds.* Curr Opin Chem Biol, 2011. **15**(1): p. 66-78.

85. Fortney, K., et al., *NetwoRx: connecting drugs to networks and phenotypes in Saccharomyces cerevisiae.* Nucleic Acids Res, 2013. **41**(Database issue): p. D720-7.

86. Lu, J.J., et al., *Multi-target drugs: the trend of drug research and development.* PLoS One, 2012. **7**(6): p. e40262.

87. Chen, W.H., et al., *Human monogenic disease genes have frequently functionally redundant paralogs.* PLoS Comput Biol, 2013. **9**(5): p. e1003073.

88. Orchard, S., H. Hermjakob, and R. Apweiler, *Annotating the human proteome.* Mol Cell Proteomics, 2005. **4**(4): p. 435-40.

89. Karathia, H., et al., *Saccharomyces cerevisiae as a model organism: a comparative study.* PLoS One, 2011. **6**(2): p. e16015.

90. Winzeler, E.A., et al., *Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis.* Science, 1999. **285**(5429): p. 901-6.

91. Giaever, G., et al., *Genomic profiling of drug sensitivities via induced haploinsufficiency.* Nat Genet, 1999. **21**(3): p. 278-83.

92. Hillenmeyer, M.E., et al., *The chemical genomic portrait of yeast: uncovering a phenotype for all genes.* Science, 2008. **320**(5874): p. 362-5.

93. Barrientos, A., *Yeast models of human mitochondrial diseases.* IUBMB Life, 2003. **55**(2): p. 83-95.

94. Schwimmer, C., et al., *Yeast models of human mitochondrial diseases: from molecular mechanisms to drug screening.* Biotechnol J, 2006. **1**(3): p. 270-81.

95. Couplan, E., et al., *A yeast-based assay identifies drugs active against human mitochondrial disorders.* Proc Natl Acad Sci U S A, 2011. **108**(29): p. 11989-94.

96. Steinmetz, L.M., et al., *Systematic screen for human disease genes in yeast.* Nat Genet, 2002. **31**(4): p. 400-4.

97. Miller-Fleming, L., F. Giorgini, and T.F. Outeiro, *Yeast as a model for studying human neurodegenerative disorders.* Biotechnol J, 2008. **3**(3): p. 325-38.

98. Ma, D., *Applications of yeast in drug discovery.* Prog Drug Res, 2001. **57**: p. 117-62.

99. Vapnik, V. and C. Cortes, *Support-vector networks.* Machine Learning, 1995. **20**(3): p. 24.

100. Joachims, T., *Training linear SVMs in linear time.* Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006: p. 9.

101. Cherry, J.M., et al., *Saccharomyces Genome Database: the genomics resource of budding yeast.* Nucleic Acids Res, 2012. **40**(Database issue): p. D700-5.

102. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.

103. Zhang, Q.C., et al., *Protein interface conservation across structure space.* Proc Natl Acad Sci U S A, 2010. **107**(24): p. 10896-901.

104. Wei, B.Q., et al., *Testing a flexible-receptor docking algorithm in a model binding site.* J Mol Biol, 2004. **337**(5): p. 1161-82.

105. Ritchie, D.W., *Evaluation of protein docking predictions using Hex 3.1 in CAPRI rounds 1 and 2.* Proteins, 2003. **52**(1): p. 98-106.

106. PDB, *http://www.rcsb.org/pdb/.* 2011.

107. Berman, H., K. Henrick, and H. Nakamura, *Announcing the worldwide Protein Data Bank.* Nat Struct Biol, 2003. **10**(12): p. 980.

108. Huttenhower, C., et al., *The Sleipnir library for computational functional genomics.* Bioinformatics, 2008. **24**(13): p. 1559-61.

109. Jirapech-Umpai, T. and S. Aitken, *Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes.* BMC Bioinformatics, 2005. **6**: p. 148.

110. Ding, C. and H. Peng, *Minimum redundancy feature selection from microarray gene expression data.* J Bioinform Comput Biol, 2005. **3**(2): p. 185-205.

111. Gist, *http://www.chibi.ubc.ca/gist/.* 2002.

112. Neapolitan, R.E., *Learning Bayesian Networks.* Prentice Hall, 2004.

113. Guan, Y., et al., *Functional genomics complements quantitative genetics in identifying disease-gene associations.* PLoS Comput Biol, 2010. **6**(11): p. e1000991.

114. Knox, C., et al., *DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.* Nucleic Acids Res, 2011. **39**(Database issue): p. D1035-41.

115. PubMed, *http://www.ncbi.nlm.nih.gov/pubmed.* 2012.

116. Whirl-Carrillo, M., et al., *Pharmacogenomics knowledge for personalized medicine.* Clin Pharmacol Ther, 2012. **92**(4): p. 414-7.

117. Park, C.Y., et al., *Functional knowledge transfer for high-accuracy prediction of under-studied biological processes.* PLoS Comput Biol, 2013. **9**(3): p. e1002957.

118. Chikina, M.D. and O.G. Troyanskaya, *Accurate quantification of functional analogy among close homologs.* PLoS Comput Biol, 2011. **7**(2): p. e1001074.

119. Kim, S.Y. and D.J. Volsky, *PAGE: parametric analysis of gene set enrichment.* BMC Bioinformatics, 2005. **6**: p. 144.

120. Gonzalez, I., et al., *An R Package to extend Canonical Correlation Analysis.* Journal of Statistical Software, 2008. **23**(12).

121. Koonin, E.V. and M.Y. Galperin, *Evolutionary Concept in Genetics and Genomics*, in *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*. 2003, Kulwer Academic.

122. Galperin, M.Y., D.R. Walker, and E.V. Koonin, *Analogous enzymes: independent inventions in enzyme evolution.* Genome Res, 1998. **8**(8): p. 779-90.

123. von Mering, C., et al., *STRING: a database of predicted functional associations between proteins.* Nucleic Acids Res, 2003. **31**(1): p. 258-61.

124. Hunter, S., et al., *InterPro in 2011: new developments in the family and domain prediction database.* Nucleic Acids Res, 2012. **40**(Database issue): p. D306-12.

125. Anand, P. and P.A. Gruppuso, *Rapamycin inhibits liver growth during refeeding in rats via control of ribosomal protein translation but not cap-dependent translation initiation.* J Nutr, 2006. **136**(1): p. 27-33.

126. Inoki, K., et al., *Signaling by target of rapamycin proteins in cell growth control.* Microbiol Mol Biol Rev, 2005. **69**(1): p. 79-100.

127. Wu, C.H., et al., *Proteinase-3, a serine protease which mediates doxorubicin-induced apoptosis in the HL-60 leukemia cell line, is downregulated in its doxorubicin-resistant variant.* Oncogene, 2002. **21**(33): p. 5160-74.

128. Bednarski, B.K., A.S. Baldwin, Jr., and H.J. Kim, *Addressing reported pro-apoptotic functions of NF-kappaB: targeted inhibition of canonical NF-*

kappaB enhances the apoptotic effects of doxorubicin. PLoS One, 2009. **4**(9): p. e6992.

129. Environmental-Protection-Agency, *Pesticide Fact Sheet.* 2003.

130. InterPro, *http://www.ebi.ac.uk/interpro/entry/IPR020846.* 2013.

131. Kiyotani, K., et al., *Pharmacogenomics of tamoxifen: roles of drug metabolizing enzymes and transporters.* Drug Metab Pharmacokinet, 2012. **27**(1): p. 122-31.

132. Sakaeda, T., et al., *Simvastatin and lovastatin, but not pravastatin, interact with MDR1.* J Pharm Pharmacol, 2002. **54**(3): p. 419-23.

133. Tufan, A., et al., *Association of drug transporter gene ABCB1 (MDR1) 3435C to T polymorphism with colchicine response in familial Mediterranean fever.* J Rheumatol, 2007. **34**(7): p. 1540-4.

134. Pawarode, A., et al., *Differential effects of the immunosuppressive agents cyclosporin A, tacrolimus and sirolimus on drug transport by multidrug resistance proteins.* Cancer Chemother Pharmacol, 2007. **60**(2): p. 179-88.

135. Momparler, R.L., et al., *Effect of adriamycin on DNA, RNA, and protein synthesis in cell-free systems and intact cells.* Cancer Res, 1976. **36**(8): p. 2891-5.

136. Sazuka, Y., H. Tanizawa, and Y. Takino, *Effect of adriamycin on DNA, RNA and protein biosyntheses in mouse tissues, in connection with its cardiotoxicity.* Jpn J Cancer Res, 1989. **80**(10): p. 1000-5.

137. Fornari, F.A., et al., *Interference by doxorubicin with DNA unwinding in MCF-7 breast tumor cells.* Mol Pharmacol, 1994. **45**(4): p. 649-56.

138. Muslimovic, A., et al., *Numerical analysis of etoposide induced DNA breaks.* PLoS One, 2009. **4**(6): p. e5859.

139. Limson, M.V. and K.S. Sweder, *Rapamycin inhibits yeast nucleotide excision repair independently of tor kinases.* Toxicol Sci, 2010. **113**(1): p. 77-84.

140. Powers, T. and P. Walter, *Regulation of ribosome biogenesis by the rapamycin-sensitive TOR-signaling pathway in Saccharomyces cerevisiae.* Mol Biol Cell, 1999. **10**(4): p. 987-1000.

141. Pestov, D.G. and N. Shcherbik, *Rapid cytoplasmic turnover of yeast ribosomes in response to rapamycin inhibition of TOR.* Mol Cell Biol, 2012. **32**(11): p. 2135-44.

142. Bolzan, A.D. and M.S. Bianchi, *Genotoxicity of streptozotocin.* Mutat Res, 2002. **512**(2-3): p. 121-34.

143. Wozniak, K., et al., *The DNA-damaging potential of tamoxifen in breast cancer and normal cells.* Arch Toxicol, 2007. **81**(7): p. 519-27.

144. Weniger, M.A., et al., *Treatment-induced oxidative stress and cellular antioxidant capacity determine response to bortezomib in mantle cell lymphoma.* Clin Cancer Res, 2011. **17**(15): p. 5101-12.

145. Wilck, N., et al., *Attenuation of early atherogenesis in low-density lipoprotein receptor-deficient mice by proteasome inhibition.* Arterioscler Thromb Vasc Biol, 2012. **32**(6): p. 1418-26.

146. Schnabl, M., G. Daum, and H. Pichler, *Multiple lipid transport pathways to the plasma membrane in yeast.* Biochim Biophys Acta, 2005. **1687**(1-3): p. 130-40.

147. Dumlao, D.S., N. Hertz, and S. Clarke, *Secreted 3-isopropylmalate methyl ester signals invasive growth during amino acid starvation in Saccharomyces cerevisiae.* Biochemistry, 2008. **47**(2): p. 698-709.

148. Purevdorj-Gage, B., et al., *The role of FLO11 in Saccharomyces cerevisiae biofilm development in a laboratory based flow-cell system.* FEMS Yeast Res, 2007. **7**(3): p. 372-9.

149. Camarasa, C., et al., *Phenotypic landscape of Saccharomyces cerevisiae during wine fermentation: evidence for origin-dependent metabolic traits.* PLoS One, 2011. **6**(9): p. e25147.

150. Gutierrez, A., et al., *Biomarkers for detecting nitrogen deficiency during alcoholic fermentation in different commercial wine yeast strains.* Food Microbiol, 2013. **34**(1): p. 227-37.

151. Surjana, D., G.M. Halliday, and D.L. Damian, *Role of nicotinamide in DNA damage, mutagenesis, and DNA repair.* J Nucleic Acids, 2010. **2010**.

152. Kruman, II, et al., *Folic acid deficiency and homocysteine impair DNA repair in hippocampal neurons and sensitize them to amyloid toxicity in experimental models of Alzheimer's disease.* J Neurosci, 2002. **22**(5): p. 1752-62.

153. Smith, M.L. and M.A. Kumar, *Seleno-L-Methionine Modulation of Nucleotide Excision DNA Repair Relevant to Cancer Prevention and Chemotherapy.* Mol Cell Pharmacol, 2009. **1**(4): p. 218-221.

154. Chen, D., et al., *Bortezomib as the first proteasome inhibitor anticancer drug: current status and future perspectives.* Curr Cancer Drug Targets, 2011. **11**(3): p. 239-53.

155. Chen, S., et al., *Genome-wide siRNA screen for modulators of cell death induced by proteasome inhibitor bortezomib.* Cancer Res, 2010. **70**(11): p. 4318-26.