IDENTIFYING AND MODELING PERTURBED NETWORKS IN CANCER THROUGH STATISTICAL AND CONSTRAINT-BASED ANALYSIS

BY

JAMES ALLAN EDDY

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Bioengineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Doctoral Committee:

Associate Professor Nathan D. Price, Chair
Professor Michael Insana
Professor Gene Robinson
Associate Professor Sheng Zhong
Assistant Professor Jian Ma

Abstract

I was introduced to systems and computational biology as an undergraduate at the University of Virginia. In pursuing projects both for independent and thesis research, I was fascinated by the potential of mathematics and computers to describe complex phenomena in biology and to help address challenges otherwise intractable with traditional experimental methods. Continuing in this field for my graduate research allowed me to focus more heavily on problems in human health and disease, while building a diverse skill set that can also be applied to studying basic biology or even microbes for industrial applications. In my Ph.D. research, I have been a part of cutting-edge computational and systems biology endeavors with Nathan Price at the University of Illinois and, most recently, the Institute for Systems Biology. My work has focused largely on modeling biological networks—using both statistical and mechanistic modeling approaches—with applications to medicine, energy, and genomic science.

The more coarse-grained of these approaches use statistical analysis of high-throughput expression data to identify molecular signatures of disease phenotypes; such signatures are indicative of aberrant function of genes or pathways. As a primary focus of my Ph.D. work, I have generated and applied tools to harness high-throughput transcriptomic data to identify molecular signatures—particularly a special class of signatures that quantifies the *relative* expression levels among multiple genes—and to help elucidate underlying mechanism of complex biological processes and properties. I have used both gene- and network-based signatures to pioneer approaches for predicting clinical phenotypes and outcomes of patient samples. Most notably, I led the development and implementation of a computational method for identifying biomolecular pathways (e.g., metabolic or signaling) that are perturbed in disease data. I used this method (described in **Chapter 4**) to identify differentially regulated and variably expressed pathways in a number of human diseases, and later investigated behavior-associated modules in a transcriptional regulatory network of the honeybee brain.

As a more functionally grounded complement to statistical approaches, stoichiometric models of biochemical reaction networks can be used to simulate disease systems in mechanistic detail. I have developed an extensive background in constraint-based analysis of biochemical reaction networks, beginning with my contribution to the genome-scale reconstruction of the pathogen *Leishmania major,* while working as an undergraduate researcher. Building off of the metabolic and modeling expertise I gained through this project—as well as the construction of major signaling pathways in yeast—I was able to help substantially in the metabolic reconstruction of the butanol producing bacteria *Clostridium beijerinckii* when I began my Ph.D. at the University of Illinois. This model is now being used to identify genetic modifications that may lead to increased butanol production for industrial bioenergy applications.

Most recently, I have helped to pioneer and automate novel network reconstruction and modeling approaches in human tissues and cell types. In particular, I have applied these methods to reconstruct and validate the first genome-scale metabolic model for the deadly brain cancer, glioblastoma multiforme. This model and others like it will be transformative for mapping the genome-to-phenotype relationship in complex diseases to elucidate underlying mechanisms and pave the way for new therapeutic strategies. Moreover, the reconstruction and analysis of this model represents a culmination of the array of mechanistic and statistical approaches used and described over the course of my Ph.D. Perturbed metabolic pathways identified by statistical tools represent robust differences between healthy and diseased states, and in turn, serve as focal points for ongoing model simulation and development.

Acknowledgements

I am writing this section as the last step before depositing this dissertation. While my tenure in grad school has been full of ups and downs, numerous people have been an integral part of the process, and so I would like to end on a positive note by thanking them. In the past (nearly) six years, I have grown immensely as both a researcher and a person; without so many stellar individuals shaping and influencing my experiences, I believe this growth would have been more modest and in less satisfying directions. In addition to a hard-earned degree, I'm extremely fortunate to conclude my graduate experience having gained a wealth of great relationships, many of which I expect to last a lifetime.

I would first like to thank the faculty with whom I have worked and interacted for the guidance, insights, and opportunities they have provided. Above all, I will always be extremely grateful to my advisor, Nathan Price. Nathan went out on a limb when he agreed to take me as his first student, based only on a short phone conversation and the recommendation of his good friend and my undergrad research advisor, Jason Papin; without that option, I most likely would not have been accepted to any graduate programs immediately after graduating UVA, and my plans would have been dramatically altered. More importantly, my time and experiences in Nathan's lab have allowed me to learn and thrive more than I could have ever expected. While Nathan's more hands-off approach often meant that I operated and progressed more independently in my research, his big-picture and ambitious thinking helped to drive a number of projects along the way. Moreover, as I've come closer to finishing my Ph.D., Nathan has provided invaluable career advice. Finally, Nathan's decision to move from the University of Illinois to ISB—while surprising and challenging at first—ultimately brought me to Seattle, a city where I truly feel at home and plan to stay.

Besides Nathan, I've had the pleasure of working with several other kind and intelligent professors, including Don Geman, Wei Zhang, and Gene Robinson. Through each of these collaborations, I learned a great deal and gained fresh perspective on new topics. I would also like to thank the remaining members of my committee, Sheng Zhong, Jian Ma, and Michael Insana; from their respective courses in the Bioengineering department to their insightful questions and feedback during my preliminary and final exams, I've enjoyed and appreciated what they've each had to share. Of course, I'd like to especially thank anyone who has ever written a recommendation or served as a reference for a job or fellowship.

My time in the Price lab has been made immeasurably more enjoyable and enriching thanks to many of the quality students and post-docs with whom I've worked. It's tough for me to name individuals, as I've really been fortunate to have spent time with so many good people (and great minds); moreover, I was always happy to be part of what I felt was a respectful and conflict-free environment. I do need to say a

special thanks to Caroline Milne, who has been my best buddy and scientific accomplice since the early days in IGB—through long nights in lab, countless frustrations with projects, exciting breakthroughs, and everyday goofing around, she's played a huge role, not only in keeping me sane, but in pushing me to work my hardest, think critically, and become the best researcher (and person) I could. Since coming to ISB, I've also had the distinct pleasure of meeting, working with, having countless great conversations (about science and life in general) with, and becoming good friends with Julie Bletz and Ben Heavner.

Maybe the most important bunch to thank of all is my closest friends, five people who I consider to be my second family: Brad Novak, Caroline (see above), Michael Rasche, and Ana Martinez. These amazing people have been with me through it all, and I've shared some of the best times of my life (so far) with them. It would take way too much space to even summarize all the reasons why I admire and value these friends so much, but I'm sure they get it. I also can't leave out my wonderful pup, Bean—he's been the best friend and companion anyone could hope for, and I'm much better off for having him around (he can't read this, but I'll make sure he gets a real good treat instead). As much as I've enjoyed my time in Seattle, I owe a special thanks to Aaron Brooks, who has been an outstanding partner for climbing, music, and beer adventures of all sorts. Finally, speaking of adventures, I want to thank Alison Driver for the incredible times we've had in the months leading up to the conclusion of my Ph.D. and also for being so patient and supportive in the process.

Last, but not least, is my family, who have been truly loving and supportive for (literally) as long as I can remember. My parents, Allan and Pam, along with my siblings, Raechel, Laura, and Chris, and my big sister, Sarah, have always rooted for me harder than anyone, and I've rarely felt more proud of myself than on the occasions where I make them proud. It may be cliché, but I would not be where I am today without them. My extended family has also been great, and my grandmothers especially took such a strong interest in my work—I wish Helen Eddy (my dad's mother) could have been here to see me reach this milestone, and on a more positive note, I really cannot thank Betty Livengood (my mom's mom) for the countless letters and gifts she's sent over the years.

There are many individuals I've left out here—family members, friends in Illinois and Seattle, all of my great buddies from Virginia, teachers, department staff, and almost certainly others. I'm grateful to everyone for the part they played, big or small, in helping me to get through grad school, and for making my life a little bit better in some way or another. Thank you!!

Table of Contents

Chapter 1. Introduction and Overview[1]

Monumental advances in molecular and cellular biology—beginning in the latter half the 20th century and continuing today—have provided an increasingly detailed portrait of human biology from the molecular to physiological levels. These advances have centered on "reductionist" experimental approaches aiming to annotate a vast array of biological components, from cells and tissues to genes and proteins. Collectively, these components represent a "parts list" for biological systems (e.g., biochemical pathways, larger interaction networks). At scales beyond a handful of interacting components, however, simple analysis techniques can become limited in providing comprehensible insight into resulting phenotypic behaviors. Systems biology is a rapidly growing discipline that employs an integrative approach to characterize biological systems, in which interactions among all components in a system are described mathematically to establish a computable model. These *in silico* models—which complement traditional *in vivo* animal models—can be simulated to quantitatively study the emergent behavior of a system of interacting components. Model development in the systems biology paradigm is enabled by the description of parts and interactions from reductionist biology, and also depends upon quantitative measurements. The advent of high-throughput experimental tools has allowed for the simultaneous measurement of thousands of biomolecules, paving the way for *in silico* model construction of increasingly large and diverse biological systems. Integrating heterogeneous dynamic data into quantitative predictive models holds great promise to significantly increase our ability to understand and rationally intervene in disease-perturbed biological systems. This promise—particularly with regards to personalized medicine and medical intervention—has motivated the development of new methods for systems analysis of human biology and disease.

Herein, I present new systems biology tools and methodologies to explore perturbed molecular networks in disease. These methods collectively aim to identify informative changes that give rise to disease, which in turn could be leveraged to improve clinical diagnosis or even guide the development of novel therapeutics. Importantly, my work has included both statistical and mechanistic approaches to modeling large, often genome-scale biological information or data. The predominant focus, both in the development and application of the tools described here has been human cancers; among these, the brain cancer glioblastoma multiforme is the central focus of metabolic modeling efforts. The goal of this chapter is to provide a high-level introduction to the central topics and themes, around which the majority of the work

[1] This chapter includes material that was reproduced with permission from the following publication: Edelman, L.B., J.A. Eddy, and N.D. Price. 2010. In silico models of cancer. Wiley Interdisciplinary Reviews: Systems Biology and Medicine. 2(4):438-459. (**Cancer Complexity & Systems Biology** section; text was collaboratively written with Luke Edelman).

described in this dissertation is based. Moreover, I have provided an overview of the chapters that follow to serve as a sort of roadmap for interested readers.

Cancer Complexity & Systems Biology

Cancer is an intrinsically complex and heterogeneous disease, making it particularly amenable to systems biology approaches. Malignant tumors develop as a function of multiple biological interactions and events, both in the molecular domain among individual genes and proteins, and at the cellular and physiological levels between functionally diverse somatic cells and tissues [1] (**Figure 1.1**). At the molecular level, genetic lesions interact synergistically to evade tumor suppression pathways, with no single mutation typically sufficient to cause transformation [2-6]. Beyond genetic mutations, transformed cells can exhibit changes in expression of hundreds to thousands of genes and proteins [7-9]. Genetic modifications observed in cancer are often accompanied by changes at the epigenetic level [10-15]. The convolution of genetic effects and epigenetic modifications illustrates the complex, nonlinear relationship between molecular state and cellular cancer phenotype, emphasizing the need for heterogeneous data integration through *in silico* models. The diversity of cancer models mirrors the broad array of molecular and physiological events characteristic of the disease (**Figure 1.2**). The most course-grained approaches use statistical analysis of high-throughput expression data to identify molecular signatures of cancer phenotypes. Such signatures are indicative of aberrant function of genes or pathways, and can be used to predict the type, stage, or grade of biopsied tumor samples. More advanced methods aim to statistically infer the structure and/or quantitative relationships among biomolecules within interaction and regulatory networks of importance in cancer. Alternatively, stoichiometric or kinetic models of biochemical reaction networks—constructed in a bottom-up, annotation-based manner—can be used to simulate in mechanistic detail the behavior of metabolism or signal transduction in cancer.

The complexity of intracellular phenomena observed in cancer is mirrored by equally intricate interactions between cells and across somatic tissues. Among the most important biological systems mediating cancer development is the local tumor microenvironment—a complex, interacting system of cells and extracellular moieties. Contributory agents include the extracellular matrix [16], cooperating tumor and proximate "host" cells [17-20], extracellular signaling factors [21-23], and the metabolic context of local tissue [24, 25]. Other important agents include the infiltrating leukocytes and cytokines of the immune system [26-28]. Human cancers also exhibit other major interactions with somatic tissues concomitant to malignant invasion, such as tumor-induced angiogenesis [29-31]. The potential response to chemotherapeutics, radiotherapy, and surgical procedures represent additional confounding factors in the cellular and physiological behavior of cancer cells [32-35]. The heterogeneous nature of the tumor

microenvironment poses substantial modeling challenges, yet ongoing research has sought to characterize these cancer systems, including continuum and discrete models.

Despite significant experimental and analytic challenges arising from cancer's complexity, modeling has already successfully led to insights into cancer biology and treatment, as will be discussed in **Chapter 2**. Some of the earliest models describing the molecular basis of cancer over half a century ago implicated the absolute number of genetic mutations as causative for malignancy [36]. Today, important efforts in sequencing the human genome [37, 38] and now individual cancers [39, 40] mean that malignant genetic transformations can be studied and modeled in the context of the entire genome.

Biological Networks & Cancer Metabolism

Reconstructing detailed *in silico* models of biochemical reaction networks (e.g., metabolic, signaling, regulatory) at the genome scale establishes a platform on which different genetic perturbations can be related to emergent malignant functions and phenotypes. Metabolism, in particular, is arguably the best understood cellular process and is highly perturbed in oncogenesis, where cancer cells have increased metabolic rate to provide the energy needed for increased proliferation [41-43]. Metabolite-based analyses to probe cancer have been performed since the 1980s [44] and have shown that cancer cells display distinct metabolic profiles, which can be characterized to diagnose the type and progression of disease, inform prognosis, and assess efficacy of therapy [45]. Metabolic phenotypes that remain consistent across cancer types reflect decreased aerobic respiration activity [46, 47] coupled with increased glycolysis [45, 46] and increased phospholipid production [46, 48]. These observations have led to targeted diagnosis and treatment strategies [45, 49]. Based on past advances resulting from metabolite analysis in cancer cells, accurate profiling of the cancer cell metabolome—and integrating this data into cell-scale metabolic models—is of high interest. Doing so should not only further elucidate the pathophysiolgy of cancer, but has potential to improve methods of diagnosis and treatment.

The most basic mathematical representation of a biochemical reaction network is a stoichiometric model, which describes the interconversion of biomolecules purely in terms of the number of reactants and products participating in each reaction. The generation of stoichiometric models and analysis of their properties is a well-established process [50-52], and genome-scale models of metabolism have been completed for a diverse range of organisms, including the prokaryote *E. coli* [53], archaea [54], and eukaryotes such as *S. cerevisiae* [55] and the protozoan *L. major* [56]. The reconstruction of a biochemical reaction network results in a database of stoichiometric equations that can be represented mathematically to form the foundation of a genome-scale, computable model. Computational tools for

constraint-based analysis (**Figure 1.3**) are then used to interrogate the properties of the reconstructed network *in silico* and to facilitate model-driven validation and refinement [57]. That is, physico-chemical and environmental constraints under which the network operates are applied in the form of *balances*, including mass, energy, and charge, and *bounds*, such as flux capacities and thermodynamic constraints [57, 58]. The statement of constraints defines a solution space comprising all non-excluded network states, thereby describing possible functions or allowable phenotypes. These methods are now being adapted for modeling human systems in greater detail. Genome-scale models have been used for applications ranging from biofuel production to drug target discovery [59, 60].

Dissertation Aim & Organization

The overarching goals of the work presented in this dissertation were to (i) develop computational tools and frameworks to quantify molecular changes in biological networks that are either causal for or reflective of disease; and (ii) explore the application of these tools—particularly those related to network expression analysis or mechanistic modeling of cellular metabolism—to identifying perturbed systems in cancer.

Throughout my Ph.D., I have had many opportunities to interact and collaborate with other scientists—including students, post-doctoral fellows, and faculty; this has both enriched my own projects and allowed me to contribute in both intellectual and hands-on facets to the work of others. As such, the voice in subsequent chapters is sometimes more fluid to best indicate the origin of methods, analyses, results, and the text describing them. As much as possible, I have used first-person singular to clarify the projects or tasks for which I was directly responsible. In some cases, I use "we" when referring to more collaborative efforts; and of course, I gladly cite my colleagues when describing their specific contributions.

The chapters in this dissertation can be briefly summarized as follows:

Chapter 1 introduces the central themes and topics for the work described in subsequent chapters.

Chapter 2 provides an introduction to and extensive overview of computational and systems biology approaches to modeling and investigating cancer.

Chapter 3 describes a particular class of statistical models, Relative Expression Analysis algorithms, and reviews their potential as well as previously demonstrated applications in aiding cancer diagnosis and prognosis.

Chapter 4 presents a novel method, Differential Rank Conservation (DIRAC), which I developed to quantify patterns of gene expression in disease on the level of biological networks.

Chapter 5 highlights several other applications of DIRAC (including in the honeybee and mouse brain), extensions of the method, and alternative gene- and network-based expression analysis approaches.

Chapter 6 describes the construction and analysis of a genome-scale metabolic network for the glioblastoma multiforme cell line, U87 MG, a more mechanistic approach to network analysis compared to the statistical tools described in **Chapters 3, 4,** and **5**.

Chapter 7 offers a summary and conclusion for the ideas and work presented here as well as a prospective contemplation about ongoing challenges and future directions of my work and the field of systems biology.

Chapter 1 Figures

Figure 1.1. Molecular and physiological complexity in cancer.

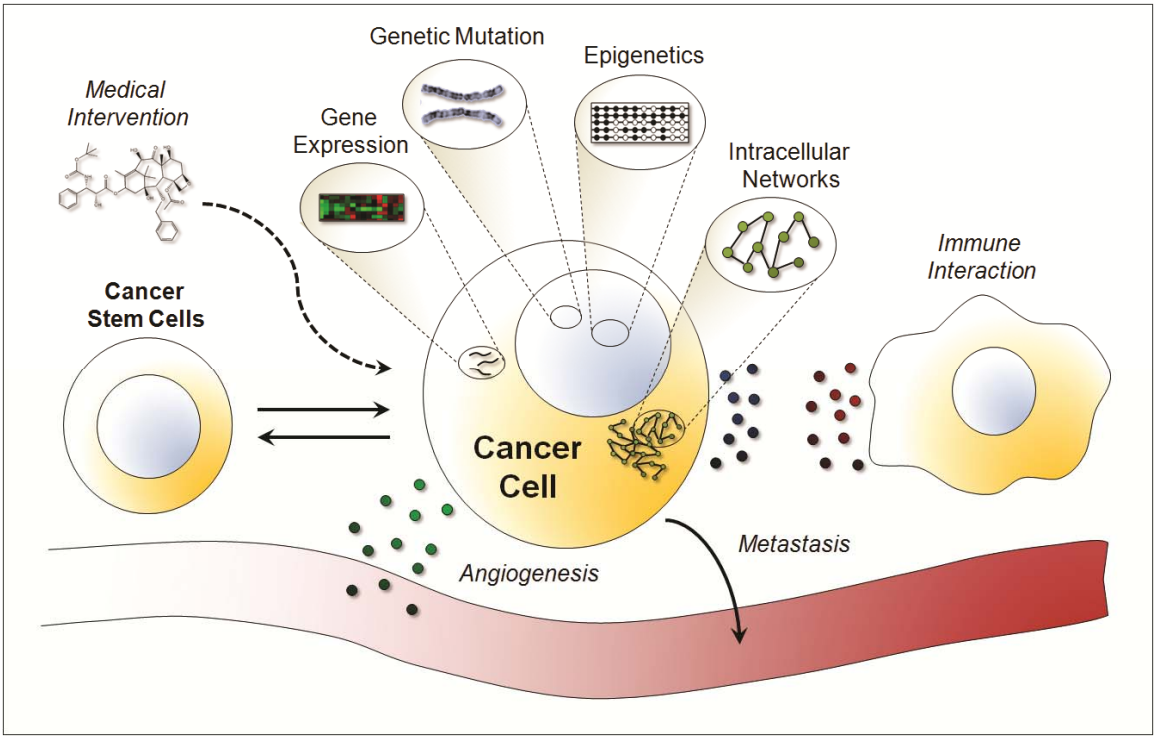


Figure 1.2. Biological scales and potential modeling approaches.

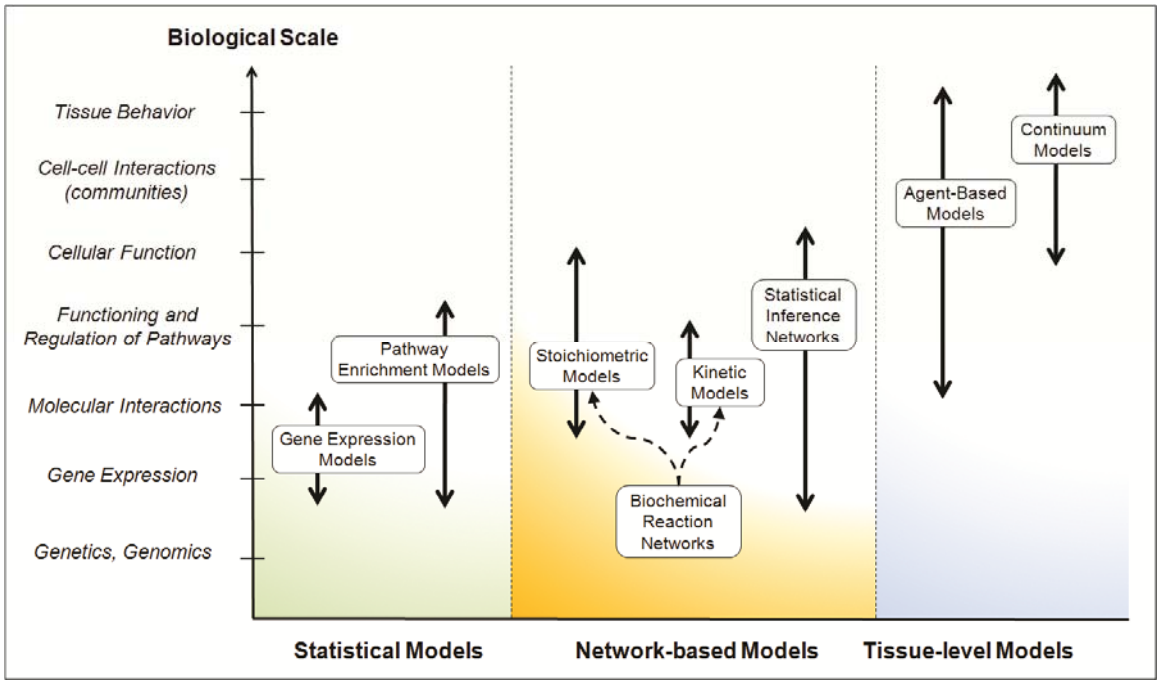
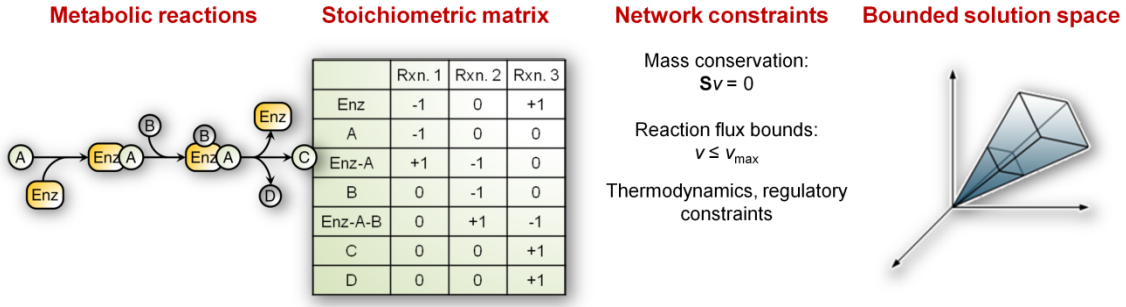


Figure 1.3. Schematic overview of constraint-based modeling.



Chapter 2. *In silico* models of cancer[2]

Cancer is a complex disease that involves multiple types of biological interactions across diverse physical, temporal, and biological scales. This complexity presents substantial challenges for the characterization of cancer biology and motivates the study of cancer in the context of molecular, cellular, and physiological systems. Computational models of cancer are being developed to aid both biological discovery and clinical medicine. The development of these *in silico* models is facilitated by rapidly advancing experimental and analytical tools that generate information-rich, high-throughput biological data. Statistical models of cancer at the genomic, transcriptomic, and pathway levels have proven effective in developing diagnostic and prognostic molecular signatures, as well as in identifying perturbed pathways. Statistically inferred network models can prove useful in settings where data overfitting can be avoided and provide an important means for biological discovery. Mechanistically-based signaling and metabolic models that apply *a priori* knowledge of biochemical processes derived from experiments can also be reconstructed where data are available and can provide insight and predictive ability regarding the dynamical behavior of these systems. At longer length scales, continuum and agent-based models of the tumor microenvironment and other tissue-level interactions enable modeling of cancer cell populations and tumor progression. Even though cancer has been among the most-studied human diseases using systems approaches, significant challenges remain before the enormous potential of *in silico* cancer biology can be fully realized.

In this chapter, I describe key examples of recent *in silico* modeling efforts in cancer. These include (i) statistical models of cancer, such as molecular signatures of perturbed genes and molecular pathways, and statistically inferred reaction networks; (ii) models that represent biochemical, metabolic, and signaling reaction networks important in oncogenesis, including constraint-based and dynamic approaches for the reconstruction of such networks; and (iii) continuum and agent-based models of the tumor microenvironment and tissue-level interactions. Finally, I examine the direction of ongoing research in cancer systems biology, and discuss opportunities for fundamental biological insights and their potential application to clinical practice.

Enabling Tools and Technology in Systems Cancer Modeling

With the initial sequencing of the Human Genome completed at the turn of the millennium, biological inquiry has been transformed by the ability to examine diverse cellular phenotypes such as cancer at the

[2] This chapter includes material that was reproduced with permission from the following publication: Edelman, L.B., J.A. Eddy, and N.D. Price. 2010. In silico models of cancer. Wiley Interdisciplinary Reviews: Systems Biology and Medicine. 2(4):438-459. (all sections; text was collaboratively written with Luke Edelman).

genome scale. Experimental tools for measurement of the cancer genome, transcriptome, and proteome are being continuously improved, including microarray platforms [61-63], next-generation sequencing technologies [64, 65], and advanced mass spectrometry [66-68]. These tools for the high-throughput interrogation of cancer biology have developed in conjunction with analytical and computer-based tools for integration and annotation of the data generated. Together, these parallel advancements are enabling increasingly sophisticated models of cancer at the molecular and cellular levels.

Software and Database Resources for Cancer Systems Biology

With the rapid advancement of technologies for high-throughput biological measurement, the construction of tools for the distribution, integration, and assessment of this raw information has become increasingly important [69]. Multiple online databases (**Table 2.1**) have been created for the storage and distribution of genome-scale data, including transcriptomics [70, 71], regulatory sequences [72, 73], and proteomics [74]. Importantly, these resources provide easily accessible high-throughput data by which cancer models can be constructed and validated. This access enables the global study of disease in a manner not previously possible and will continue to be transformative for systems biology and systems medicine. Models of cancer at different scales of size and complexity often employ *a priori* biological information as constraints to impose parsimony on model architecture. Other web-based services provide annotation of both individual genetic elements [75-79] and their collective associations within distinct biochemical pathways or functional modules [80-85]. Additionally, published mathematical models and required parameters (e.g., reaction kinetics) can be obtained for diverse biological systems [86-88]. Among the major challenges for both statistical and biochemical cancer modeling is knowledge of protein interactions [89]. Multiple public repositories of human and cancer protein interaction networks are currently available based on yeast-two-hybrid data and other experimental methods [90-93]. Additionally, standardized markup languages for biochemical and signaling networks [94-97], as well as software packages [98, 99] have been developed to enable rapid porting of diverse datasets, annotations, and models between different programs and formats. The Systems Biology Markup Language (SBML) in particular is a machine-readable format for representing models of biological processes that has been evolving over the past several years [96]. SBML is becoming more widely adopted, both in the reporting and distribution of models, and in the import and export capabilities of software toolboxes.

The application of advanced analytical techniques to the examination of high-throughput experimental data will continue to facilitate assembly of cancer models with increasing biological precision. However, several practical challenges, which impede the successful integration of data acquired from different experimental sources, or analyzed through different analytical methods, remain to be addressed. For

example, a recent report found that curated protein interaction datasets integrating multiple literature sources can be highly error-prone [100]. The fusion of data across multiple studies, performed at different times or by different laboratory groups, poses a substantial challenge for systems-level analyses that rely so strongly upon large and information-rich datasets. The Minimum Information for Biological and Biomedical Investigations (MIBBI) project is an example of ongoing efforts focused on developing and sharing "minimum information checklists" for experimental data. Such checklists place more stringent guidelines on the quality and content of published data [101]. Many journals and funding agencies already require authors reporting microarray data to comply with the Minimum information About a Microarray Experiment (MIAME) checklist [102]. Other minimal information guidelines which are gaining support in their respective communities include Minimum Information About a Proteomics Experiment (MIAPE) [103], Core Information for Metabolomics Reporting (CIMR), and Minimal Information About a RNAi Experiment (MIARE). The ability to estimate and compare the quality of different data integration schemes in the context of protein interactomics has also been the subject of recent investigation [104, 105]. As these challenges continue to be addressed going forward, they promise the development of ever more powerful and biologically accurate *in silico* representations of human cancer.

Statistically Derived Models of Cancer

Statistical models of cancer can be broadly divided into those that employ unbiased statistical inference, and those that also incorporate *a priori* constraints of specific biological interactions from data (**Figure 2.3**). Statistical models of cancer biology at the genetic, chromosomal, transcriptomic, and pathway levels provide insight about molecular etiology and consequences of malignant transformation despite incomplete knowledge of underlying biological interactions. These methods can help to elucidate key biomolecular events and pathways in oncogenic processes. They thus represent an important paradigm for both the fundamental characterization of cancer systems and the discovery of molecular targets for diagnostic or therapeutic applications. Several studies have sought to infer the structure of small and large-scale biomolecular networks in human cells. These methods can describe previously known biological pathways and can also be used to generate novel hypotheses about potential unknown interactions, which can then be studied more thoroughly through experiments. However, multiple experimental and analytical challenges still remain in the statistical study of cellular and physiologic cancer systems, which must be overcome to yield increasingly useful insights into cancer biology.

Gene Expression Models for Cancer Diagnosis, Stratification, and Prognosis

The discrete mutational events found in the cancer genome and epigenome substantially modulate transcriptional profiles within cancer cells. Models based on these perturbed gene expression signatures can be applied for the diagnosis of disease subtypes as well as for stratification of tumor grade or prediction of therapeutic response (**Figure 2.1**). These models apply deterministic or probabilistic rules to model the molecular context through which a cell or tissue progresses to malignancy. Different transcriptional classifiers have been developed for the discrimination of cancer type, subtype, and grade [106-108], including hierarchical clustering [109, 110], k-means clustering [111], support vector machines [112], artificial neural networks [113, 114], and classifiers based on the relative expression of gene pairs [115-118]. Transcriptomic signatures, which represent oncogenic processes as a function of gene expression levels, have also been applied to model relapse and overall survival in diverse cancers, including non-small-cell lung cancer and pediatric leukemia [119-122]. Additionally, transcriptional classifiers have been used to predict tumor response to chemotherapeutic agents [123-125], which represents one of the most important clinical determinations in oncology practice.

These models often begin with genome-scale microarray data and, through either computational or combined computational and experimental analyses, derive prognostic classifiers comprised of a smaller number of highly relevant transcripts. The statistical analysis of molecular signatures has potential utility for informing small molecule, radiological, and surgical cancer treatment choices that cannot be filled by standard histopathological or clinical analyses. While these transcriptomic models have not entered widespread clinical use, they embody a promising direct application of cancer models to medicine. Furthermore, such models may be more widely adopted as nucleotide measurement technology itself becomes more common in non-research contexts. Transcriptional models of cancer also provide an accessible characterization of the molecular processes that mediate tumor initiation, progression, and sensitivity to medical interventions.

Pathway Enrichment Models and Biological Discovery

The genomics revolution has prompted substantial efforts towards comprehensively annotating biological functions of and associations among identified genes [78, 126]. Several curated databases have been established for grouping genes into modules of defined activity according to physical or functional association, such as Gene Ontology [81] and the KEGG database [127]. Multiple computational methods have been developed for the purpose of mapping genetic and transcriptional activity into these groups [128, 129]. These tools use diverse numerical methods to determine gene groupings whose activity, as represented by gene expression, is substantially modulated through malignant transformation. For example, Gene Set Enrichment Analysis [130] and related tools have been applied to identify pathway

perturbations in human cancers based on transcriptomic data, including tumors of the breast and prostate [131-133]. An alternative approach combines human protein-protein interaction network data with gene expression profiles for the classification of breast and lung cancer samples [7, 134]. These related studies have considered the cumulative activity of discrete pathways, as defined by protein interaction data, to discriminate between phenotypes. This approach yielded results of similar predictive accuracy to the more common individual gene-based classifiers discussed previously, and demonstrates the importance of integrating various data sources for biological modeling. Similar pathway-based analysis methods also have been applied as prognostic indicators of clinical tumor progression—and the most important pathways identified through these computational methods often overlap significantly with those known experimentally to be oncogenic [135-137].

These frameworks have been applied to diverse cancer systems and serve as a robust source of biological discovery at the genome scale. That said, mapping genetic activity into discrete, human-defined gene sets clearly does not accurately recapitulate the full biological complexity encoded by historical evolutionary constraints. Rather, it is a useful tool that leads to further, more detailed studies. Indeed, a large portion of fundamental eukaryotic cell biology remains poorly characterized. This includes complex events such as non-coding transcription and epigenetic phenomena, but also more basic information such as the specific components and mechanisms of various signaling cascades. As such, pathway annotations themselves are subject to incomplete information and biases in curation. Nonetheless, these techniques, while coarse-grained, enable the simultaneous consideration of many cellular pathways and processes using high throughput data.

Network-Based Statistical Inference

In silico cancer models based on statistical analysis of genetic aberrations, transcriptional activity, and pathway enrichment can provide important insights into the molecular phenomena that generate and propagate malignant cells. However, to characterize specific biomolecular mechanisms that drive oncogenesis, genetic and transcriptional activity must be considered in the context of cellular networks that ultimately drive cellular behavior [138-141]. In microbial cells, network inference tools have been developed and applied for the modeling of diverse biochemical, signaling, and gene expression networks [142-145]. Due to the much larger size of the human genome compared to microbes, and the substantially increased complexity of eukaryotic genetic regulation, inference of transcriptional regulatory networks in cancer presents increased practical and theoretical challenges. Additional complexity arises from the various data sources needed for studying human systems such as immortalized cell lines, primary cell lines, animal models, and tissue biopsies, which each provide imperfect recapitulation of *in vivo* cancer

biology and varying degrees of experimental malleability. Nonetheless, several efforts to infer cancer networks—at both the level of individual pathways and at the genome scale—have revealed important information regarding the molecular basis of oncogenesis.

Network Inference at the Pathway Level

This scale of challenge to construct quantitatively predictive network models at the genome scale in humans is not likely to be met in the near future [146], but more modest inference problems focusing on pathway-based subnetworks continue to prove useful. These studies examine different forms of biological data to infer causal relationships between genes at the transcriptional or signaling levels. For example, a multiparametric flow cytometry approach recapitulated a cell surface signaling pathway in the activation of human CD4+ leukocytes using a Bayesian network inference procedure [143]. Other studies have examined apoptotic signaling networks following experimental perturbation with pro-apoptotic TNF and other agents, using an innovative method in which downstream molecular perturbations are mapped into concordant "principal signaling components" [147, 148]. These studies have shown that a small number of canonical signaling patterns can accurately describe even a complex cellular response such as apoptosis. Furthermore, by comparing signaling in different cell types with differing phenotypic responses, it was found that differential upstream signaling—at the level of kinase activation, for example—most strongly mediates cell-specific stimulus response, but that different signals are propagated through common intracellular networks.

A recent examination of receptor tyrosine kinases found that the complex and overlapping cross-talk involved in such signal transduction can be explained substantially by linear combinations for docking affinities for downstream proteins [149]. This study represents a systems-level characterization of an important cellular system that is too complex to be fully explicable using traditional "univariate" biological modeling. Another recent study considered the cross-talk between phosphorylation pathways to predict the effects of inhibitor treatment on cancer cell migration phenotype [150]. This study characterized off-target and cross-talk effects of kinase inhibitor treatment to predict the effects of combinatoric treatments to forestall cell migration—an important malignant phenotype in cancer. Using different experimental and computational tools, these pathway-centric inference procedures have revealed important insights into diverse systems-level events in cancer that can have significant applications to pharmaceutical development and target selection.

These network inference studies incorporate *a priori* information in the design of the experiment itself—for example, in guiding the selection of what to measure (genes, proteins, and/or phosphorylation events) and the planning of perturbation experiments to informatively test the system. This model-guided

approach constitutes a reduction and focusing of network search space that increases the fidelity of network inference. These methods are not easily extendable to the inference of larger and more complex networks, as the larger number of variables introduces a much larger amount of possible interactions, with associated experimental noise and statistical uncertainty. As larger networks are examined, the task of "model identifiability" becomes an increasingly refractory but critically important challenge.

Network Inference at the Genome Scale

The inference of biological networks at the genome scale involves several challenges for eukaryotic systems, and thus most network inference methods have been developed initially using unicellular organisms. One example is the "Inferelator" computational suite, which has been used to infer a genome-scale transcription regulatory network in the unicellular archaeon Halobacterium [151, 152]. This study constructed a network model of transcription able to determine the level of expression of each gene given the expression state of transcriptional regulators. The model was able to predict quantitatively the response of the organism to novel stimuli that were not used to train the model itself. This type of quantitative inferred genome-scale transcriptional regulatory network in humans remains a difficult and ongoing challenge.

Nonetheless, efforts have been made to craft network-based statistical models of human cells, including in breast cancer and lymphoma, in which the architecture of regulatory networks for a portion of the human genome is characterized [153-156]. These models have applied Bayesian network inference methods [157] to discriminate physical or functional interactions between several thousand genes. The related probabilistic Boolean network formalism [158] has been used to construct regulatory networks in the context of several cancer systems, including glioma [159]. One study of melanoma adapted Boolean network analysis to include the use of "seed" genes, in which groups of genes believed to interact were used as a subnetwork starting point which was then iteratively grown to incorporate other related genes [160]. Many network inference methods appear robust when validated on synthetic data, but exhibit substantially reduced performance when applied to experimental gene expression datasets [161-163]. These results speak to the substantial complexity of gene regulation networks—even in the context of unicellular organisms—and to the requirement for large datasets in network inference tasks [164].

Statistically inferred network models can be used to study the topology of complex cellular systems, and elucidate important genetic interactions and control points. Such statistical networks map the status of regulatory agents into qualitative interactions with cognate targets and discrete network states. Going forward, the acquisition of larger high-throughput datasets encoding different facets of regulatory interactions, combined with innovative methods for their integration and analysis, will enable the

construction of increasingly precise numerical network models. The important ability to reliably model the behavior of cancer cells, for example upon response to treatment or genetic evolution, would hold significant implications for the development and administration of cancer treatments.

Biochemical Reaction Networks

In contrast to statistically inferred networks, biochemical reaction networks are constructed to represent explicitly the mechanistic relationships between genes, proteins, and the chemical interconversion of metabolites within a biological system (**Figure 2.2**). In these models, network links are based on pre-established biomolecular interactions rather than statistical associations; significant experimental characterization is thus needed to reconstruct biochemical reaction networks in human cells. These biochemical reaction networks require, at a minimum, knowledge of the stoichiometry of the participating reactions. Additional information such as thermodynamics, enzyme capacity constraints, time-series concentration profiles, and kinetic rate constants can be incorporated to compose more detailed dynamic models.

Stoichiometric Models of Biochemical Reaction Networks

The most basic mathematical representation of a biochemical reaction network is a stoichiometric model. Stoichiometric models describe the interconversion of biomolecules purely in terms of the number of reactants and products participating in each reaction. The generation of stoichiometric models and analysis of their properties is a well-established process [50-52], and genome-scale models of metabolism have been completed for a diverse range of organisms, including the prokaryote *E. coli* [53], archaea [54], and eukaryotes such as *S. cerevisiae* [55] and the protozoan *L. major* [56]. Most important for cancer research is the recent reconstruction of known human metabolism at the genome-scale [165, 166]. Methods have also been developed for reconstructing signaling networks [167-169], transcriptional and translational networks [50], and regulatory networks [170, 171]; these models are fundamentally analogous to reconstructed metabolic networks (**Figure 2.3**). The reconstruction of a biochemical reaction network results in a database of stoichiometric equations that can be represented mathematically to form the foundation of a genome-scale, computable model. Computational tools for constraint-based analysis are then used to interrogate the properties of the reconstructed network *in silico*, and to facilitate model-driven validation and refinement [57]. Physico-chemical and environmental constraints under which the network operates are applied in the form of *balances*, including mass, energy, and charge, and *bounds*, such as flux capacities and thermodynamic constraints. The statement of constraints defines a solution

space comprising all non-excluded network states, thereby describing possible functions or allowable phenotypes. These methods are now being adapted for modeling human systems in greater detail.

Constraint-based analysis of biochemical reaction networks has been applied to a number of human systems. Using the reconstruction of the human mitochondrial metabolic network [172], linear programming and random sampling were applied to identify candidate steady states of the network under normal, diabetic, ischemic, and dietetic conditions [173]. In a related study, Monte Carlo sampling of flux spaces was used to study the effects of enzymopathies on the human erythrocyte metabolic network [174]. A stoichiometric formalism was also used to reconstruct reactions comprising the JAK-STAT signaling network in the human B-cell [175]. In this study, a topological constraint-based approach was employed to evaluate network cross-talk, redundancy in signaling inputs and outputs, and to delineate correlated reaction sets or systemic modules. The recent completion of a global reconstruction of the human metabolic network [165] represents a significant milestone in human systems biology—and clears a path forward for reconstructing metabolic models for all of the 200 cell types in the human body and their modified forms in various cancers. This reconstruction comprises 1,496 genes and 3,748 reactions divided into 88 metabolic pathways. In addition to the typical network capabilities determined by constraint-based modeling, the initial genome-scale reconstruction has enabled analysis of relationships between network topology and human metabolic diseases [176]. An independently reconstructed human metabolic network [166]—comprising 2322 and 2823 reactions among 70 different pathways—was also used to demonstrate the potential of systems modeling in human metabolism to aid in drug discovery [177]. An important effort going forward will be to resolve the discrepancies between different versions of the human reconstructions, establishing a consensus global reconstruction of human metabolism.

Modeling Metabolism in Human Cancer

The global human metabolic reconstruction provides a basis for the known set of metabolic reactions catalyzed by human proteins. However, the utility of these models for cancer research going forward depends upon overcoming several challenges. First, further refinement of the global human metabolic map is essential to increase its accuracy. Second, each of the approximately 200 cell types in the human body exhibits only a portion of the full metabolic capability contained in the genome [178] (**Table 2.2**). The high percentage of undetermined activities for metabolic enzymes in human tissues clearly shows how much more we have to learn about even this very well studied cellular process. Effectively representing which portions of the global human metabolic network are active in any given cell type, and at what level, is thus of critical importance. Cancers in particular are known to exhibit diverse metabolic phenotypes compared to their progenitor cells, typically with an increased rate of overall metabolic

activity to support their increased growth and the highest metabolic activity observed in the most aggressive malignancies [179, 180]. Multiple other hallmarks of cancer including angiogenesis, metastasis, evasion of apoptosis, and avoidance of immune detection have been previously linked to human tumor metabolism [41, 43]. Metabolic targets have also been used in cancer chemotherapy [181-183]. For these reasons, metabolic networks in human cancer have the potential to be a rich focus area for systems modeling going forward.

Kinetic Models of Biochemical Reaction Networks

Detailed dynamic models are needed to accurately quantify many important molecular processes, such as feedback and feedforward regulation, competitive inhibition, post-translational modification, and transcriptional regulation. At scales much smaller than the genome, biochemical reaction networks have been used as the basis for forming dynamic differential equation models with the addition of kinetic rate constants [167, 184, 185]. Many of the fundamental considerations and challenges associated with constructing kinetic models of cell signaling pathways have been previously reviewed in detail [186, 187]. Briefly, these models are most commonly formulated as a set of ordinary differential equations (ODEs). In these ODE-based pathway models, the rates of production and consumption of individual biomolecular species is described in terms of mass action kinetics (elementary reactions with forward and reverse rate constants). As the full kinetic detail for reactions in biological pathways is often unknown, simplifying assumptions such as the Michaelis-Menten approximation or time-scale separation can be applied. Kinetic models are advantageous in that they do not employ a steady-state assumption as constraint-based models typically do, and thus can simulate detailed dynamic behavior. Dynamic models are also able to simultaneously account for both the concentrations of compounds and flux through reactions. Capturing dynamics however, requires more parameters, and thus these models are more data intensive to create as well as more prone to overfitting when parameters must be estimated. Nevertheless, smaller scale dynamic models have been used successfully to study in mechanistic detail key pathways related to human cancer. Most of the pathways and molecular components listed below do not function independently to promote oncogenesis, but rather are often connected through signaling cross-talk, feedback mechanisms, and other forms of up- or downstream regulation. These examples are not an exhaustive list, but serve to demonstrate the benefits of detailed kinetic modeling of biological pathways.

Aberrant activity of the transcription factor NF-κB has been linked to oncogenesis, tumor progression, and resistance to chemotherapy. A computational ODE model was employed to determine the role of inhibitor of NF-κB kinase (IκB) isoforms in the temporal control of NF-κB [167]. Analysis of the model revealed that IκBα is responsible for strong negative feedback and fast turn-off of the NF-κB response to

IKK stimulation, while IκBβ and IκBϵ reduce oscillations in the signaling module and stabilize NF-κB response during longer stimulation. The same model was expanded and used to study dynamics of IκB–NF-κB signaling module when cells were stimulated by lipopolysaccharide (LPS) via Toll-like receptor 4 (TLR4) [184, 188]. Modeling of two TLR4-dependent signaling pathways suggested that one pathway required the expression tumor necrosis factor-α (TNFα) to activate NF-κB.

In another study, the negative feedback loop between the tumor suppressor p53 and the Mdm2 oncogene was examined experimentally and computationally [189]. Different mathematical models of the system were analyzed to draw connections between predicted and experimentally observed behavior. The models indicated that low-frequency noise in protein production rates was an important source of variability in oscillations of the p53-Mdm2 system. Another model was built for the single-cell response of p53 to radiation-induced DNA damage [190]. In agreement with experiments, simulations showed that by assuming stochasticity in the initial number of double-strand break sites and the DNA repair process, p53 and Mdm2 exhibit coordinated oscillatory dynamics with radiation stimulation.

Numerous components of the phosphatidylinositol-3-kinase (PI3K)/AKT pathway are targeted by malignant perturbations such as amplification, mutation, and translocation in cancer patients. Activation of the PI3K pathway results in aberrant cell growth and survival, metastatic competence, and therapy resistance. An excellent review of the challenges, opportunities, and progress in modeling and exploiting the PI3K pathway can be found in Hennessy et al., 2005 [191]. For example, inhibitors of the mTOR pathway have been demonstrated to directly target PI3K signaling. However, due to the intricate role of mTOR in regulation of PI3K/AKT signaling, detailed understanding of the pathway derived from computational modeling is needed to determine the consequences of inhibition of particular components.

A number of models have been used to examine mechanisms that govern mitogen-activated protein kinase (MAPK) pathway dynamics. MAPK pathways are involved in the activation of many cancer-promoting genes and respond to a number of different cancer-associated stimuli, such as epidermal growth factor (EGF). A computational model of general MAPK signaling pathways was used to study the role of negative feedback mechanisms in generating complete signal adaptation (i.e., desensitization to stimulus) [192]. Another model was used to simulate EGF-induced MAPK activation [193]. Analysis of model simulations provided quantitative evidence that feedback inhibition of the MAPK cascade is the most important factor in determining the duration of cascade activation. It was also found that differences in EGF and NGF-stimulated signaling could be accounted for by differential feedback regulation of the MAPK cascade. Similarly, a model was constructed to examine the signal-response relationships between

EGF binding and the activation of downstream proteins in the cascade [194]. Model predictions showed that EGF-induced responses are remarkably stable over a 100-fold range of ligand concentration.

Another model used dynamic simulations of gene expression and signaling networks to study behavior in human cancer cells [195]. While dynamic interaction networks of cancer systems are challenging to construct and simulate, such models are essential to capture the molecular complexity of oncogenesis. Taken together, these studies provide powerful examples of how building a model that correctly explains data and makes verifiable predictions can lead to deepened insight into biological mechanism and function.

Microenvironment and Tissue-Level Models

In silico models of cancer can be built not only for intracellular networks, but also at larger length scales. Alternative computational methods must be applied to consider the interface between cancers and the tissue contexts in which they reside. These settings exhibit complex interactions with multiple factors of different function and scale, including extracellular biomolecules, a spatially intricate and dynamic vasculature, and the immune system [196]. Models of cancer at the tissue level that account for these functionally divergent parameters can be broadly divided into "continuum" models and discrete or "agent-based" models [197-199]. The latter are often applied when the number of individual interacting units, such as cancer cells, is constrained to remain small; the former is more practical at population scales where agent-based modeling can be computationally prohibitive. Both methods can integrate information about the biological context in which cancers develop, and thus represent a multi-scale consideration of oncogenesis as it occurs within somatic tissues [200].

Continuum Models

Extracellular parameters can be represented as continuously distributed variables to mathematically model cell-cell or cell-environment interactions in the context of cancers and the tumor microenvironment. Systems of partial differential equations have been used to simulate the magnitude of interaction between these factors, including the effects of hypoxia on cell cycle progression [201], the impact of mechanical forces on tumor invasiveness [202], and extracellular matrix interactions [203]. Recent studies have examined cell population dynamics within colonic crypts in colorectal cancer [204, 205]. These models consider interactions between stem cells, differentiating cells, and differentiated cell populations to quantitatively predict tissue-level invasion and the growth of tumor mass. Other models have represented solid tumors as a multi-phase system of both bound and "mobile" forms [206, 207]. Such "mixture" models consider differential growth and apoptosis rates as well as mass transfer and

regulatory interactions between phases. Alternative models have considered nonlinear and combinatoric effects of multiple factors, including nutrient availability and mechanical parameters [208, 209] and the effects of mutation rate on invasion and metastasis [210]. These numerical systems embody a robust method to incorporate the effects of somatic biological phenomena into computational representations of cancer. Continuum-based models are thus powerful tools to simulate and characterize interactions between intracellular and extracellular factors in oncogenic processes.

Cellular Automata and Agent-Based Models

Multivariate continuum models are able to represent the effects of several physiological or biochemical events on cancer development. However, *in situ*, these factors are highly heterogeneous, and interact discontinuously with tumor cells [211, 212]. Cellular automata models represent cancer cells as discrete entities of defined location and scale, interacting with one another and external factors in discrete time intervals according to predefined rules [213]. Agent-based models expand the cellular automata paradigm to include entities of divergent functionalities interacting together in a single spatial representation, including different cell types, genetic elements, and environmental factors [214]. With sensitivity to starting conditions and the ability to incorporate probabilistic interactions at each time step, these models can exhibit similar stochastic behaviors to those observed in oncogenesis *in vivo* [215-217]. Phenomena that have been modeled using agent-based models include three-dimensional tumor cell patterning [218], immune system surveillance [219], angiogenesis [220-222], and the kinetics of cell motility [223]. Another recent model integrated diverse parameters such as extracellular signals, blood flow, and tissue degradation to simulate the spatiotemporal formation of tumor vasculature [224].

Increasingly, "hybrid" models have been created which incorporate both continuum and agent-based variables in a modular approach. For example, a recent study considered continuous extracellular biomolecule distributions and discrete cell locations to simulate the interaction between intracellular decision-making processes and malignant growth [225]. Another recent model incorporated a continuous model of a receptor signaling pathway, an intracellular transcriptional regulatory network, cell-cycle kinetics, and three-dimensional cell migration in an integrated, agent-based simulation of solid brain tumor development [226]. The interaction between cellular and microenvironment states have also been considered in a multiscale model that predicts tumor morphology and phenotypic evolution in response to such extracellular pressures [227]. These and other techniques that incorporate multiple, nested scales of interacting biology embody promising paradigms to understand cancer as a cascade of information across levels of size and complexity. The ability to interrogate cancer across multiple biological agents and

compartments presents a unique framework to elucidate oncogenic processes and to evaluate potential therapeutic interventions through digital simulation prior to experimental deployment.

Conclusions & Future Directions

The ability to model the initiation and progression of human cancer at the level of molecules, cells, and tissues is important to improve cancer diagnosis and treatment. In **Chapter 1**, I have discussed the biological motivations for *in silico* models of cancer with regards to the molecular complexity observed in the disease, which is best understood in the context of systems-level interactions. Similarly, tumors interface with an intricate environment of several cell types and extracellular factors, an additional layer of complexity that will require systems-based approaches to address reliably. Powerful new technologies enable the high-throughput biological measurements critical to understanding these processes, including nucleotide sequencing, proteomics, and protein interaction assays. Additionally, the standardization and distribution of this data through common file formats and internet databases empowers the research community to build efficiently upon previous studies and to cumulatively expand our knowledge of the disease. Predictive models serve as compact and rigorous representations of large-scale hypotheses for the functioning of disease-perturbed networks and promise to advance rational therapy design in the coming future of systems medicine.

Diverse computational methods have been developed to analyze different elements of cancer biology, and to elucidate multicomponent, systems-level contributors to cancer development. Statistical models consider cancer cell biology at the pathway or genome scale to elucidate genetic or transcriptional signatures associated with oncogenesis. The ability to identify molecular signatures that can inform diagnosis and treatment selection is already beginning to take effect in the clinic, for example with molecularly targeted therapies such as imatinib (Gleevec) for selected subpopulations of cancer [228]. At a finer resolution, quantitative pathway-level models offer explicit representation of intracellular signaling and metabolic events in cancer, and important scientific successes have been observed with these mechanistic pathway models. Integration of many molecular pathways for quantitative genome-scale modeling remains an ongoing challenge, with statistical inference of cellular networks currently possible only in a very limited context in human cancers. At the physiological level, continuum and agent-based models efficiently describe the interaction between cancer cells and the surrounding tissue environment. Cumulatively, these models are able to integrate many different sources of experimental data to simulate and observe diverse aspects of cancer biology. Translation of these models to the clinic, while currently a slow process, suggests a substantial promise of systems medicine that is just beginning to be explored.

The future prospects for *in silico* models of cancer are substantial, and this field is clearly still in its infancy. With rapidly advancing technologies for biologic interrogation at the genetic, transcriptional, proteomic, and cellular levels, the body of data upon which models can be trained is expanding even faster. Coupled with the persistent exponential rise of computing power, the "raw materials" of systems biology will continue to improve at an exhilarating rate. With the broad availability of annotations for cellular pathways and biomolecular interactions, specific constraints can now be imposed across the genome on model architectures to enhance dimensional parsimony and statistical resolution. Additionally, a broad array of tools for numerical simulation of both individual biological modules and genome-scale phenomena has been developed. Going forward, the fusion of experimental, informational, and analytical modalities will provide a powerful analytical framework for *in silico* modeling of human cancer.

In silico cancer modeling presents significant opportunities to investigate oncogenesis across biological scales and systems. These powerful methods should help to accelerate the development of diagnostic and therapeutic technologies for clinical medicine. Considerable improvement in the resolution, scale, and predictive power of these models must first be achieved, regarding which substantial challenges remain. Nonetheless, it may ultimately be possible to simulate oncogenesis and malignant invasion accurately from the scale of genetics to physiology. With this technology, distinguishing signatures of cancers could be discovered automatically for purposes of early diagnosis, prognosis, and treatment planning. Additionally, and perhaps most promisingly, with reliable digital representations of cancer, the effects of therapeutic interventions at both the molecular and surgical scales could be predicted *in silico* without exposing patients to risk. This innovation would greatly accelerate the development of safe and targeted anticancer therapeutics, and offer hope of medical treatments for diseases that remain refractory to current clinical technologies. Ultimately, while confronting substantial experimental and analytical challenges, *in silico* models of cancer are advancing, and promise to strongly enhance both the fundamental understanding of cancer and its treatment in the clinic.

Chapter 2 Figures & Tables

Figure 2.1. Schematic overview of statistical modeling in human disease.

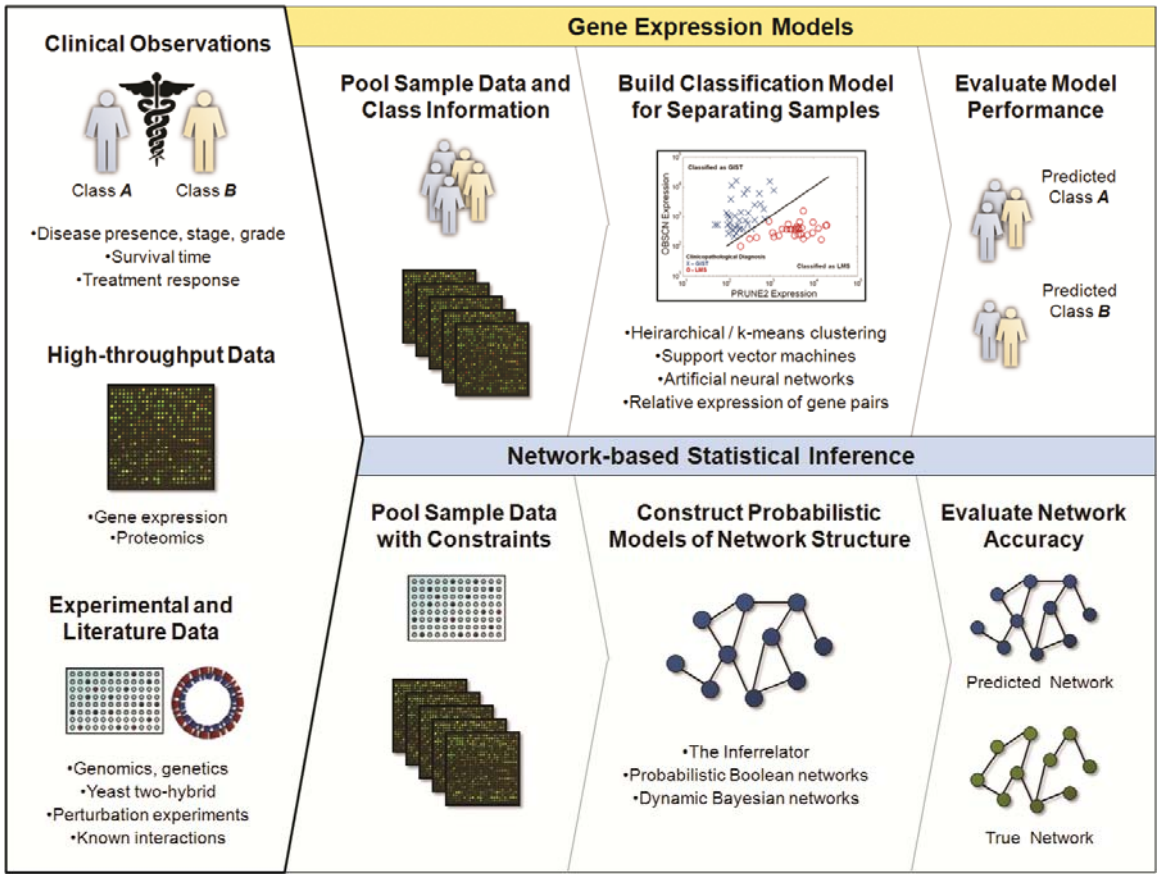


Figure 2.2. Comparison of biochemical reaction network and statistical network models.

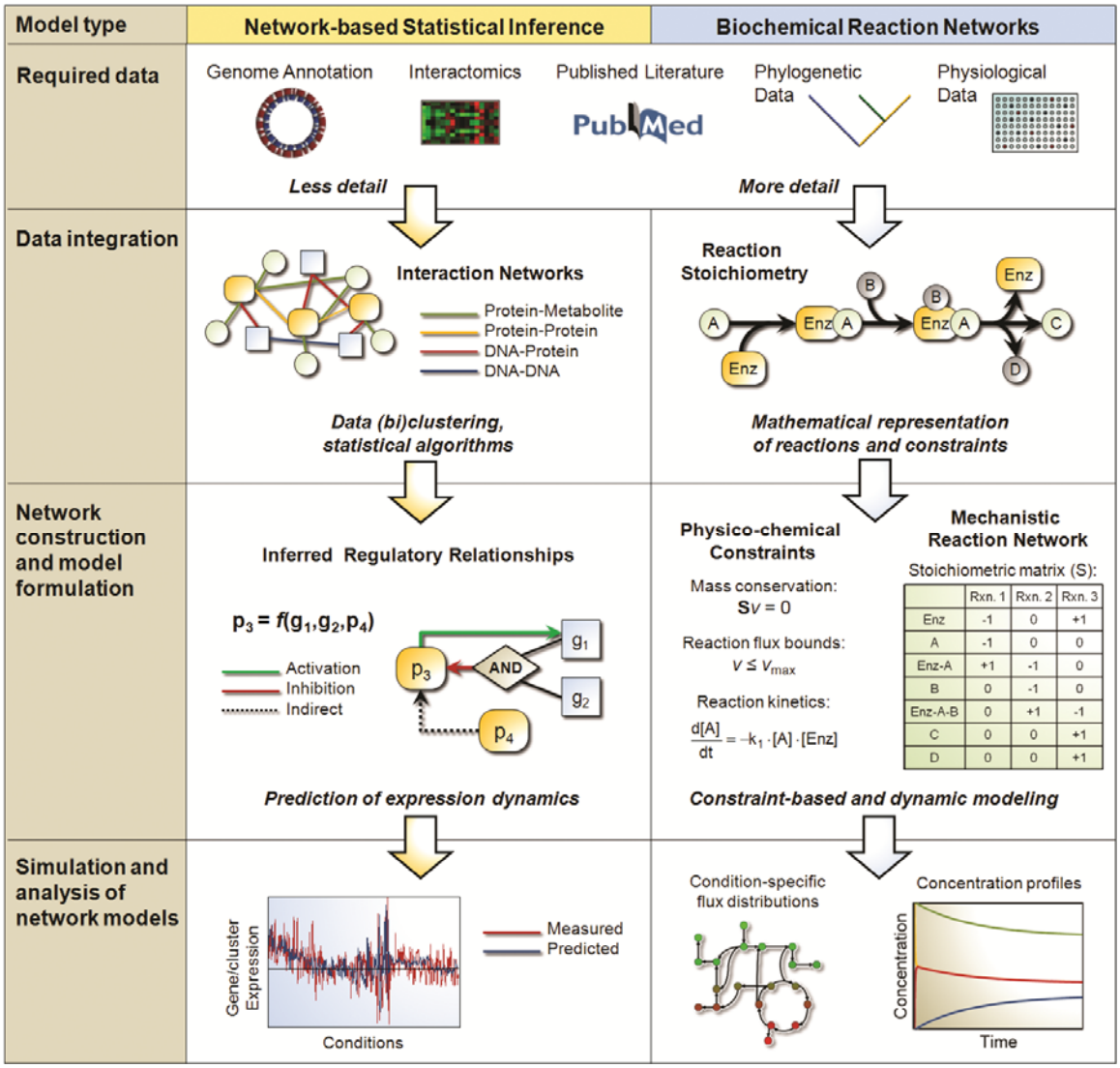


Figure 2.3. Mathematical representation of reaction links in biochemical networks.

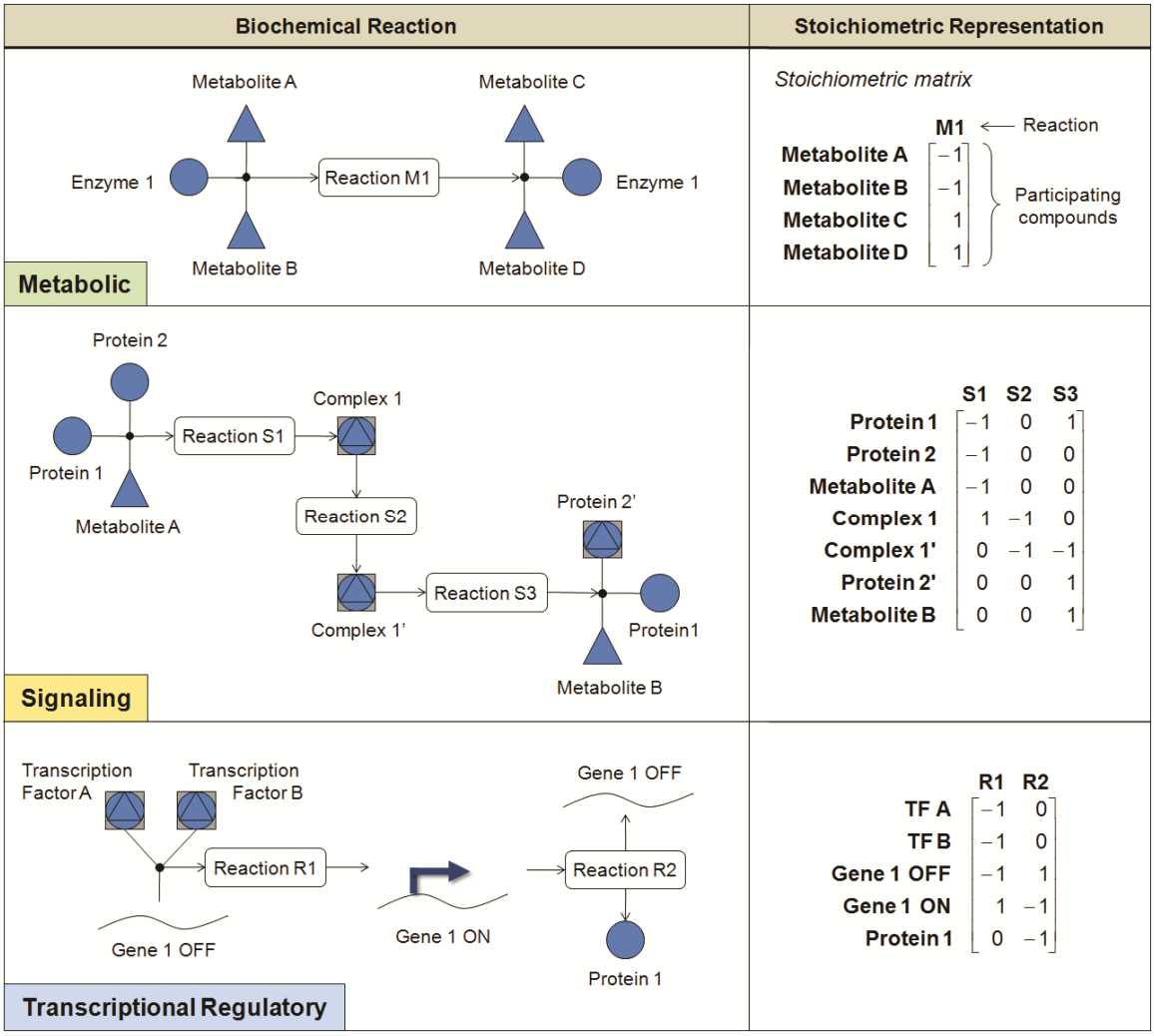


Table 2.1. Databases and software for building systems models of cancer.

Resource	Database	Refs
Genome sequence data	Ensembl	[78]
	UCSC Genome Browser	[126]
Genome annotation data		
Genetic elements	Entrez Gene	[229]
	Gene Ontology Annotation (GOA) Database	[75, 77]
	Universal Protein (UniProt) Knowledgebase	[76]
	Genome Reviews	[79]
Biochemical pathways & functional associations	Kyoto Encyclopedia of Genes and Genomes (KEGG)	[80, 127]
	Gene Ontology (GO)	[81]
	The SEED	[84]
	MetaCyc	[82]
	BioCyc	[83]
	TransportDB	[85]
Regulatory Sequences	Eukaryotic Promoter Database (EPD)	[72]
	Transcriptional Regulatory Element Database (TRED)	[73]
Model, model parameter repositories	Kinetic Data of Bio-molecular Interactions Database	[88]
	BioModels Database	[86]
	Database of Quantitative Cellular Signaling (DQQCS)	[87]
Protein interaction networks	Database of Interacting Proteins (DIP)	[91]
	Molecular INTeraction Database (MINT)	[92]
	Mammalian Protein-protein Interaction Database (MIPS)	[93]
High-throughput genome-scale data		
Transcriptomics	Gene Expression Omnibus (GEO)	[70]
	Standford Microarray Database (SMD)	[71]
Proteomics	Proteomics Identifications Database (PRIDE)	[74]
Visualization and data management software packages	Cytoscape	[99]
	The Gaggle	[98]

Table 2.2. Predicted tissue-specific activity of metabolic genes in the human metabolic reconstruction (data from [178]).

Tissue	Active (of 1165)	Inactive	Undetermined
Liver	12.4%	19.2%	68.4%
Brain	7.0%	32.1%	60.9%
Heart	4.2%	22.0%	73.8%
Kidney	12.9%	21.5%	65.7%
Lung	3.9%	33.6%	62.5%
Pancreas	3.8%	35.5%	60.8%
Prostate	2.8%	30.0%	67.1%
Spleen	1.1%	38.5%	60.4%
Thymus	1.0%	37.3%	61.6%
Skeletal Muscle	4.6%	28.0%	67.4%

Chapter 3. Relative Expression Analysis for Molecular Cancer Diagnosis and Prognosis[3]

The enormous amount of biomolecule measurement data generated from high-throughput technologies has brought an increased need for computational tools in biological analyses. Such tools can enhance our understanding of human health and genetic diseases, such as cancer, by accurately classifying phenotypes, detecting the presence of disease, discriminating among cancer sub-types, predicting clinical outcomes, and characterizing disease progression. In the case of gene expression microarray data, standard statistical learning methods have been used to identify classifiers that can accurately distinguish disease phenotypes. However, these mathematical prediction rules are often highly complex, and they lack the convenience and simplicity desired for extracting underlying biological meaning or transitioning into the clinic. In this chapter, I survey a powerful collection of computational methods for analyzing transcriptomic microarray data that address these limitations. Relative Expression Analysis (RXA) is based only on the relative orderings among the expressions of a small number of genes. Specifically, I provide a description of the first and simplest example of RXA, the *k*-TSP classifier, which is based on *k* pairs of genes; the case $k = 1$ is the TSP classifier. Given their simplicity and ease of biological interpretation, as well as their invariance to data normalization and parameter fitting, these classifiers have been widely applied in aiding molecular diagnostics in a broad range of human cancers. I review several studies that demonstrate accurate classification of disease phenotypes (e.g., cancer vs. normal), cancer subclasses (e.g., AML vs. ALL, GIST vs. LMS), disease outcomes (e.g., metastasis, survival), and diverse human pathologies assayed through blood-borne leukocytes. The studies presented demonstrate that RXA—specifically the TSP and *k*-TSP classifiers—is a promising new class of computational methods for analyzing high-throughput data, and has the potential to significantly contribute to molecular cancer diagnosis and prognosis.

Expression Analysis for Molecular Signature Identification

High-throughput measurements in biology (e.g., transcriptomics, proteomics, metabolomics) provide an enormous amount of information, but only implicitly, in the form of raw expression values. Harnessing this information means converting it to knowledge and, for the purposes of classification, useful decision rules. In particular, this conversion can enable a greater understanding of cancer and drive advances in personalized medicine. A systems-level approach, which employs computational and statistical tools to reveal and evaluate patterns with diagnostic or prognostic value, is critical to fully exploiting these new

[3] This chapter includes material that was reproduced with permission from the following publication: Eddy, J.A., J. Sung, D. Geman, and N.D. Price. 2010. Relative expression analysis for molecular cancer diagnosis and prognosis. Technology in Cancer Research & Treatment. 9(2):149-159. (all sections; text was collaboratively written with Jaeyun Sung and Donald Geman).

technologies. In particular, molecular signatures derived from patterns in gene expression microarray experiments have great potential to detect the presence of disease, to discriminate among cancer sub-types, to predict clinical outcomes, and to provide insight into specific changes that occur during disease progression.

Perhaps the most evident challenge for developing useful molecular signatures is to identify classifiers that are accurate for a specific study or platform and that are also robust across a wide range of settings. Previous studies have aimed to identify sets of individual genes ("signatures") whose differential expression is highly correlated with phenotypic changes (e.g., genes that may be over- or under-expressed in cancer relative to normal). In these cases, increased or decreased absolute mRNA concentration levels above some threshold (i.e., more than would be statistically expected by chance for a gene on the microarray) are put forth as candidates for disease-induced (or causing) perturbations. Unfortunately, the statistically significant genetic changes often depend largely on the context of the microarray experiment. Even when thresholds are tuned to produce statistically significant results, findings can depend heavily on a number of factors, such as the experimental design and the type of data normalization. Consequently, there may be little to no overlap in the molecular signatures identified from one platform to another, or by extension, from one clinical setting to another.

A less evident, but equally important, challenge for phenotype classification using gene expression data is to develop techniques that not only yield accurate and robust decision rules, but also provide rules that are easy to interpret and might contribute to biological understanding. Advanced statistical learning and pattern recognition methods are routinely applied to transcriptomics and other high-throughput data types. These include neural networks [230-232], decision trees [233-235], boosting [234, 236], and support vector machines [237, 238]. In many cases, these methods achieve good classification performance, with sensitivities and specificities above ninety percent. However, they generally result in extremely complex decision rules based on nonlinear functions of many gene expression values. Therefore, whereas advanced methods may be more accurate than those based on the patterns of individual genes, they usually produce decision rules that are virtually impossible to interpret. Furthermore, as the number of variables (transcripts) far exceeds the number of observations in most microarray studies, building more complex classifiers entails a greater risk of over-fitting the training data and poor generalization.

An important potential benefit of simple and interpretable decision rules is to provide insight into the underlying biological differences between phenotypes. Notably, malignant phenotypes in cancer arise from the net effect of interactions among multiple genes and other molecular agents within biological networks. Genes in networks operate in a combinatorial manner—the actions of one gene greatly

influence the actions of other genes. This often limits the information that can be gleaned from the expression patterns of individual genes. As an alternative approach, studying gene expression in the context of networks may yield greater insight into mechanisms and functional changes associated with disease. Recently, methods for analyzing microarray data have focused not on individual genes, but instead on biologically meaningful pathways or networks [128, 239-241]. These frameworks have been applied to diverse cancer systems and serve as a robust source of biological discovery [240, 242].

At scales smaller than biological networks or even pathways, assessing the relationships among a small number of genes—for example, the patterns of interactions among just two or three genes—can provide useful information about biomolecular processes. One way to probe the interactions among several genes is to study their *relative* expression, i.e., the ordering among the expression values, rather than their absolute expression values. One then searches for characteristic perturbations in this ordering from one phenotype to another. The simplest form of such an interaction is the ordering of expression among two genes, in which case one seeks to identify typical "reversals"—pairs of genes for which one of the two possible orderings is usually present in one phenotype and rarely present in the other. We refer to the family of such rank-based methods as Relative Expression Analysis (RXA). This methodology is characterized by replacing each expression level across all genes by its corresponding rank within a single microarray profile.

Here, I focus on RXA methods that involve a small number of gene pairs, each exhibiting a characteristic "relative expression reversal" between the phenotypes or classes of interest. Aggregating the decisions from a few such pairs, even just one, is surprisingly powerful. Basing decisions on one pair is called the top-scoring pair (TSP) classifier [243] and on k pairs is called the k-TSP classifier [116]. Thus, in TSP, a sample is classified based on a decision rule that only involves comparing the ranks, hence the relative expression levels, of two genes within a profile. For the k-TSP classifier, the decision rule combines a disjoint set of TSPs by simple majority voting. Other RXA methods include those based on the six possible orderings among three genes (the top-scoring triplet classifier [244]) and comparing the average ranks in two groups of genes [117]. Herein I review the TSP and k-TSP computational methods, focusing on their utility for aiding molecular diagnostics in a broad range of human cancers. This review is largely restricted to applications with transcriptomic data, as this is the most plentiful and has been the most used to date. However, RXA is generally applicable to any ordinal data type, such as protein expression, DNA copy number, chromosomal position, and so forth.

Relative Expression Analysis

Microarray Data and Analysis

For those readers less familiar with computational approaches to microarray analysis, I first describe the typical features of microarray data and common procedures relevant to the results presented here. Whereas I discuss computational analysis of microarray data in the context of RXA (**Figure 3.1**), the notation, representation of the data, and basic steps are similar for other approaches. Computational analysis of microarray data typically involves two steps. First, a classifier is trained on a collection of microarray profiles (samples) referred to as the training set. This involves selecting a subset of genes and choosing a mathematical algorithm (decision rule) to apply to the selected genes in order to determine the phenotype of a new sample. Of course the goal is to identify an algorithm that is predicted to work well on a new dataset, and the second step is then to evaluate the performance of the classifier on held-out data. Usually, the algorithm works quite well on the training data, and hence validation is essential.

Microarray data are typically represented as a matrix of G rows of genes and N columns of samples (e.g., different tumors, tissues, patients, time points). The n^{th} column of this matrix is therefore a G x 1 vector representing the expression profile $\mathbf{x}_n$ of the n^{th} sample. Each profile contains expression values for gene one (g_1) through gene G (g_G). The expression level of the gene g_i is denoted by X_i. In addition, each sample is labeled by a phenotype $Y \in \{\text{A, B, } \ldots\}$. For example, $y_n = \text{A}$ indicates that the n^{th} sample belongs to phenotype A. The labeled dataset to be used for classifier training is $F = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$.

As mentioned above, the simplest method for classifying expression profiles based on the relative ordering of expression values is the top-scoring pair (TSP) algorithm for distinguishing between two phenotypes A and B. In TSP, a particular pair of genes i and j is selected during training and the decision rule is simple maximum likelihood: for the sample to be classified, choose the class, A or B, for which the observed ordering between the expression values of g_i and g_j is the most likely. Notice that the observed ordering is either $X_i < X_j$ or $X_i > X_j$ (we can assume at this point that ties are broken at random). The pair which is chosen is the one that achieves the highest "score" among all pairs of genes. This score is a quantitative measure of the degree of relative expression reversal estimated from the data and used for classifier training, as explained in the following section. For the k-TSP classifier, the decision rules are conceived in the same manner as in the TSP classifier, but use a combination of gene-pair markers to obtain potentially better classification accuracy. There are currently three software implementations available for researchers who wish to apply these methods: one in Perl [116], one in R [245], and one, developed by myself and fellow graduate student John Earls, that provides a graphical user interface wrapped around Python functions [246].

The TSP and *k*-TSP classifiers are parameter-free methods that are invariant to all normalization techniques that are monotonic transformations of the original expression values within each chip or microarray. That is, if the data are processed in such a way that if gene g_i is expressed more than gene g_j before normalization (original data) and it is still expressed more after normalization (processed data), then the TSP and *k*-TSP classifiers derived from the original and processed data are the same. It is in this sense that these classifiers are "invariant" to normalization. Moreover, the TSP and *k*-TSP classifiers are especially favorable in terms of the simplicity of the decision rule and the small number of genes involved in classification. They are easy to implement in practice, as the classifier only requires measurement of the expression of small numbers (at most $2k$) of genes using techniques such as RT-PCR. They also remain context-independent by not requiring any parameter tuning or data pre-processing based on genes outside of the pairs involved. Furthermore, because data normalization is not required, RXA classifiers have been shown to be useful in the integration of data across different studies and platforms for the purpose of increasing sample size and facilitating meta-analysis of microarray data [247].

Training RXA Classifiers

In relative expression approaches, the features selected are *pairs* of genes. Consider first TSP. Because only gene pairs are considered, it is possible to completely enumerate all possible pairs and select the "best" ones using the training data. The natural criterion is performance, which anticipates how the pair of genes will be used for classification. As a result, one then selects the pair of genes g_i and g_j for which the difference $|\text{Prob}(X_i < X_j \mid Y = \text{A}) - \text{Prob}(X_i < X_j \mid Y = \text{B})|$ is maximized. This can be shown to be equivalent to choosing the pair of genes that achieves the lowest error rate on the training set. In many cases, there is a single pair of genes achieving the top score. Otherwise, in order to select a unique pair of genes, a secondary score is applied, which is based on the average difference in expression values over all samples. An important feature of the top-scoring pair of genes is that it may not be the case that both genes are highly differentially expressed on the basis of their individual *t*-statistics; in fact, one gene may serve as a "pivot" for the other.

Depending on which of the two probabilities $\text{Prob}(X_i < X_j \mid Y = \text{A})$ or $\text{Prob}(X_i < X_j \mid Y = \text{B})$ is larger, the decision rule is either:

Rule 1: If expression gene i < expression gene j, THEN class A, ELSE class B.
Rule 2: If expression gene j < expression gene i, THEN class A, ELSE class B.

In the case of *k*-TSP, the classifier is constructed from the *k* top-scoring pairs of genes. Each pair votes for class A or class B the same way as in TSP, and the class with the majority vote is chosen. Effectively, this is the maximum likelihood rule: choose the class for which the *k* observed orderings are the most likely. Usually, the pairs are constrained to be "disjoint," meaning that a gene cannot appear in more than one pair, and the number of pairs (k) is determined by cross-validation up to some limit (e.g., $k_{max} = 10$) in order to keep the total number of genes manageable. Consequently, the size of the gene "signature" is two for TSP and $2k$ for *k*-TSP. Unlike other methods, once the signature is determined so is the classifier. That is, there are no parameters to tune, which reduces over-fitting the training data.

Testing RXA Classifiers

Classifier training is followed by performance evaluation on a test dataset. The gold standard for testing any predictive method is to use an independent dataset collected solely for testing. However, due to the scarcity of data, the test set usually consists of samples selected from the original training dataset and set aside. Even in this case, repeated training and testing, known as cross validation, is preferred due to small sample sizes. Such procedures involve splitting the original training dataset F into two smaller sets: the set of samples on which the classifier is trained, F_{train}; and the set of samples on which the classifier is tested, F_{test}. Importantly, no information from F_{test} can be utilized when learning the classifier on F_{train}. The data is repeatedly split into training and test groups, and the cross-validated accuracy is the average classifier performance across all test groups. Leave-one-out cross validation (LOOCV) is commonly used, in which the total of N samples is divided into a training set of size $N - 1$ with the test set consisting of the single remaining sample. While error estimation with LOOCV is known to have high variance relative to the true error [248], it is particularly useful for TSP and *k*-TSP because there is a technique [116] that yields a very significant reduction in the computation involved in looping over all pairs of genes in each loop of cross validation.

A number of different metrics can be used to measure the performance of classifiers. Particularly common measures include sensitivity, specificity, and overall accuracy. These metrics are most easily understood for experiments with a case (e.g., cancer) and a control (e.g., normal), but can be extended to any binary phenotype comparison as well as extended to multiclass problems by decomposing them into sets of binary comparisons. If a classifier correctly predicts that a cancer profile belongs to the cancer class, this is known as a true positive (TP), and the probability of correctly labeling future cancer samples is the sensitivity of the classifier (also known as the true positive fraction). Similarly, a true negative (TN) is when a classifier correctly labels a normal sample, and the probability of doing this on new samples is the specificity of the classifier. Importantly, the sensitivity and specificity computed on the samples used for

training are upwardly biased and not predictive of cross-validated rates. Finally, overall accuracy can be defined in several ways; perhaps the simplest is the average of sensitivity and specificity.

RXA in the Study of Cancer

Cancer Studies Using Relative Expression Values Before TSP and k-TSP

Gene-pair relative expression markers, specifically in the form of a two-gene expression-level ratio, have been previously used for disease classification and prognosis. Gordon *et al*. [249] successfully distinguished between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung based on ratios of expression. Although genetically disparate, the tissues of MPM and ADCA can be difficult to distinguish based on established histopathological methods. Gorden *et al.* [249] tested the fidelity of ratio-based diagnosis in differentiating between the two tissue types in 181 samples (31 MPM and 150 ADCA). First, the investigators used a training set of 32 samples (16 MPM and 16 ADCA) to identify differentially expressed genes based on various methods (fold changes, standard *t*-tests, expression cutoffs, etc.). They then formed 15 ratios using individual or combinations of those genes that showed the highest significance in inversely correlated expression levels. Any single ratio of the 15 examined was at least 90% accurate in predicting diagnosis for the remaining 149 samples (e.g., test set). They then examined (in the test set) the accuracy of multiple ratios combined to form a simple diagnostic tool. Using two and three expression ratios, the investigators found that the differential diagnoses of MPM and lung ADCA were 95% and 99% accurate, respectively. Whereas, in this study, these gene-pairs are not combined in the same way as TSP, they are sensitive to normalization and parameter choices. Still, their work illustrates the utility and discriminatory power of gene pairs in important clinical diagnoses.

Ma *et al.* [250] found that a two-gene expression ratio derived from a genome-wide, oligonucleotide microarray analysis of estrogen receptor (ER)-positive, invasive breast cancers predicts tumor relapse and survival in patients treated with tamoxifen. Tamoxifen is one of the most commonly used medications in the treatment of early-stage and metastatic ER-positive breast cancer [251, 252]. When administered to women with surgically treated ER-positive breast cancer, tamoxifen therapy reduces the annual risk of recurrence by 40-50%, leading to a 5.6-10.9% improvement in 10-year survival [253]. However, 25-66% of women diagnosed with ER-positive breast tumors fail to show a prolonged response or develop early resistance to adjuvant therapy [251, 254]. Currently, there are no markers that reliably predict clinical outcome of cancer patients treated with tamoxifen. Therefore, a reliable means to accurately predict tamoxifen treatment outcome is crucial for early-stage breast cancer management.

In the tamoxifen study conducted by Ma *et al.* [250], a set of 60 patients with receptor-positive primary breast cancers were treated with tamoxifen alone. The results from gene expression profiling of the extracted tumor tissues before therapy indicated that the homeobox gene (HOXB13) was over-expressed in patients who experienced disease recurrence, whereas the interleukin-17B receptor gene (IL-17BR) and EST gene were over-expressed in those with no evidence of recurrence after a 5-year treatment period. The investigators evaluated the prognostic utility of each of these three genes by itself and in combination with genes that have opposing patterns of expression between the two classes. Results from *t*-test and ROC analyses revealed that a two-gene ratio of HOXB13 over IL-17BR had a stronger correlation with treatment outcome than any of the genes alone with AUC values reaching 0.84, and was able to accurately predict tumor recurrence in adjuvant tamoxifen-treated patients

This observation was also confirmed in real-time quantitative PCR analysis, where the predictive accuracy of the two-gene ratio was 81%. Furthermore, the expression ratio of HOXB13 over IL-17BR outperformed existing biomarkers for prognosis of breast cancer, such as patient age, tumor size, grade, and lymph node status. In this study pre-dating any formal RXA classification approaches, Ma *et al*. [250] demonstrated the utility of a two-gene expression biomarker in identifying a subset of patients with early-stage ER-positive breast cancer who are at a risk for tumor recurrence even with tamoxifen therapy. Such a biomarker provides a potential means to identify patients appropriate for alternative therapeutic regimens in early-stage breast cancer.

Comparative Analysis of TSP and k-TSP Performance in Cancer Classification

Geman *et al.* [243] introduced the TSP method and demonstrated its efficacy on several gene expression datasets involving breast, prostate and leukemia cancers. The phenotype classification problems considered were: (i) predicting the status of lymph nodes (affected vs. non-affected) in patients with breast tumors using data from [255]; (ii) classifying sub-types of leukemia (AML vs. ALL) using data from [256]; and (iii) distinguishing prostate tumors from normal profiles using data from [257]. The reported accuracies for TSP results were based on LOOCV, and comparison to randomly permuted data was made to estimate the statistical significance for each classifier.

In predicting the status of lymph nodes in the breast cancer dataset, a cross-validation classification rate of 79% was achieved from 49 patient samples. The authors also mention a separate study where estimated error rates for these data—based on LOOCV and using a wide variety of common machine learning techniques—are summarized for varying numbers of pre-filtered genes [255]. Other methods, more complex than TSP and using many more genes, did not result in better classification rates, and the low

accuracy observed in all methods applied to date is probably a function of the complexity and similarity of the phenotypes being separated. In the case of separating AML from ALL, the TSP classifier correctly classified 68 samples out of 72 samples in cross validation. In comparison, the study in Golub *et al.* [256] used a fifty-gene classifier to predict 65 samples correctly out of 72.

In addition to demonstrating improved performance in classifying breast cancer and leukemia samples, Geman *et al*. [243] also investigated the ability of TSP to detect the presence of prostate cancer. In a previous study, Singh *et al*. [257] found a strong correlation between patterns of gene expression of prostate cancer and various clinical and pathological aspects of the disease. The top-scoring gene pair using the TSP algorithm on their data could discriminate non-tumor versus prostate tumor samples at a prediction rate of 95%. Hence, the classification rates using TSP were comparable to the best results reported previously in the literature, often incorporating hundreds of genes or more in complex decision rules.

The performance of TSP and *k*-TSP classifiers were compared with those of other machine learning methods on 19 gene expression datasets involving human cancers in a study by Tan *et al*. [116]. The study investigated a number of publicly available datasets, with sample sizes ranging from 33 to 327 for each disease phenotype within a particular dataset. The collection of datasets comprised various studies of human cancer, including colorectal, leukemia, lung, prostate, breast, central nervous system, lymphoma, bladder, melanoma, renal, uterus, pancreas, ovary, and mesothelioma. The classification performance of TSP and *k*-TSP was compared to that of decision trees (DT), Naïve Bayes (NB), *k*-nearest neighbor (*k*-NN), support vector machines (SVM), and prediction analysis of microarrays (PAM), which is essentially linear discriminant analysis. The TSP and *k*-TSP techniques were also extended beyond binary classification to the multiclass setting, where several well-known aggregation strategies, such as "one-vs.-all" and "one-vs.-other," were applied to combine the results of binary sub-problems into one final decision rule.

In this study, LOOCV was used in order to estimate the classification rate. The best classifier based on the average accuracy for the binary classification problems used in this study was *k*-TSP (92.01%), followed by SVM (91.18%), PAM (88.91%), and TSP (88.26%). The differences in accuracies were small, so it was concluded that all four methods perform classification similarly. The authors also elucidate the biological meaning of the classifiers by showing the connections between the genes in the markers and their corresponding cancer types. For the multiclass problems, TSP achieved an average accuracy of 85.12% over 10 problems, somewhat less than PAM (88.50%) and SVM (88.10%), which performed the best overall but used hundreds or thousands of genes.

In the initial variant of RXA, Geman *et al*. [243] showed that the TSP classifier provides decision rules that are highly accurate in binary classification problems and involve very few genes. Tan *et al*. [116] compared the TSP and *k*-TSP approach to other machine learning techniques on a broad source of human cancer gene expression data. The performance of TSP and *k*-TSP on both binary and multiclass problems were comparable to those of the other techniques, while no single method was found to have the best performance across all datasets. TSP and *k*-TSP were thus shown to have comparable accuracy to state-of-the-art methods, involve fewer genes and yield transparent, context-independent classifiers that are invariant to most forms of data normalization.

Specific Cancer Studies Using TSP or k-TSP

TSP-based classification methods have been applied to a number of specific cases of predictive studies in cancer. These studies can be broadly divided into those that identify classifiers for disease diagnosis and studies that develop relative expression classifiers for disease prognosis. Specifically, diagnosis can refer to determination of the presence or absence of disease, the particular sub-type of a disease, or in some cases the stage of disease. In contrast, prognosis aims to predict the outcome of patients with the disease. Examples of disease prognosis include response to treatment, survival time, and tumor metastasis. Importantly, a number of the studies presented here demonstrate not only the power of TSP methods to accurately classify microarray profiles, but also their utility for integrating microarray datasets from different sources and even across different measurement technology platforms.

Gene-pair Classifiers for Diagnosis

Gastrointestinal stromal tumors (GISTs) and leiomyosarcomas (LMSs) are common mesenchymal tumors with similar phenotypic features. A whole-genome gene expression study of 68 well-characterized tumor samples identified a two-gene relative expression classifier using TSP that distinguished GIST and LMS with 99.3% accuracy on microarray samples and 97.8% accuracy in cross validation [258]. The classifier, which predicts GIST when OBSCN > C9orf65 and LMS otherwise, was validated using RT-PCR on 20 samples from the original dataset and on 19 independent samples, achieving 100% accuracy. Immunostaining for the Kit protein marker is currently the best test to differentiate GIST and LMS. Using expression of c-Kit to classify samples (with a cutoff determined by 1D linear discriminate analysis) achieved only 87.3% accuracy. That is, as some GIST samples have low Kit expression and some LMS samples have high Kit expression, testing for levels of the protein marker was more prone to error than predictions based on the OBSCN/C9orf65 expression ratio.

The TSP classification method is invariant to standard procedures for monotonic data normalization, as it relies only on the ranks of gene expression values within the microarray. As such, using TSP for classification enables the integration of microarray profiles from multiple datasets, thereby increasing the sample size of the training data and the predictive potential of the classifiers. Xu *et al*. [247] identified a TSP marker for prostate cancer (HPN > STAT6) that achieves high accuracy, sensitivity, and specificity on two datasets from different platforms. Performance of the HPN-STAT6 TSP marker—trained on integrated microarray data—was better than other TSP classifiers trained on individual datasets. In training the classifier, three microarray datasets from different prostate cancer studies were integrated and TSP was applied to analyze both individual and integrated datasets. It was found that TSP markers vary between individual datasets, but as more samples are added to the integrated training dataset, TSP selection becomes consistent. Stability analysis was also performed to calculate the appearance frequency of markers (i.e., how often the same TSP markers were selected) when samples were randomly removed from the dataset. The TSP marker was tested on an independent cross-platform dataset, comprising prostate tumor expression values from both Affymetrix and spotted cDNA platforms. Samples in the independent test set were classified with 93.8% accuracy, 91.7% sensitivity, and 97.7% specificity.

Gene-pair Classifiers for Cancer Prognosis

Xu *et al*. [259] integrated three independent microarray datasets containing 358 total samples for prediction of distant metastases in breast cancer. All samples in the integrated dataset were obtained from lymph-node-negative patients who had not received adjuvant systemic treatment. Gene expression data was directly merged using 22,283 probe sets on the Affymetrix HG-U133A microarray, and the top 200 "features" were selected as gene pairs with the highest TSP scores. In accordance with clinical treatment guidelines defined by the St. Gallen (Switzerland) expert consensus and the NIH, the goal of the authors was to achieve the highest possible specificity while maintaining high sensitivity (~90%). The optimal signature size (80 pairs, 112 distinct genes) was determined in *k*-fold cross validation, and a likelihood ratio test (LRT) for classification based on this signature achieved 88.6% sensitivity and 54.6% specificity in an independent external test set of 154 samples. Because the LRT assumes statistically independent gene pairs, the decision rule amounts to weighted voting among the gene pair classifiers and hence is very similar to *k*-TSP.

Over-expression of the Src tyrosine kinase in pancreatic cancer is thought to play a significant role in tumor development and progression. The *in vivo* efficacy of an orally active small molecule Src inhibitor AZD0530 was investigated in a collection of pancreatic tumor xenografts [260]. The *k*-TSP algorithm was applied to gene expression profiles from the tumors in order to identify predictive biomarkers of

response to AZD0530. Tumor growth index (TGI) was used to morphologically classify xenografts as sensitive (TGI < 50%) or resistant (TGI > 50%) to AZD0530 treatment. In the training set of 16 xenografts (3 sensitive, 13 resistant), the TSP classifier LRRC19 > IGFBP2 most accurately predicted cases as sensitive (and correspondingly predicted cases as resistant when LRRC19 ≤ IGFBP2).

The *k*-TSP classifier achieved an estimated LOOCV accuracy of 97.8% on the microarray dataset. The two-gene predictor was tested and validated on eight independent xenografts not included in the original training set and achieved an overall accuracy of 87.5%, specificity of 83.3%, and sensitivity of 100%. RT-PCR was performed on the two genes in the eight independent xenografts, showing the relative expression of LRCC19 and IGFBP2 was the same as measured by microarray gene expression in all cases. This stability across different measurement platforms is critical for application in the clinic and represents an advantage of methods based on RXA.

A two-gene expression ratio (RASGRP1/APTX) has been found that accurately predicts response to the drug tipifarnib in patients with acute myeloid leukemia (AML) [261]. The TSP algorithm was applied to transcriptional profiles of bone marrow samples from newly diagnosed AML patients—including 13 responders and 13 patients with progressive disease, achieving 92.3% sensitivity and 100% specificity (96% accuracy) in LOOCV. External validation of the two-gene classifier was performed in an independent dataset of 54 samples from patients with relapsed or refractory AML (10 responders, 44 with progressive disease). When applied to the independent test set, the classifier predicted tipifarnib response with sensitivity of 80% and specificity of 52.3%. This reduction in accuracy compared to LOOCV may derive from the initial very small sample set not being sufficient to represent the amount of variance in the population, and thus further data collection and classifier development is needed. Still, the results are encouraging considering the subtle difference of the phenotypes being considered and the small amount of training data.

In another study, Weichselbaum *et al*. [262] applied *k*-TSP to a previously determined gene expression signature—the IFN-related DNA damage signature (IRDS)—in order to develop a therapy-predictive marker of adjuvant chemotherapy for metastatic breast cancer. 78 breast cancer patients were divided into two IRDS status groups (IRDS(+) and IRDS(–)) using hierarchical clustering of microarray data. The *k*-TSP classifier was trained using 49 genes in the IRDS along with 534 previously-defined intrinsic breast cancer genes, with the optimal number of gene pairs determined using 10-fold cross validation. Each of the seven selected gene pairs in the *k*-TSP classifier contained one IRDS gene and a second gene for comparison. Classification was based on a majority vote, where samples were classified as IRDS(+) if expression of the IRDS gene was higher than the other gene in at least four of the seven pairs.

For the purpose of employing a non-binary measure for survival analysis, the number of positive-scoring gene pairs was used to define a TSP IRDS score. Specifically, the sum of pair-wise comparisons in which the IRDS gene was more highly expressed defined an ordinal scale from zero to seven, with seven representing the most IRDS(+)-like pattern. To examine the IRDS as a predictive marker for therapy outcome, a dataset of 295 patients with early stage breast cancer was analyzed based on the TSP IRDS score. A multivariable Cox proportional-hazards model for metastatic risk when an interaction with chemotherapy is considered revealed a hazard ratio of 1.2—signifying a 1.2-fold increased risk of metastasis for each incremental increase in the TSP IRDS score from 0 to 7. These statistically significant results suggested that an association of the IRDS with clinical outcome depends on the use of adjuvant chemotherapy.

Broad Application of TSP in Disease Diagnosis and Prognosis

A more recent study has shown that two-transcript classifiers have the potential to reliably classify diverse human diseases [263]. In this study, the investigators sought to assess the effectiveness of the TSP approach in the identification of diagnostic classifiers in a number of human diseases including bacterial and viral infection, cardiomyopathy, diabetes, Crohn's disease, and transformed ulcerative colitis through analysis of both local diseased tissue and the immunological response assayed through blood-borne leukocytes. The results of this study showed that several diseases of solid tissues could be reliably diagnosed through TSP classifiers based on the blood-borne leukocyte transcriptome. The TSP method identified multiple predictive gene pairs for each phenotype, with LOOCV accuracy ranging from 70 to nearly 100 percent. Performance compared favorably with that of pre-existing transcription-based classifiers, and in some cases approached the accuracy of current clinical diagnostic procedures. Thus, this study provided further evidence that the TSP classifier represents a simple yet robust method to differentiate between phenotypic states based on gene expression profiles of diverse human pathologies. The experimental simplicity of this method results in measurements that can be easily translated to clinical practice.

Beyond TSP and k-TSP

Top-Scoring Pair of Groups

In an effort to identify a robust common cancer signature, Xu *et al*. [117] performed a large-scale meta-analysis of cancer gene expression datasets in order to identify a universal cancer signature and validated their signature using a variant of RXA to separate cancer from normal samples across a wide range of cancers. More specifically, the authors integrated nearly 1,500 microarray gene expression profiles from

26 published cancer datasets across 21 major human cancer types using two different Affymetrix microarray platforms. Michiels *et al*. [264] had shown that molecular signatures are strongly dependent on the samples in the training data and advocated the use of repeated random sampling for signature validation. In [117], the authors applied an RXA method, referred to as the top-scoring pair of groups (TSPG) classifier, combined with a repeated random sampling strategy to identify of a common cancer signature consisting of 46 genes. The TSPG classifier is an extension of the TSP classifier from two individual genes to two groups of genes. Being an RXA method, it is based entirely on the internal ranking of the genes in the signature. The signature is divided into two disjoint groups, and the average rank is computed for each group and two averages are compared. The decision rule is again maximum likelihood; to choose the class for which the observed ordering between the two rank averages is most likely. It can also be shown that TSPG is a special case of *k*-TSP, where *k* is the product of the two group sizes. Given a new expression profile, the classifier was found to discriminate most human cancers from normal tissues, including a validation on six different independent test datasets generated from different Affymetrix microarray platforms. Upon further validation, this cancer signature may be used to improve understanding of cancer pathogenesis and therapeutic targets, and hence lead to the development of effective treatment regimens.

Top-Scoring Triplets

Lin *et al*. [244] proposed an extension of TSP which bases prediction entirely upon the relative expression ordering among three genes, referred to as the "top-scoring triplets" (TST). The decision rule is to select the class, which makes the observed ordering the most likely. In many cases, one gene serves as a "reference" whose expression falls between the expressions of two differentially expressed genes. The objective is to achieve a more discriminating decision mechanism than TSP but without sacrificing interpretability. The investigators explored the different roles the three genes play in the decision mechanism from previous cancer studies, and also applied this methodology to two problems in breast cancer: a cross study validation based on predicting ER status and a clinically relevant application to predicting germ-line BRCA1 mutations. Further analysis on protein-protein interactions among the triplets of genes aided in understanding the biological roles of the classifiers.

Conclusions and Future Directions

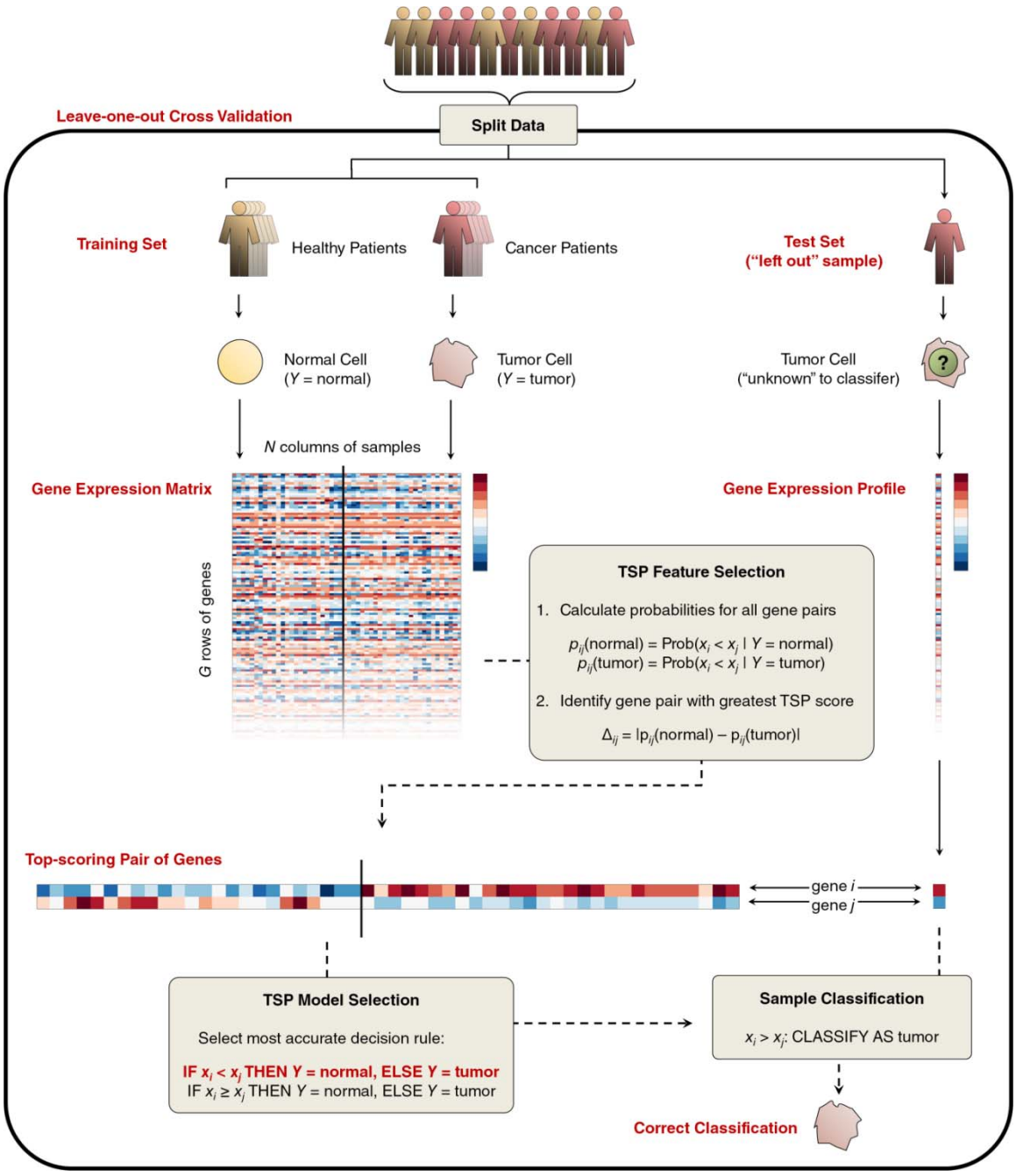
The advent of high-throughput measurement technologies for the comprehensive, rapid, and inexpensive detection of biomolecular signatures in human cells, tissues, and serum has led to the generation of a tremendous amount of raw, unprocessed information. However, analyzing and interpreting these data in

order to enhance our understanding of human health and genetic diseases (i.e., cancer) continues to be a challenge in the scientific community. In the case of gene expression microarray data, standard statistical learning methods have been used to identify decision rules that can accurately distinguish disease phenotypes. These techniques have been shown to produce accurate classifiers, but still lack the convenience and simplicity desired for extracting any underlying biological rationale for the decision rules.

In this chapter, I have provided a detailed description of the concepts and methodologies of the TSP and *k*-TSP classifiers, two bioinformatics techniques for gene expression-based molecular classification based on the analysis of relative expression values. Due to the simplicity of the classifier and ease of biological interpretation, as well as its independence to data normalization and parameter fitting, the TSP and *k*-TSP methods have been applied in several studies to perform molecular classification of various pathologies, primarily cancer**.** These methods, as I have shown above, display highly accurate classification performance in distinguishing a broad range of disease phenotypes (e.g., cancer vs. normal), cancer subclasses (e.g., AML vs. ALL, GIST vs. LMS), disease outcomes (e.g., metastasis, survival), and diverse human pathologies assayed through blood-borne leukocytes. I have also shown that natural extensions of the basic TSP and *k*-TSP methods can incorporate more genes and allow for indirect microarray data integration and hence large-scale meta-studies. Further work on RXA includes the use of biological network information (e.g., as described in **Chapter 4** and **Chapter 5**) for phenotype classification and biological discovery as well as decision tree-based strategies for classification of multiple disease phenotypes.

Chapter 3 Figures

Figure 3.1. Schematic overview of phenotype classification with the top-scoring pair (TSP) algorithm in cross validation.



Chapter 4. Identifying Tightly Regulated and Variably Expressed Networks by Differential Rank Conservation (DIRAC)[4]

A powerful way to separate signal from noise in biology is to convert the molecular data from individual genes or proteins into an analysis of comparative biological network behaviors. One of the limitations of previous network analyses is that they do not take into account the combinatorial nature of gene interactions within the network. In this chapter, I report here a new technique, Differential Rank Conservation (DIRAC), which permits one to assess these combinatorial interactions to quantify various biological pathways or networks in a comparative sense, and to determine how they change in different individuals experiencing the same disease process. I drove the development and implementation of DIRAC and analysis of results, with constructive input from Don Geman from Johns Hopkins University and biological insights from Lee Hood at the Institute for Systems Biology. This approach is based on the *relative* expression values of participating genes—i.e., the ordering of expression within pathway profiles—and is part of the RXA family of methods described in **Chapter 3**. DIRAC provides quantitative measures of how network rankings differ either among networks for a selected phenotype or among phenotypes for a selected network. I examined disease phenotypes including cancer subtypes and neurological disorders and identified networks that are tightly regulated, as defined by high conservation of transcript ordering. Interestingly, I observed a strong trend to looser network regulation in more malignant phenotypes and later stages of disease. At a sample level, DIRAC can detect a change in ranking between phenotypes for any selected network. Variably expressed networks represent statistically robust differences between disease states and serve as signatures for accurate molecular classification, validating the information about expression patterns captured by DIRAC. Importantly, DIRAC can be applied not only to transcriptomic data but also to any ordinal data type.

Network Expression Analysis and Differential Rank Conservation

Molecular signatures based on the measured abundance of biomolecules (e.g., mRNA, proteins, metabolites) have the potential to discriminate among disease subtypes, to predict clinical outcomes, or to provide insights into the mechanistic underpinnings of disease progression. Moreover, with sufficient data, these signatures begin to enable the identification of perturbed networks that reflect core aspects of the disease process—and thus could provide insights into functionally relevant drug targets as well as new approaches to diagnostics [7, 139]. However, distinguishing signal from noise in high-throughput data such as mRNA microarray experiments presents a significant challenge. This noise commonly results

[4] This chapter includes material that was reproduced with permission from the following publication: Eddy, J.A., N.D. Price, and D. Geman. 2010. Identifying tightly regulated and variably expressed networks by Differential Rank Conservation (DIRAC). PLoS Computational Biology. 6(5):e1000792. (all sections).

from technical issues in data production and the integration of datasets from different platforms, laboratories, or even experiments within a lab. Noise in high-throughput data also stems from biological variability in the sources, such as genetic polymorphisms, different stages of the biological process, disease stratification, and stages of disease progression. In the study of human disease processes, this variability poses a unique hurdle as there are often only data for a single point in time; when comparing data between individuals who appear to have the same disease, one does not know whether the observed differences reflect disease subtypes or different stages for a single disease type.

A fundamental tenant of systems approaches to biology and medicine is that dynamically changing biological networks mediate physiological, developmental, and disease processes, and that the key to understanding these processes is translating network dynamics into phenotypes. As such, a powerful method to mitigate some forms of biological noise (hence increasing the utility of high-throughput data as a diagnostic and scientific tool) is to convert the molecular data from individual genes or proteins into an analysis of comparative biological network behaviors. Typically, studies search for a small number of individual genes whose differential expression is highly correlated with phenotypic changes. However, malignant phenotypes in many diseases arise from the net effect of interactions among multiple genes and other molecular agents within biological networks. For example, cooperating oncogenes interact synergistically to evade tumor suppression mechanisms such as cell-cycle arrest and apoptosis [2, 5]. The combinatorial nature of such disease-induced perturbations leads to a highly complex picture of the underlying biological processes. As such, the biological insight gleaned from the expression patterns of individual genes is often limited. Other pitfalls associated with individual gene expression analysis have been proposed and discussed elsewhere [7, 239, 265].

The importance of studying network behavior—evident in most phenotypes, disease or otherwise—is particularly well documented for cancer. Research has linked modulated function on the level of either metabolic networks [41, 43, 266] and/or signaling networks [267-269] to cancer hallmarks including angiogenesis, increased growth, metastasis, and evasion of immune detection. Similarly, recent global genomic analyses in glioblastoma multiforme [270, 271] and pancreatic cancers [272] have revealed both varying numbers and frequencies of genetic alterations within distinct core networks of each disease. In light of these findings, microarray data analysis methods have begun to shift towards identifying biologically meaningful pathways or networks. One can consider all pathways to in fact be part of interconnected biological networks, and henceforth in this chapter, I use the term network rather than pathway. In general, network regulation controls the expression levels of related genes responding to specific conditions. Existing tools for network-based expression analysis commonly investigate informative patterns of up-regulation or down-regulation (i.e., increases or decreases in expression) of

genes in different disease states. For example, the widely used gene set enrichment analysis (GSEA) platform identifies networks that are significantly enriched for individual genes that are highly correlated with a phenotype [128, 239]. Other methods employ a single statistic to represent the collective activity of a network (e.g., mean or median gene expression) [7, 134]; perturbed levels of network activity (i.e., collective up- or down-regulation) are then examined to identify those networks most differentially expressed between phenotypes. These frameworks have been applied to diverse cancer systems and serve as a robust source of biological discovery [7, 9].

Studying cellular regulation of networks in terms of "unidirectional" changes may, however, overlook subtle, yet influential, changes in the *relationships* among the genes within a network. This drawback directly reflects the combinatorial operation of genes in networks, in which the actions of one gene greatly influences the actions of other genes. By accounting for these combinatorial interactions we can begin to alleviate the signal-to-noise issues in disease-perturbed networks (as well as dynamically changing networks mediating physiology or development). In particular, even the elementary interactions captured by the relative orderings among two or three genes have been shown to provide powerful biomarkers for separating phenotypes [261, 262, 273]. With methods that aim to identify statistically significant up- or down-regulation of genes or networks, results will also depend largely on the context of the microarray experiment. Cellular regulation in a case with a number of up- or down-regulated genes in one phenotype versus another manifests as an increase in absolute expression levels above some threshold, relative to all other genes on the microarray. Even when thresholds are tuned to produce statistically significant results, the findings are still based on indirect measurements, (i.e., fluorescence) and therefore may depend heavily on the experimental set up, type of data normalization and other factors. In addition to the technical limitations of microarray experiments, biological context can greatly impact results. For instance, if nearly all genes are differentially expressed between two phenotypes, then no single network will be statistically "enriched" for change. It is also possible that neither individual network genes nor any network as a whole will display notable over- or under-expression in response to environmental or disease-related stimuli. The importance of accounting for combinatorial gene interactions again becomes clear, and to do so without need to reference all of the genes on the microarray.

I led the development a new method called Differential Rank Conservation (DIRAC), which considers combinatorial behavior and provides quantitative measures of how network expression differs within and between phenotypes. The DIRAC approach assesses cellular regulation of a network in the context of the *relative levels of expression* for participating genes. For each microarray, the expression values of the network genes are ordered from highest expression (ranked first) to lowest expression (ranked last);

regulation is then quantified entirely by the *rankings* of genes within a selected network. Consequently, DIRAC identifies and measures network-level perturbations from a completely novel perspective, namely the "combinatorial comparisons" of network genes as opposed to increases or decreases alone, allowing one to study how this ordering changes in different conditions—and thus begin to infer the consequences of combinatorial gene interactions. As a result, this approach has two key advantages over tools that measure absolute changes in expression levels. First, it accounts for gene-gene interactions; second, the results do not depend on the other genes on the microarray or on the method of normalization used. These are both critical points in dealing with signal-to-noise issues. Notably, as DIRAC treats each network independently, it can still identify perturbed networks even when every gene on the microarray is differentially expressed (in contrast to enrichment measures).

The DIRAC strategy for representing network rankings uses pairwise comparisons of gene expression levels. Such pairwise comparisons can yield two-gene predictors with simple decision rules for classification of expression profiles [116, 274]. These decision rules have in turn resulted in highly accurate two-gene diagnostic classifiers based on relative expression reversals that have proven effective for molecular identification of cancer [116, 261, 262, 273, 274]. DIRAC extends the relative expression reversal concept to networks. However, analyzing sample-to-sample changes for every possible distinct ordering of gene expression values within a network is not computationally feasible; there are simply too many possible orderings, i.e., permutations. Knowing the states of all pairwise orderings is equivalent to knowing the full ranking, which motivates this particular representation. For each distinct pair of genes within a network, we consider a binary variable indicating whether or not the mRNA abundance of the first gene is less than that of the second gene; in fact, we restrict attention to the probability of this event within a phenotype for each pair of genes. In this way, we avoid the combinatorial complexity of permutations and represent the "expected" ordering of network genes for a given phenotype as a binary template. Unlike the probabilities of full orderings, pairwise frequencies are reliably estimated with typical sample sizes, while still capturing a great deal of information about network regulation. We subsequently compute a *matching score* to signify how closely each sample's network ordering matches a phenotype-specific template.

We can use DIRAC at the population level to quantify conservation differences between networks for a given phenotype. Specifically, DIRAC allows us to use rankings to identify and contrast tightly and loosely regulated network types of a single phenotype:

i. a network is considered *tightly regulated* within its phenotype if the ranks of network genes are mostly unchanged among samples;

ii. a network is considered *loosely regulated* if the ranks of network genes are greatly varied between samples of the same phenotype.

Tightness of regulation for a selected network is best understood as the allowed variation in gene expression levels observed across the population. This offers an advantage over studying up- or down-regulation only because it indicates the level of control across samples in a population. In this work we use the DIRAC approach to identify networks that are tightly regulated in a number of human cancers neurological disorders. As networks under tight control in a particular phenotype may be necessary to maintain a specific cellular function, tightly regulated networks that change across phenotypes may provide insight into processes such as disease progression.

Additionally, DIRAC can be applied at the sample level to identify conservation differences between phenotypes for a specified network. At this level the DIRAC method can identify variably expressed networks that reveal statistically robust differences between disease states, leading to highly accurate classification of expression profiles from various diseases. When used to separate expression profiles the DIRAC method is noteworthy because it (i) is independent of microarray data normalization; (ii) results in a simple yet efficient classifier for phenotype distinction; and (iii) appears to be comparable in accuracy to state-of-the-art classification methods. Learning the regulation of gene rankings within different states allows us to discover molecular signatures composed of related genes that distinguish phenotypes, identify networks most involved in disease transitions, and assist identification of potential therapeutic targets. Importantly, while I focus on gene expression in the present study, the method can be generalized to any ordinal dataset, and thus can be applied to such biological data types as proteomics, gene copy number, chromosomal position, and so forth.

Overview of DIRAC Methods

The DIRAC approach was used to evaluate regulation of gene ordering within networks in different diseases. For each microarray sample in each phenotype studied, we characterized the ordering of network genes (i.e., network ranking) in terms of comparisons between the expression values of pairs of genes. Based on the comparison statistics, we defined a *rank template* for each network and phenotype representing the expected (i.e., most common) pairwise ordering of gene expression for that network in that phenotype. We employed a simple measure—a *rank matching score* (R)—to determine how well the network ranking in each individual sample (i.e., expression profile) matched the ordering defined in the rank template. Averaging R over all samples within a phenotype yields a network-specific *rank conservation index* (μ_R), which represents how well, *on average*, all samples in the same phenotype match

the corresponding rank template. Alternatively, comparing two rank matching scores for the same sample leads to a highly discriminating *rank difference score* (Δ) that allows one to determine the most variably expressed networks between two phenotypes. The calculation of these quantities is illustrated in **Figure 4.1**.

Several prototypical scenarios arise from these measures. In one scenario (**Figure 4.2, left**), conservation indices are used to measure the consistency with which network rankings are maintained in a population, and are used to identify tightly regulated networks in each phenotype. One situation, where all samples have similar network rankings, yields a large rank conservation index and indicates the network is tightly regulated. A second situation, where the ordering of network genes is highly varied, yields a small rank conservation index and indicates the network is loosely regulated. In a second prototypical scenario, the DIRAC method detects changes in ranking (i.e., shuffling of gene expression values) between phenotypes for a selected network (**Figure 4.2, right**). The top networks selected by DIRAC based on the difference score can be used to classify gene expression profiles by phenotype.

We first applied DIRAC to investigate network rankings using gene expression profiles obtained from patients with different stages of prostate disease. The gene expression data, originally reported by Yu et al. [275] and publically available in the NCBI Gene Expression Omnibus (GDS2545), contains 108 human prostate samples: 18 samples of normal prostate tissue (NP) from organ donors, 65 primary prostate tumor (PT) samples, and 25 metastatic prostate tumor (MT) samples. The findings for normal prostate and prostate cancer samples presented below represent the main features of the DIRAC method and can be similarly obtained for any disease expression data.

In addition to the more detailed prostate cancer analysis, we examined a number of other disease phenotypes including cancer subtypes and neurological disorders, and we identified tightly regulated and variably expressed networks in each. For each dataset, we grouped expression levels of genes into 248 human signaling networks, defined according to the BioCarta gene sets collection in the Molecular Signatures Database (MSigDB) [239]. In order to ensure that the networks examined were as complete as possible, we used gene synonym information from NCBI to replace unmatched names in each dataset with those belonging to networks in the BioCarta collection. This step led to an average increase of 5% in the fraction of network genes (1296 total across 248 networks) for which a corresponding expression value was found.

Population-Level DIRAC

The population-level analysis is centered on the rank conservation index (μ_R), defined for each network and each phenotype. This index represents the degree of conservation in the rankings of the expression levels of the network genes, averaged over samples of the phenotype.

Tightly Regulated Networks in Normal Prostate and Cancerous Prostate

For a given phenotype, the extent of gene ranking conservation within networks will vary across networks. The ten most tightly regulated networks in normal prostate (NP), primary prostate tumors (PT), and metastatic prostate tumors (MT), as measured by rank conservation indices, are shown in **Table 4.1**. Large rank conservation index values indicate similar gene orderings among all samples of each phenotype in these networks, and hence tight regulation. This suggests that the combinatorial gene interactions in each network are quite similar among different patients.

Identifying networks that are tightly regulated in some phenotypes and loosely regulated in others suggests that the level of control across samples in a population may change dramatically, reflecting the nature of the disease process. While identifying changes in tightness of regulation of networks can provide insight into molecular differences between phenotypes, some networks may be tightly regulated in all phenotypes examined. For example, we found that the G-protein signaling (GS) network is the most tightly regulated network in normal prostate (NP), as well as in primary (PT) and metastatic prostate tumors (MT). The GS network comprises major signaling proteins downstream of G-protein coupled receptors, including both the catalytic (*PRKACA*) and regulatory (*PRKAR1A*) subunits of the cAMP-dependent protein kinase C (*PKC*). PKC family members phosphorylate a wide variety of protein targets and are known to be involved in diverse cellular signaling networks, such as those associated with cell adhesion, cell transformation, cell cycle checkpoint, and cell volume control. In 18 NP samples, the pairwise orderings among the six GS network genes matched the corresponding normal prostate rank template identically for all 15 pairs in the network (μ_R = 1.000). Similarly, network rankings in PT samples and MT samples matched the respective templates for 98.9% (μ_R = 0.989) and 99.5% (μ_R = 0.995) of all pairwise orderings on average. We also found that a single network ranking was shared by the majority of NP samples (100%), PT samples (83%), and MT samples (92%); in particular, therefore, the GS network rank template was identical in all three phenotypes. Furthermore, the remaining samples in PT and in MT displayed only a single mismatch in pairwise orderings compared to the template.

There are several possible explanations for observing tight regulation of certain network rankings in a phenotype. In the simplest case, the genes in a network may be expressed at greatly disparate magnitudes, making a change in their relative expression rankings less likely. We can see that this is most likely true for the GS and FOSB networks, both of which displayed the highest rank conservation for all three

prostate phenotypes. The average gene-to-gene expression variance across all samples for these networks fell between 1.14-1.58, roughly three times the average gene-to-gene variance for all 248 networks (~0.41). As such, a change in the relative ordering among genes in these networks would require a more dramatic change in the expression of individual genes. Networks like GS and FOSB are therefore analogous to "housekeeping" genes, as the ranking of genes in each is expected to remain the same in most samples.

Alternatively, small variation in ordering—nearly the same ranking in all samples of the same phenotype—could indicate that a network is critical to maintaining some specific cellular function. This is more likely in cases with less gene-to-gene expression variance within a network; if pairwise orderings can be easily altered by small changes in expression but remain consistent, some force such as selective pressure might drive the cell to minimize fluctuation in the expression of network genes. We found that the SET network is tightly regulated in NP samples, but displays much smaller gene-to-gene variance than networks like GS and FOSB. The SET network—also known as the granzyme mediated apoptosis pathway—comprises a total of 11 genes (illustrated in **Figure 4.3**), and is an important cytotoxic T cells mechanism for fighting tumors and virus-infected cells [276]. While the SET network displays greater variation in ranking among NP samples than GS or FOSB (μ_R = 0. 945), 16 out of 18 samples show only five or fewer mismatches compared to the 55 pairs in the rank template. We hypothesize that expression of genes within the SET network is highly consistent in NP samples to maintain proper function of cellular defense mechanisms.

Tightly regulated networks in disease phenotypes might also lead to useful hypotheses about cell behavior. The RAN network, similar to SET, is tightly regulated across MT samples, and shows relatively low gene-to-gene variation within the network. The RAN network contains five genes: regulator of chromosome condensation (*CHC1*), Ras-related nuclear protein (*RAN*), RAN binding proteins 1 and 2 (*RANBP1* and *RANBP2*), and RAN GTPase activating protein 1 (*RANGAP1*). In MT samples, on average, the pairwise orderings among the five RAN network genes matched the corresponding MT rank template for 96.0% of all pairs in the network (μ_R = 0.960). This network is involved in the export of mRNA transcripts from the nucleus to the cytosol for subsequent translation. Although it is unclear what advantage tight regulation of the RAN network may confer upon metastatic prostate tumors, there is clearly little variation in network ranking. Importantly, the mutation rates in cancer cells are increased 200-400 fold—providing ample opportunity for changes to be fixed by natural selection or random fixation (if the change is not selectively advantageous or disadvantageous).

We can learn more by examining the tightness of regulation for the same network in different phenotypes. The SET network in PT samples has a rank conservation index equal to 0.909, which is significantly lower than in NP samples (P-value < 0.05); similarly, μ_R for SET in MT samples is equal to 0.891. As seen in **Figure 4.3**, the decreased network rank conservation in PT and MT is due to a greater number of samples with rankings different from the respective templates (i.e., more samples with greater numbers of mismatches). The increased variation in network ranking seen in the two stages of prostate cancer might indicate that the biological function associated with SET genes (i.e., granzyme mediated apoptosis) plays a lesser role in behavior of these cells, and is therefore under looser control. It is also possible that in primary and metastatic prostate tumors—obviously more malignant phenotypes compared to normal prostate—the SET network becomes *deregulated*, and that this higher deregulation contributes to its malignancy. Alternatively, an increase in mutation rates with malignancy might have resulted in more random fixations.

These rank conservation indices estimate population statistics based on limited sample sizes (on the order of 20-100, as seen in **Table 4.2**), and hence some variation is expected even if the true indices were the same. For instance, the difference in the rank conservation index for the SET network between NP and PT or between NP and MT could be a small-sample effect and would need to be validated with either a larger study or by a suitable permutation test (see ***Deregulation of Network Ranking in Disease*** section below). However, even these small-sample estimates generate specific hypotheses, such as an increase in disorder in the more malignant phenotype, which can then be meaningfully explored by examining a variety of datasets and phenotypes—discussed in the following sections.

Deregulation of Network Ranking in Disease

As described for the SET network above, certain networks may be tightly regulated in one phenotype, but not in another. The SET network appears to be relatively tightly regulated in normal prostate but more loosely regulated in both primary and metastatic prostate tumors. Cases such as this represent the deregulation of a network in one phenotype relative to another. We used the difference in rank conservation index values between phenotypes as the basis for identifying the most deregulated networks. For example, in comparing NP samples to MT samples, we first calculated the rank conservation index for all networks in both phenotypes. Next, we identified the networks with the greatest absolute difference in index values between NP and PT (i.e., highly conserved in one class but not in the other). Based on sample permutation tests, we found that 67 out of 248 networks had a significant difference in index values (P-value < 0.05; see **Detailed Methodology**). The network with the largest conservation difference—the FIBRINOLYSIS network—was more tightly regulated in NP (μ_R = 0.891) than in MT (μ_R

= 0.736) (**Table 4.3**). The FIBRINOLYSIS network comprises 12 genes and breaks down fibrin clots formed during coagulation. It has previously been reported that patients with metastatic prostate cancer occasionally exhibit enhanced fibrinolytic activities with symptoms of bleeding, epistaxis or other forms of hemorrhage [277]. Deregulation of the FIBRINOLYSIS in MT samples might therefore be directly linked to malignant features of the disease. However, without further information it is impossible to discern whether loose regulation of this network is a causative mechanism in MT, or occurs as a downstream effect of some other perturbation in tumor progression.

Upon inspecting the remaining differentially regulated networks between NP and MT, we found that in fact, 57 out of 67 significantly deregulated networks identified showed tighter regulation in NP than in MT (**Figure 4.4J**). The strong majority of networks more tightly regulated in the NP (*P*-value = 5.14×10^{-8} from a binomial distribution; see **Table 4.4**) lends evidence to the theory that deregulation of network ranking is in some way related to increased malignancy. As such, the DIRAC approach may be useful both in the stratification of disease and/or in assessment of the progression of disease. To explore this hypothesis further, we examined a number of gene expression datasets available for public download from the NCBI Gene Expression Omnibus (**Table 4.2**). These datasets included expression profiles from multiple cancers such as breast, ovarian, and blood (leukemia), as well as diseases of the brain/nervous system, skin, and intestinal tract (note: the leukemia datasets G and H were excluded from this particular comparison, as there is no clear evidence for which subtype—AML or ALL—is more malignant). We repeated the procedure described for NP and MT for each binary phenotype comparison from the expression data. In all but one case out of nine, the less malignant phenotype had a greater number of high conserved (tightly regulated) networks (**Figure 4.4**). That is, a network appears much more likely to become deregulated in worse cases of disease. Importantly, the dataset for the one exception—comparing Marfan syndrome and normal fibroblasts—contained expression values for only ~4,000 genes (compared to 20,000 or more in most of the other datasets). Due to the small number of genes, many of the networks contained significant caps, which may have produced less robust results. Still, the overall trend seen in **Figure 4.4** suggests that in malignant phenotypes, networks are often more loosely regulated, with greater variation in expression ranking of participating genes from sample to sample. The global pattern of increased disorder with malignancy highlights the utility of studying gene expression ordering within networks, and also reveals a striking phenomenon that could drive future investigation and may lead to new understandings of gene expression in disease.

Global Regulation of Networks Across Phenotypes

Averaging rank conservation indices over all the networks provides a measure of global regulation of networks in different phenotypes. For example, networks in normal prostate are more highly conserved on average (0.903) than networks in metastatic prostate cancer (0.884). This difference suggests that the more malignant cancer subtype (MT) may have greater overall variation in network rankings among different samples. We used the gene expression datasets described above to compare global regulation of network rankings among a number of phenotypes. For each phenotype, we calculated rank conservation indices for all networks and used the average conservation as a rough measure of how tightly or loosely regulated networks *tend* to be in each case.

We used the average index value to order phenotypes from highest to lowest global conservation. Phenotypes with the highest average conservation primarily have tightly regulated networks across samples in the population. For example, most networks in non-bipolar cortex and bipolar cortex were found to have conservation index values greater than 0.95 (seen as bright colors on the heatmap in **Figure 4.5**) for average values of 0.956 and 0.955, respectively. In contrast, many networks in the two breast cancer phenotypes (r—responsive to treatment; nr—non-responsive to treatment) have rank conservation indexes less than 0.80 (dark colors on the **Figure 4.5** heatmap). In this case, the low global conservation—average index values of 0.835 and 0.826 in (r) breast cancer and (nr) breast cancer, respectively—suggests that network rankings in these disease phenotypes have looser regulation and greater variation. Based on a one-way ANOVA, the estimated overall *P*-value for the ordering of phenotypes in **Figure 4.5** is zero.

Interestingly, the trend of lower conservation in more malignant phenotypes described in the previous section seems to persist even from a coarser, global perspective. For example, networks in the less malignant adenoma-like ovarian tumors are more highly conserved on average (0.947) than in more malignant carcinoma-like ovarian tumors (0.913). The same was seen when examining all three prostate phenotypes, where normal prostate is more tightly regulated overall than primary (p) prostate cancer, which itself is more tightly regulated than metastatic (m) prostate cancer. Even for the most highly conserved phenotypes (non-bipolar and bipolar cortex) and lowest conserved phenotypes (breast cancers), networks are more tightly regulated on average in the less malignant phenotype of each pair. We also observed interesting differences observed based on tissue-type, where expression ranking of networks in brain and ovarian tissue displayed higher conservation on average than prostate tissue, which is in turn is more highly conserved than in blood and in breast tissue. Thus, at least two global trends must be considered in evaluating network deregulation: (i) the severity of the disease, and (ii) the tissue of origin.

Sample-Level DIRAC

In order to identify variably expressed networks between two selected phenotypes, we designed a rank difference score (Δ), calculated for each sample based on rank matching scores. For a particular network, this measure indicates the similarity between the ordering of network genes in a sample to the template of one class versus the template of the other. The difference score ranges from -1 to 1, with positive values suggesting the first phenotype, and negative values suggesting the second, culminating in simple rules for classifying an expression profile. Our purpose in introducing the rank difference score was two-fold: (i) to identify variably expressed networks between two selected phenotypes; and (ii) to validate the DIRAC approach to network identification, and the emphasis on combinatorial interactions, by demonstrating the discriminative power of the networks identified.

Variably Expressed Networks in Prostate and Prostate Cancers

As shown in **Figure 4.6**, this positive versus negative trend holds for most samples in MT and NP across all networks. To determine the most variably expressed networks between MT and NP, we (i) defined rank templates for each phenotype; (ii) calculated rank matching scores and evaluated the rank difference score for each sample; and (iii) chose the networks with the largest estimated classification rate. Specifically, the classification rate for a network is defined as the average of sensitivity and specificity for predicting sample classes in the training data (i.e., apparent accuracy).

To evaluate whether variably expressed networks represented meaningful differences between MT and NP gene expression profiles, we used permutation-based testing to assess the statistical significance of estimated network classification rates (see **Detailed Methodology**). A total of 176 networks significantly differentiated between expression profiles of MT and NP (P-value < 0.05), the top ten of which appear in **Table 4.5**. Among these differentially expressed networks, we estimated that only 6.7% (between 11 and 12 of the 176 total) are likely to have been found by chance rather than based on true differences between the phenotypes, as determined by the FDR.

The principal features governing the rank difference score, and also an example of its application to molecular classification, are illustrated in **Figure 4.7** for the MAPK network, which we identified as one of the most differentially expressed networks between normal prostate and metastatic prostate tumors. Here, $R(\mathbf{x}_n)$ denotes the rank matching score for a profile $\mathbf{x}_n$, and superscripts indicate the network and phenotype of the rank template (e.g., $R^{(\text{MAPK,MT})}(\mathbf{x}_n)$ represents the rank matching score for a sample when compared to the ordering defined in the MT template). The rank difference score is the difference in matching score values for a particular sample: $R^{(\text{MAPK,MT})}(\mathbf{x}_n) - R^{(\text{MAPK,NP})}(\mathbf{x}_n)$. This measure captures low

variance of network ranking within phenotypes, but disparate rankings between phenotypes. The rank difference values calculated for the MAPK network for all samples are shown in **Figure 4.7**, along with the corresponding phenotype predictions (i.e., MT where positive, NP if negative). Interestingly, MAPK signaling has been previously reported to be involved in the cancerous transformation of prostate cells [278, 279].

DIRAC-based Classification of Disease Phenotypes

The top networks selected by DIRAC based on the difference score (i.e., the single best network for separating each different pair of phenotypes) were used to classify gene expression profiles in cross-validation. Specifically, we used leave-one-out cross validation (LOOCV) to estimate how accurately the top networks selected predicted the phenotype of future samples (**Figure 4.8**). Importantly, all processes including defining rank templates, calculating rank difference scores, and selecting the best network were done within cross-validation, using only the training samples (i.e., no information from test samples was used to train classifiers). For comparison, we selected the top G_m differentially expressed genes—where G_m is equal to the number of genes in the top network selected by DIRAC—and used the top-scoring pair (TSP) algorithm [116, 245] and support vector machines (SVM) [280, 281] to classify samples in each of the datasets. We found that our method performed well in a number of the datasets, with estimated accuracies between 92-96% in gastrointestinal sarcoma, ovarian cancer, leukemia, and prostate cancer--including comparisons between normal prostate and cancer as well as different stages of prostate cancer (**Figure 4.8**). In cases with poor accuracies, such as responsiveness of breast cancer to therapy, bipolar disorder, and Marfan syndrome, we observed that other methods also failed to accurately classify samples, suggesting that these phenotypes are inherently difficult to separate based on the available expression data.

Overall, we found that classification, when restricted to only the genes in the top network (as determined by DIRAC), is nearly as accurate as using the overall G_m most differentially expressed genes (in TSP or SVM). Our foremost goal was not to propose a new classifier, but to aid in biological discovery and hypothesis generation; the classification accuracy simply affirms the robustness of the network rank regulation measure. Specifically, the classification experiment validates DIRAC by demonstrating the importance of combinatorial interactions: the potential loss of discriminating power in individual genes is countered by discriminating interactions.

Implications for Systems Medicine

Systems medicine approaches assume that disease arises from disease-perturbed biological networks in the relevant organ or organs. These disease-perturbed networks alter the envelopes of information that they express—and these changes encode the pathophysiology of the disease. Moreover, the altered patterns of information can elucidate new strategies for diagnosis or therapy. Future drugs will likely be designed to re-engineer disease-perturbed networks to behave in a more normal fashion, or at least to abrogate their most deleterious consequences. This will require a new drug target identification approach, and re-engineering disease-perturbed networks appropriately will almost always require multiple drugs. Likewise, the perturbed nodal points in disease-perturbed networks can be expressed as proteins in the blood—where the disease-altered levels of expression may reflect the disease process. These disease-altered blood proteins will create unique blood fingerprints specific for each disease process, and thus provide powerful diagnostics. These advances rely upon the proper identification of disease-perturbed networks. To date, most of the evaluation of networks has employed lists of transcripts that are perturbed from the levels of their counterparts in normal organs. This listing, as with genome-wide association (GWAS) studies, misses the key fact that disease-perturbed networks must be assessed in the context of the combinatorial interactions of their nodal components.

Our method is the first approach that begins to account for the combinatorial behavior of interacting genes, mRNAs and/or proteins. Using DIRAC-based calculations allows us to begin to assess the key disease-perturbed networks that may aid in the approach to diagnosis and therapy. We also stress that these methods will almost certainly prove powerful in the stratification of disease types. The example of gastrointestinal stromal tumors (GIST) and leiomyosarcomas (LMS), histologically indistinguishable, but clearly classifiable by a primitive version of DIRAC, is striking. We believe this will be a powerful approach in, for example, distinguishing various types of neurodegenerative diseases, as well as the stratification of complex diseases such as Alzheimer's. Notably exciting, some of the key transcripts used in this classification process actually encoded proteins secreted into the blood. Findings of this nature could lead to the use of altered blood levels of proteins for diagnosis without the need to sample disease tissues. Emerging technologies will make these measurements possible at the single cell level, exposing other exciting possibilities for diagnosis using the strategies outline above. We predict the application of DIRAC as a powerful clinical tool in the advancing proactive, rather than reactive, new medicine—the so-called P4 medicine (predictive, personalized, preventive and participatory)—where blood and single-cell diagnostics will be the foundation of the P4-medicine revolution.

Conclusions

In this study, we demonstrated a novel method to identify highly discriminative biological networks based on differing patterns of gene expression ranking within networks. These results provide a coarse, but meaningful, glimpse into patterns of network regulation for different phenotypes based on combinatorial relationships between the involved genes. For example, when comparing two disease states, it appears to be very common (although not universal) for network rankings to be more varied—or less tightly regulated—in the more pathological disease. This increased disorder associated with malignancy might be expected, as mutations and other altered behavior of biomolecules lead to breakdown of typical functioning in biological networks. Rank conservation index values calculated in DIRAC represent a quantitative means to study and further verify this notion. Importantly, this method not only identifies perturbed networks, but does so in such a way that it can classify samples. Thus, predictive accuracy becomes a strong measure for the validity of the perturbed network as a reproducible hallmark of the disease phenotype. Such high predictive accuracy in classification adds much stronger evidence that biologically meaningful network differences are found than only a low *P*-value or FDR, which simply measure how likely the result derives from chance. Measures of global regulation can also give useful information for designing research to identify expression-based classifiers of disease. For instance, it would be more fruitful to search for clear molecular signatures with tightly regulated phenotypes. In cases with mostly loosely regulated networks, the greater variation from sample-to-sample would pose a more difficult challenge for identifying reliable classifiers. Studying rank regulation of biologically relevant networks thus offers a promising tool for measuring network behavior within and across different populations. Looking forward, the results obtained through this approach should provide increased insight into phenotypic processes of importance in biology and medicine.

Detailed Methodology

The methods and analyses presented here were performed entirely in Matlab. Source code files are available for download at http://www.igb.uiuc.edu/labs/price/downloads.

Microarray Data

Given the list $\{g_1, \ldots, g_{Gm}\}$ of G_m genes within a network m on a microarray, we let $\mathbf{X} = (X_1, \ldots, X_{Gm})$ denote the corresponding expression profile, where X_i is the expression level of gene g_i. Our data then consists of a G_m x N matrix; the n^{th} column represents the expression profile $\mathbf{x}_n$ of the n^{th} sample, $n = 1, \ldots, N$. In addition, each sample is labeled by a phenotype $Y \in \{A, B, \ldots, K\}$. The labeled training set is $F = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$. Expression profiles $\mathbf{X}$ and phenotype labels Y are regarded as random variables,

and the elements of F represent independent and identically distributed samples from some underlying probability distribution of ($\mathbf{X}$, Y).

Our analysis is based entirely on the *ranks* within each expression profile. With G_m genes, there are $G_m!$ possible orderings for the expression values. The networks we consider typically have tens or hundreds of genes; consequently, working directly with individual permutations is not feasible. For example, any estimated distribution over permutations using training data would be highly singular. Instead, we base the analysis entirely on pairwise comparisons.

Rank Template Matching for Networks

Knowing the ordering of the gene expressions within each network expression profile is equivalent to knowing all of the pairwise orderings, i.e., whether $X_i < X_j$ or $X_i > X_j$ for each distinct pair of genes $1 \le i, j \le G_m$ within the network m. Evidently, there are $G_m(G_m - 1)/2$ such pairs. For example, if there are $G_m = 4$ genes, then there are six distinct ordered pairs: {(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3; 4)}. In order to define a template representing the expected ranking of network genes within a phenotype, we consider the probabilities $\Pr(X_i < X_j \,|Y = k)$ for each pair of genes $g_i < g_j$ and for each phenotype k. We estimate these probabilities from the training set by computing the fraction of samples in each phenotype for which gene g_i is expressed less than gene g_j. The rank template for a fixed network m and phenotype k is the binary vector $T^{(m,k)}$ of length $G_m(G_m - 1)/2$ where the i,j^{th} component is 1 if $\Pr(X_i < X_j \,|Y = k) > 0.5$ and 0 if $\Pr(X_i < X_j \,|Y = k) \le 0.5$. The calculation of a rank template is illustrated in **Figure 4.1**.

Given an expression profile $\mathbf{x}_n$ for the network m, there is then a natural measure for how well the sample matches the template $T^{(m,k)}$. The rank matching score of sample n is denoted by $R^{(m,k)}(\mathbf{x}_n)$ and is defined to be the fraction of the $G_m(G_m - 1)/2$ pairs for which the observed ordering within $\mathbf{x}_n$ matches the template—the orderings expected for phenotype k. See **Figure 4.1** for an illustration of a rank matching score.

Rank Conservation Indices

Averaging the rank matching score over all the samples in a phenotype k yields a rank conservation index denoted by $\mu_R^{(m,k)} = E(R^{(m,k)}|Y = k)$. This index is estimated by averaging the scores $R^{(m,k)}(\mathbf{x})$ over all the samples ($\mathbf{x}$, y) in the training set for which $y = k$. Whereas the rank matching score is a sample-based statistic, i.e., it is defined for each expression profile, the rank conservation index is a population statistic. The rank conservation index can be seen as a measure of the stability in rankings among the network genes in the phenotype. Two extreme cases correspond to (i) pure random shuffling of the expression values in the phenotype from sample to sample, in which case $\mu_R^{(m,k)} \approx 0.5$; and (ii) all samples displaying

exactly the same ordering, in which case $\mu_R^{(m,k)} \approx 1$. In general, there are many gene pairs g_i and g_j which are expressed on different scales, and hence $x_i < x_j$ across nearly all samples and phenotypes. As a result, one generally finds $\mu_R^{(m,k)} >> 0.5$. This index is similar to entropy in the sense that values of $\mu_R^{(m,k)} \ll 1$ indicate a highly disorganized state in which there is a great deal of variation among the rankings in phenotype k from sample to sample and values of $\mu_R^{(m,k)} \approx 1$ indicate a highly ordered state in which samples have very similar, and hence predictable, orderings among the genes.

Rank Difference Scores

Consider two phenotypes Y = A, B, and a fixed network m. If network m is tightly regulated in one phenotype, the samples from that phenotype, say Y = A, will have high $R^{(m,A)}$ values on average. But if $\mu_R^{(m,k)}$ is large for both k = A and k = B, and if the two rank templates $T^{(m,A)}$ and $T^{(m,B)}$ are significantly different, then the samples from phenotype Y = A will generally have low values for the statistic $R^{(m,B)}$ as well as high values for the statistic $R^{(m,A)}$, and vice-versa for the samples from phenotype Y = B. We capture this phenomenon, namely low variance of network ranking within a phenotype, but high variance between phenotypes, with a single statistic calculated for each sample: the difference $\Delta^{(m)}(\mathbf{x}_n) = R^{(m,A)}(\mathbf{x}_n) - R^{(m,B)}(\mathbf{x}_n)$. Clearly, $-1 \le \Delta^{(m)}(\mathbf{x}_n) \le 1$ with positive (respectively, negative) values providing evidence that the phenotype of sample n is Y = A (resp., Y = B). As a result, the difference score provides a classifier for phenotype identification based on the degree of regulation of the genes in network m. A new sample n is predicted to belong to phenotype Y = A if $\Delta^{(m)}(\mathbf{x}_n) > 0$ and to phenotype Y = B if $\Delta^{(m)}(\mathbf{x}_n) \le 0$. The classification rate for network m is then: $\eta(m) = \Pr(\Delta^{(m)}(\mathbf{X}) > 0 | Y = A) * \Pr(Y = A) + \Pr(\Delta^{(m)}(\mathbf{X}) \le 0 | Y = B) * \Pr(Y = B)$. The calculation of a rank difference score was shown in **Figure 4.1**.

For example, if Y = A denotes prostate cancer and Y = B denotes normal prostate, and if we assume that the two phenotypes are *a priori* equally likely, then $\eta(m)$ is simply the average of sensitivity and specificity relative to identifying cancer. In order to determine the most variably expressed networks between two given phenotypes, we calculate rank templates for each phenotype, evaluate the differential score for each sample in the training set and choose the networks with the largest estimated classification rate.

One previously reported method, k-TSP, classifies expression profiles based on k pairs of genes with the most significant expression reversals among all assayed genes [116]. The classifier based on the rank difference score is also based on k pairs of genes, with k equal to the distance between the two rank templates. To see this, notice that upon computing the difference $\Delta^{(m)}(\mathbf{x}_n)$ for pathway m and phenotypes A and B, the gene pairs (i,j) for which $T^{(m,A)}(i,j) = T^{(m,B)}(i,j)$ cancel out. The DIRAC-based classifier therefore reduces to voting among the gene pairs whose probabilities straddle 0.5—i.e., satisfy $\Pr(X_i < X_j$

$|Y = A) < 0.5 < \Pr(X_i < X_j | Y = B)$ or vice versa. However, these k pairs of genes are those in the "top-scoring network" as determined by DIRAC rather than the most discriminating k pairs overall (as would be identified by k-TSP).

Significance Testing

Procedures for estimating statistical significance are described below for metastatic prostate tumors (MT) and normal prostate (NP). Identical procedures were used for all binary phenotype datasets studied.

I. Deregulated Networks Based on the Difference in Rank Conservation Indices

Under the null hypothesis that no systematic difference in gene expression profiles exists between MT and NP, (i) the original phenotype labels were randomly re-assigned to samples, and rank conservation indices were calculated for all networks in each phenotype; (ii) the absolute difference in rank conservation index values between the two phenotypes was calculated for each network (i.e., $\theta(m) = |\mu_R^{(m,\mathrm{MT})} - \mu_R^{(m,\mathrm{NP})}|$ for the m^{th} network); (iii) the first two steps were repeated for 1,000 permutations to generate a null distribution of rank conservation difference values; and (iv) the significance level for $\theta(m)$ representing deregulation of a network between MT and NP was measured as the probability of observing differences in rank conservation greater than or equal to $\theta(m)$ in the null distribution.

II. Classification Rate for Networks Based on the Rank Difference Score

Under the null hypothesis that no systematic difference in gene expression profiles exists between MT and NP, (i) the original phenotype labels were randomly re-assigned to samples, and rank difference scores were calculated for each sample in all networks; (ii) sample classes in the permuted dataset were predicted as MT or NP based on whether the difference score was positive or negative, respectively, and scores were assigned to each network as measured by the estimated classification accuracy (i.e., $\eta(m)$ for the m^{th} network); (iii) the first two steps were repeated for 10,000 permutations to generate a null distribution of network classification rates; and (iv) the significance level for the $\eta(m)$ in predicting MT and NP profiles was measured as the probability of observing classification rates greater than or equal to $\eta(m)$ in the null distribution. To address the issue of multiple-hypothesis testing, we also estimated the false discovery rate (FDR) for each significance level, representing the fraction of our selected features which we would expect to be false positives.

Evaluating Classification Performance

We used leave-one-out cross validation to estimate the (generalization) error rate of each classification method studied. Importantly, for each classification method tested, all processes were done using only the

training samples without including any information from the test sample. Within each iteration of the cross validation loop, expression profiles in the original training data $F = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ are divided into two groups: a training set (F_{train}) and a test set (F_{test}). The classifier is trained on the $N - 1$ samples of F_{train} and then used to predict the phenotype of the remaining "left out" sample in F_{test}. The overall cross validation classification rate after N total train-test divisions and predictions is calculated as the average of sensitivity and specificity. Details for training and testing with each type of classifier are described below.

I. DIRAC

Rank templates, rank matching scores, and rank difference scores are calculated uniquely for each new instance of the training set F_{train}. The single best network is chosen based on the classification rate for samples of F_{train}, and the rank templates for this network are then used to assign two rank matching scores to the remaining sample comprising F_{test}. If the difference in matching scores is positive, the sample is predicted to be of phenotype A, otherwise it is classified as phenotype B.

II. TSP

The top-scoring pair (TSP) algorithm is described in detail elsewhere [116]. Here, we first filtered the total number of transcripts in F_{train}, keeping only the top G_m most differentially expressed genes (DEGs), where G_m is equal to the number of genes in the best network selected by DIRAC. The top features (i.e., DEGs) were selected based on the Wilcoxon ranksum test. Searching among all possible pairwise combinations of genes in the reduced dataset, we identified a single best pair (X_i and X_j) for which the difference $|\Pr(X_i < X_j \mid \mathrm{A}) - \Pr(X_i < X_j \mid \mathrm{B})|$ is maximized (or alternatively, $|\Pr(X_i > X_j \mid \mathrm{A}) - \Pr(X_i > X_j \mid \mathrm{B})$. The phenotype of F_{test} is then predicted by comparing the expression levels for this gene pair.

III. SVM

Prior to training a support vector machine (SVM) classifier on the samples of F_{train}, we also filtered down to the top G_m DEGs within each cross validation loop, where G_m is equal to the number of genes in the best network selected by DIRAC. The SVM was then trained on the expression values of these G_m genes using a Gaussian kernel, and then used to predict the phenotype of F_{test}.

Chapter 4 Figures & Tables

Figure 4.1. Overview of Differential Rank Conservation (DIRAC) methods.

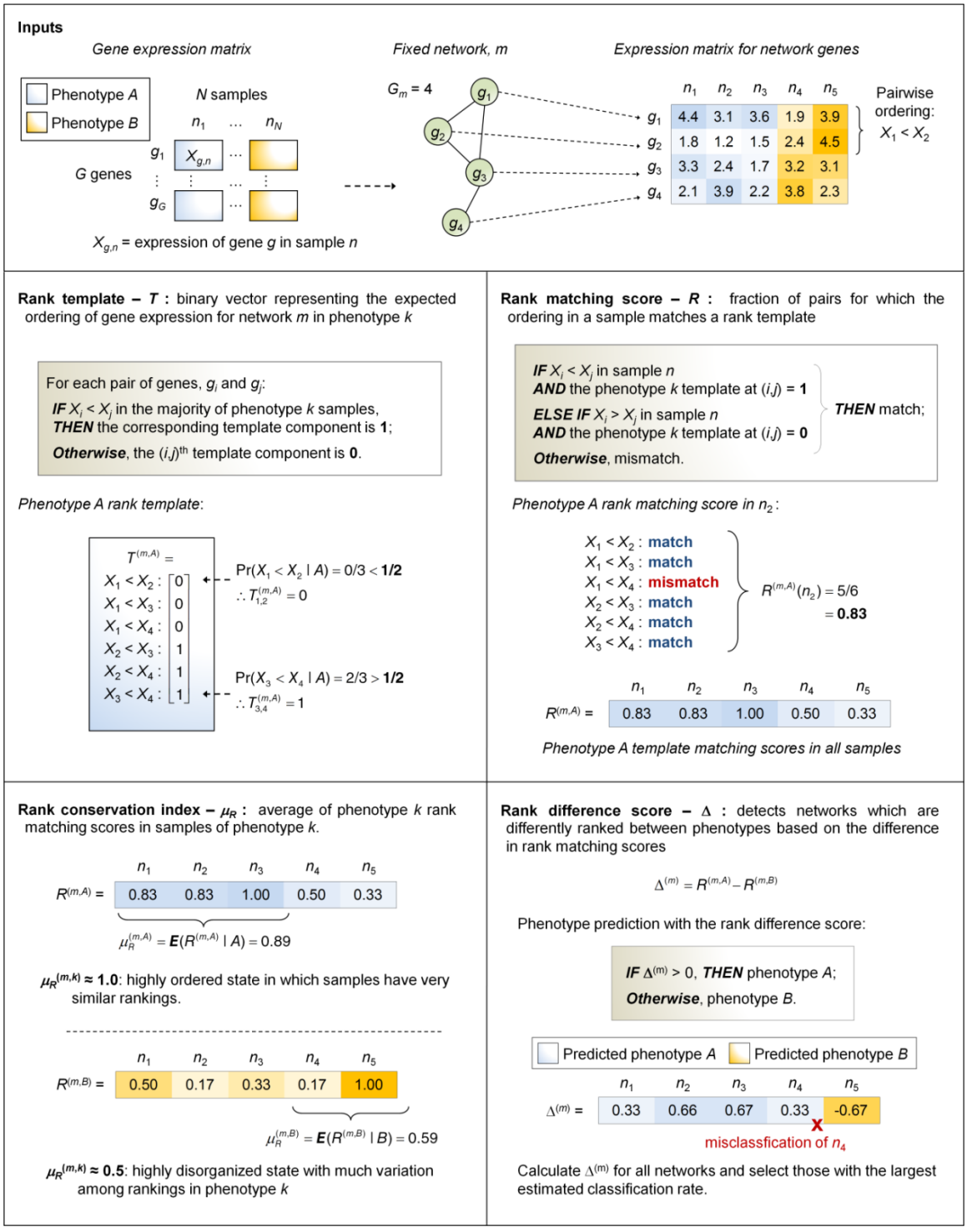


Figure 4.2. Prototypical scenarios observed for networks in DIRAC.

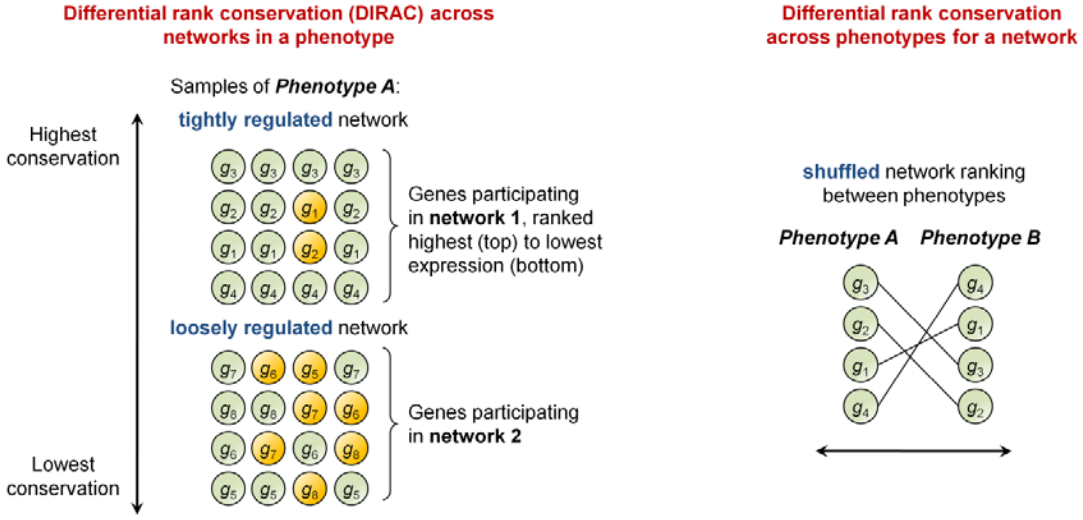
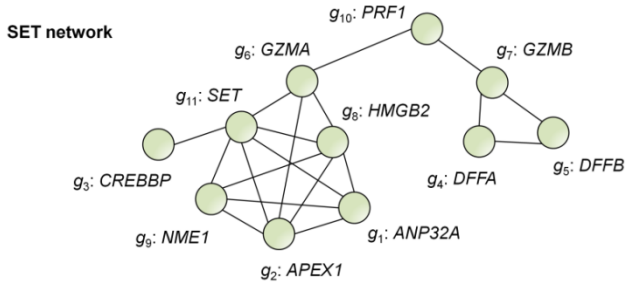


Figure 4.3. Example of a tightly regulated network in normal prostate.



Rank conservation of SET pathway in normal prostate

Pairwise orderings	Rank template (1 = true; 0 = false)	Network rankings observed among normal prostate samples															
$g_1 < g_2$	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1
$g_1 < g_4$	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
$g_1 < g_9$	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	1
$g_1 < g_{11}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
$g_2 < g_9$	1	1	1	1	0	0	1	1	1	0	1	0	0	0	0	1	1
$g_2 < g_{11}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
$g_3 < g_6$	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
$g_3 < g_8$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
$g_3 < g_{10}$	1	1	1	1	1	1	0	1	1	1	1	0	1	0	1	1	1
$g_4 < g_8$	0	0	1	1	0	0	1	1	0	1	1	1	0	0	0	1	0
$g_4 < g_{10}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
$g_5 < g_6$	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1
$g_5 < g_7$	1	1	1	1	0	1	1	0	0	1	0	1	0	0	1	1	1
$g_6 < g_7$	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1
$g_6 < g_8$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
$g_6 < g_{10}$	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
$g_7 < g_8$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
$g_7 < g_9$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
$g_7 < g_{10}$	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1
$g_9 < g_{10}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
$g_9 < g_{11}$	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1
Mismatches (out of 55 pairs):		1	1	2	2	2	2	2	2	3	3	3	4	4	5	8	8
Rank matching score:		0.982	0.982	0.964	0.964	0.964	0.964	0.964	0.964	0.945	0.945	0.945	0.927	0.927	0.909	0.855	0.855
Frequency observed (out of 18 samples):		3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

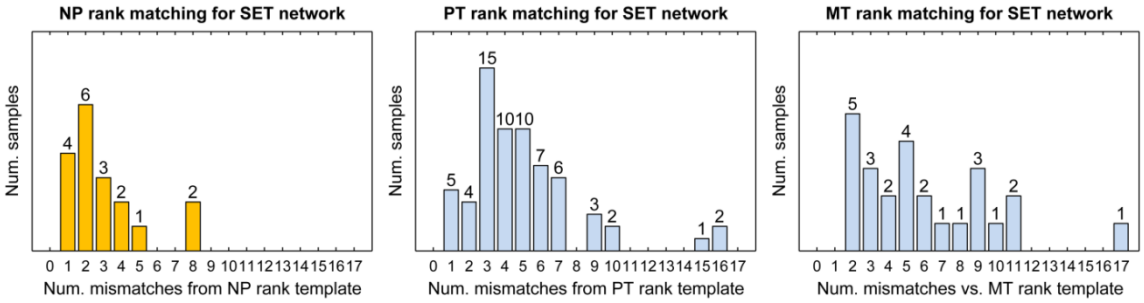


Figure 4.4. Deregulation of networks in disease.

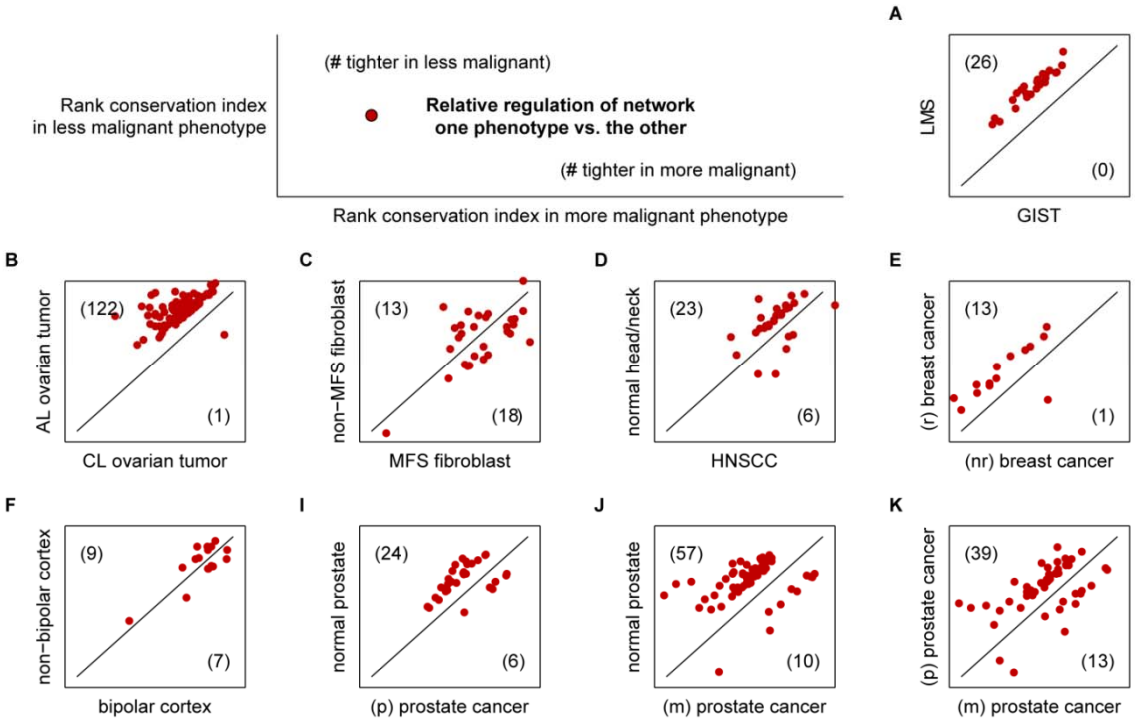


Figure 4.5. Diverse rank conservation of networks across phenotypes.

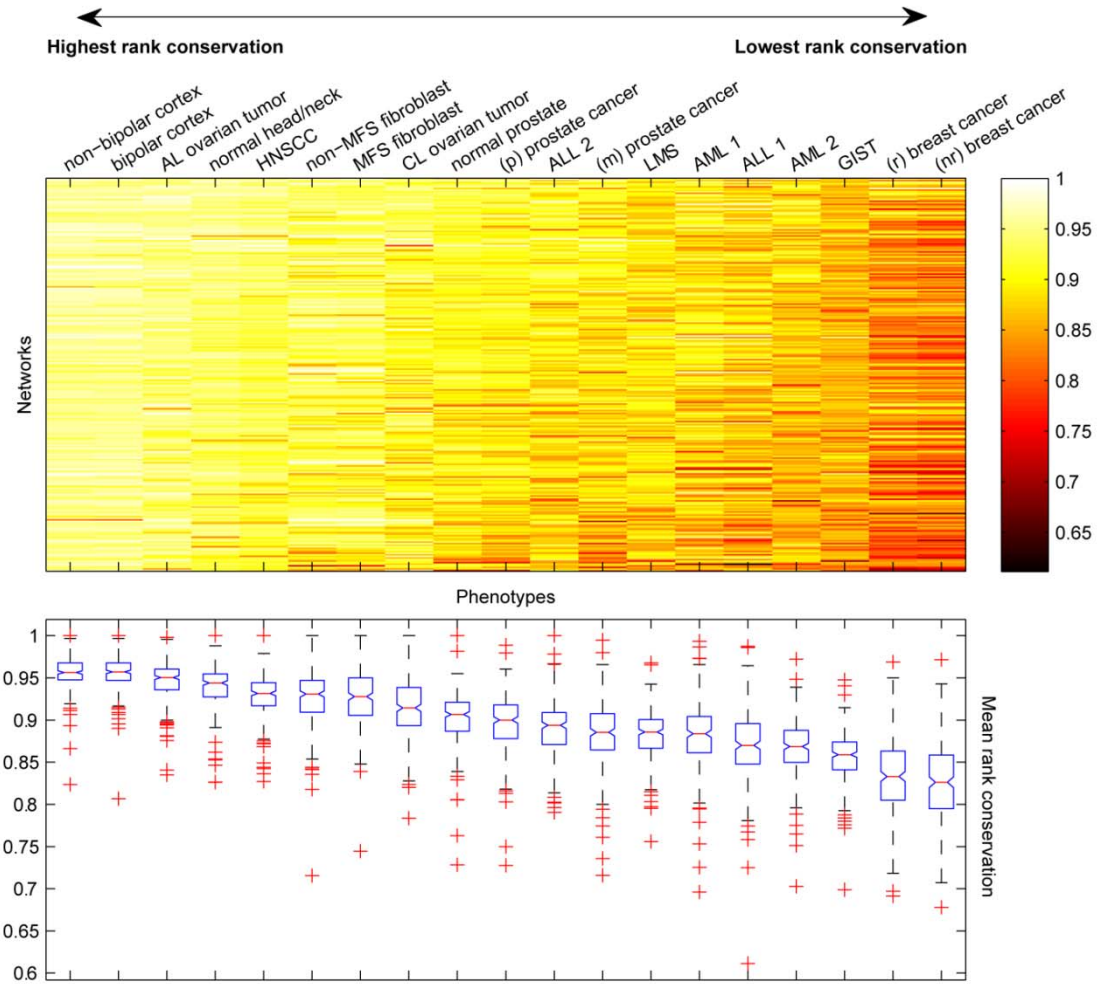


Figure 4.6. Differential rank conservation across all networks for a set of two prostate phenotypes.

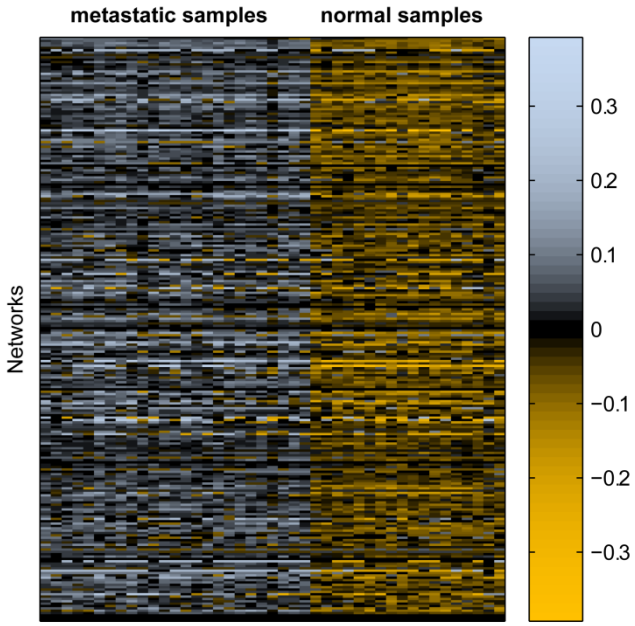


Figure 4.7. Differential rank conservation of the EDG1 network in metastatic prostate cancer and normal prostate.

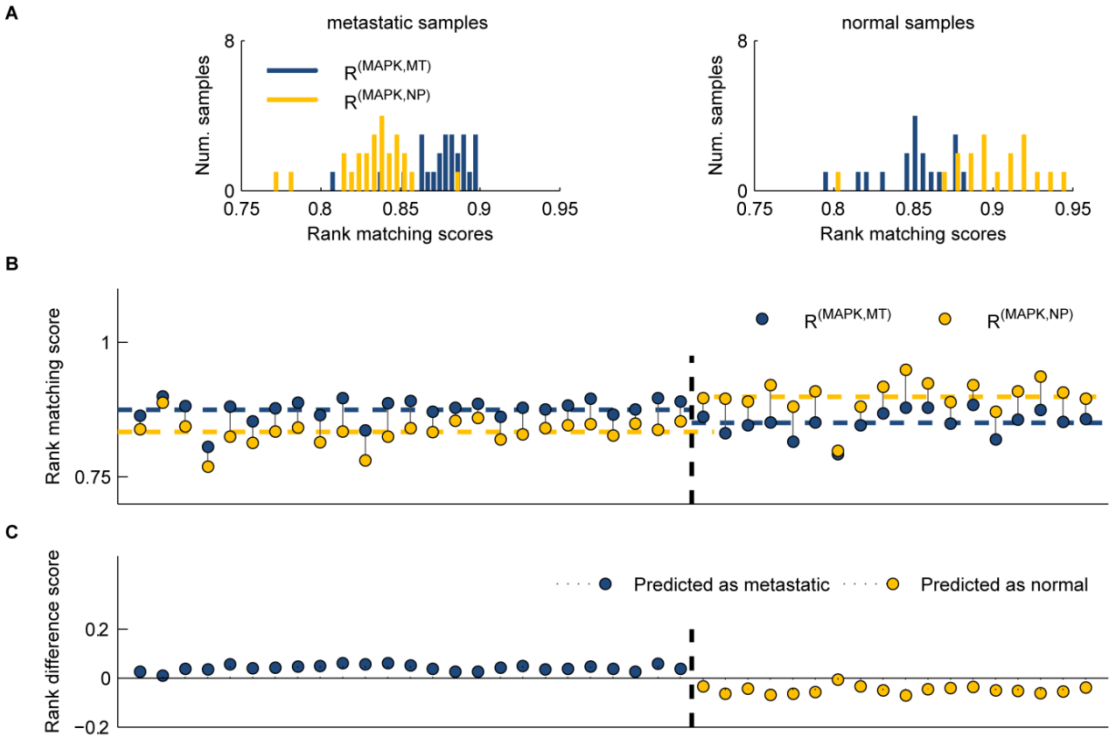


Figure 4.8. Classification with DIRAC compared to other methods.

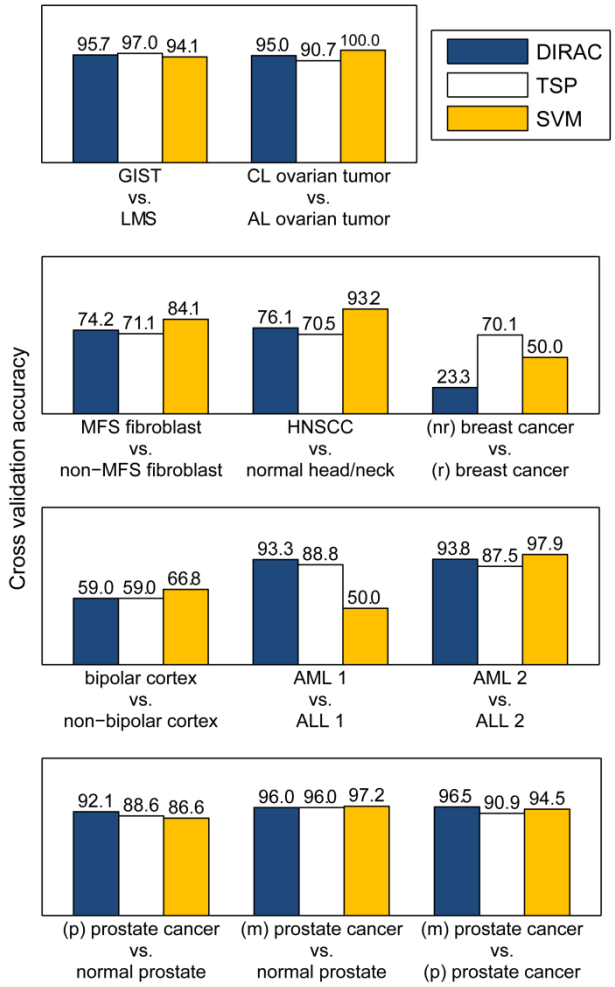


Table 4.1. Most tightly regulated networks in normal prostate and primary and metastatic prostate tumors, as indicated by rank conservation index values.

Tightly regulated networks in NP				
Network name	**Num. genes**	**Num. gene pairs**[a]	**Avg. variance in NP**	**μ_R in NP**
GS	6	15	1.328	1.000
FOSB	4	6	1.141	0.981
AKAP13	7	21	0.796	0.955
AGPCR	11	55	0.811	0.955
RNA	8	28	0.453	0.948
CACAM	12	66	0.551	0.947
NDKDYNAMIN	17	136	0.619	0.946
ETC	8	28	0.350	0.946
SET	11	55	0.537	0.945
SKP2E2F	10	45	0.339	0.943
Tightly regulated networks in PT				
Network name	**Num. genes**	**Num. gene pairs**	**Avg. variance in PT**	**μ_R in PT**
GS	6	15	1.270	0.989
FOSB	4	6	1.525	0.979
AKAP13	7	21	0.880	0.960
ARGININEC	6	15	0.548	0.960
PLK3	8	28	0.672	0.951
CDC42RAC	15	105	0.547	0.946
RNA	8	28	0.489	0.946
CREM	7	21	0.563	0.944
BOTULIN	4	6	0.850	0.944
AGPCR	11	55	0.771	0.943
Tightly regulated networks in MT				
Network name	**Num. genes**	**Num. gene pairs**	**Avg. variance in MT**	**μ_R in MT**
GS	6	15	1.322	0.995
FOSB	4	6	1.575	0.980
CREM	7	21	0.659	0.966
S1P	6	15	0.465	0.963
RAN	5	10	0.371	0.960
SLRP	4	6	1.227	0.960
BOTULIN	4	6	0.722	0.953
AKAP13	7	21	0.787	0.947
SARS	10	45	0.819	0.939
RAB	10	45	0.441	0.937

[a]The number of gene pairs is equal to $G_m(G_m - 1)/2$, where G_m is the number of genes in the network.

Table 4.2. Human disease gene expression datasets studied with DIRAC.

Dataset	Ref	Samples	Tissue type	Disease/source (subtypes)[a]	Short name[b]	Subtype samples
A	[273]	68	gastrointestinal sarcoma	gastrointestinal stromal tumor	GIST	37
				Leiomyosarcoma	LMS	31
B	[282]	43	ovarian tumors	carcinoma-like ovarian tumor	CL ovarian tumor	20
				adenoma-like ovarian tumor	AL ovarian tumor	23
C	[283]	101	skin fibroblasts	Marfan syndrome subjects	MFS fibroblast	60
				control subjects	non-MFS fibroblast	41
D	[284]	44	head and neck skin cells	head and neck squamous cell carcinoma	HNSCC	22
				normal head and neck skin cells	normal head/neck	22
E	[250]	60	primary breast cancer tumor	patients non-response (cancer recurred) to treatment	(nr) breast cancer	28
				patients responsive (disease-free) to treatment	(r) breast cancer	32
F	[285]	61	dorsolateral prefontal cortex and orbitofrontal cortex	Bipolar disorder patients	bipolar cortex	30
				control patients	non-bipolar cortex	31
G	[286]	72	blood and bone marrow	acute myeloid leukemia	AML 1	25
				acute lymphocytic leukemia	ALL 1	47
H	[287]	48	blood and bone marrow	acute myeloid leukemia	AML 2	24
				acute lymphocytic leukemia	ALL 2	24
I	[275]	83	normal and tumorgenic prostate	primary prostate tumors	(p) prostate cancer	65
				normal prostate tissue	normal prostate	18
J	[275]	43	normal and metastatic prostate	metastatic prostate tumors	(m) prostate cancer	25
				normal prostate tissue	normal prostate	18
K	[275]	90	prostate tumor	metastatic prostate tumors	(m) prostate cancer	25
				primary prostate tumors	(p) prostate cancer	65

[a]For each set of expression profiles, the two subtypes are listed in order from most to least malignant (e.g., tumor type with worst prognosis or cancer versus control).
[b]Short names are used to reference specific phenotypes in subsequent figures.

Table 4.3. Most differentially regulated networks between three stages of prostate disease.

Differentially regulated networks (PT vs. NP)						
Network name	**Num. genes**	**Num. gene pairs**[a]	**μ_R in PT**	**μ_R in NP**	**Abs. difference in μ_R**	***P*-value**
TCRA	12	66	0.859	0.928	0.069	5.85E-04
TCRMOLECULE	5	10	0.871	0.939	0.068	6.69E-04
EIF2	7	21	0.854	0.915	0.061	1.33E-03
TERC	6	15	0.877	0.933	0.056	2.29E-03
NEUTROPHIL	8	28	0.848	0.901	0.053	3.33E-03
GLYCOLYSIS	8	28	0.879	0.929	0.050	4.57E-03
ACE2	11	55	0.835	0.885	0.050	4.72E-03
FIBRINOLYSIS	12	66	0.847	0.891	0.044	9.17E-03
INTRINSIC	22	231	0.852	0.896	0.044	9.45E-03
CLASSIC	10	45	0.886	0.930	0.044	9.74E-03
Differentially regulated networks (MT vs. NP)						
Network name	**Num. genes**	**Num. gene pairs**	**μ_R in MT**	**μ_R in NP**	**Abs. difference in μ_R**	***P*-value**
FIBRINOLYSIS	12	66	0.736	0.891	0.156	-6.66E-16
EXTRINSIC	12	66	0.716	0.870	0.155	-6.66E-16
INTRINSIC	22	231	0.761	0.896	0.135	2.02E-05
CLASSIC	10	45	0.829	0.930	0.100	2.90E-04
TERC	6	15	0.843	0.933	0.091	6.21E-04
ION	5	10	0.892	0.806	0.086	8.35E-04
COMP	14	91	0.832	0.914	0.082	1.20E-03
NEUTROPHIL	8	28	0.819	0.901	0.082	1.21E-03
ARF	15	105	0.829	0.911	0.081	1.32E-03
PEPI	5	10	0.808	0.889	0.081	1.34E-03
Differentially regulated networks (MT vs. PT)						
Network name	**Num. genes**	**Num. gene pairs**	**μ_R in MT**	**μ_R in PT**	**Abs. difference in μ_R**	***P*-value**
EXTRINSIC	12	66	0.716	0.856	0.140	-6.66E-16
FIBRINOLYSIS	12	66	0.736	0.847	0.111	8.06E-06
INTRINSIC	22	231	0.761	0.852	0.091	4.03E-05
ION	5	10	0.892	0.803	0.089	6.05E-05
PEPI	5	10	0.808	0.895	0.087	6.85E-05
ARGININEC	6	15	0.880	0.960	0.080	1.65E-04
LEPTIN	8	28	0.807	0.727	0.080	1.73E-04
NOTCH	4	6	0.853	0.931	0.077	2.42E-04
PLC	8	28	0.800	0.859	0.059	1.74E-03
BETAOXIDATION	6	15	0.864	0.922	0.058	1.86E-03

[a]The number of gene pairs is equal to $G_m(G_m - 1)/2$, where G_m is the number of genes in the network.

Table 4.4. Statistical significance of network deregulation in malignant phenotypes.

Dataset	**# tighter in less malignant**	**# tighter in more malignant**	**Outcome**	**Binomial *P*-value**
A	26	0	1	0.00
B	122	1	1	0.00
C	13	18	0	0.76
D	23	6	1	0.00
E	13	1	1	6.10E-05
F	9	7	1	0.23
I	24	6	1	1.62E-04
J	57	10	1	3.41E-10
K	39	13	1	6.38E-05
Total	326	62	8	0.00
		Binomial *P*-value for outcomes:	0.002	

Table 4.5. Most variably expressed networks between different stages of prostate cancer.

Variably expressed networks (PT vs. NP)					
Network name	**Num. genes**	**Num. gene pairs**[a]	**Template difference**[b]	**Apparent accuracy**	***P*-value**
KERATINOCYTE	46	1035	0.070	0.981	< 1.0E-07
TOLL	31	465	0.073	0.945	1.21E-05
MAPK	83	3403	0.064	0.941	2.02E-05
MET	35	595	0.103	0.941	2.02E-05
FCER1	36	630	0.059	0.931	6.85E-05
INTEGRIN	34	561	0.094	0.923	1.21E-04
AT1R	34	561	0.096	0.922	1.25E-04
ERK	29	406	0.037	0.921	1.29E-04
CARDIACEGF	17	136	0.118	0.920	1.33E-04
IL1R	28	378	0.071	0.915	1.81E-04
Variably expressed networks (MT vs. NP)					
Network name	**Num. genes**	**Num. gene pairs**	**Template difference**	**Apparent accuracy**	***P*-value**
MAPK	83	3403	0.111	1.000	< 1.0E-07
DEATH	29	406	0.128	1.000	< 1.0E-07
IL2RB	35	595	0.096	1.000	< 1.0E-07
HIVNEF	53	1378	0.148	1.000	< 1.0E-07
MET	35	595	0.165	1.000	< 1.0E-07
NO1	27	351	0.125	1.000	< 1.0E-07
NFAT	47	1081	0.164	1.000	< 1.0E-07
PPARA	50	1225	0.100	1.000	< 1.0E-07
ACTINY	19	171	0.123	1.000	< 1.0E-07
FCER1	36	630	0.111	0.990	< 1.0E-07
Variably expressed networks (MT vs. PT)					
Network name	**Num. genes**	**Num. gene pairs**	**Template difference**	**Apparent accuracy**	***P*-value**
FCER1	36	630	0.119	0.985	< 1.0E-07
TCR	44	946	0.103	0.969	< 1.0E-07
BCR	33	528	0.133	0.969	< 1.0E-07
HIVNEF	53	1378	0.119	0.969	< 1.0E-07
MET	35	595	0.126	0.969	< 1.0E-07
PDGF	27	351	0.128	0.957	< 1.0E-07
BIOPEPTIDES	37	666	0.107	0.957	< 1.0E-07
MAPK	83	3403	0.100	0.954	< 1.0E-07
IL2RB	35	595	0.087	0.954	< 1.0E-07
AT1R	34	561	0.111	0.954	< 1.0E-07

[a]The number of gene pairs is equal to $G_m(G_m - 1)/2$, where G_m is the number of genes in the network.
[b]The template difference represents the Hamming distance between two binary rank template vectors.

Chapter 5: Exploration of Gene Expression Analysis Methods and Applications in Cancer and the Brain[5]

Results from DIRAC provide a coarse, but meaningful glimpse into patterns of network regulation for different phenotypes, based on combinatorial relationships between the involved genes. As such, this method can be a powerful first step towards characterizing network-level expression differences in new disease data. One interesting application has been investigating the molecular patterns that underlie differences in aggression of human astrocytomas. Working with fellow graduate student, Chunjing Wang, I have found striking correlations between dysregulation in biological pathways at the gene expression level and the up- or down-regulation of individual genes that may contribute to genetic heterogeneity. Despite these observations of high variability in astrocytomas—most prominent in the highest grades—we were also able to ascertain robust network signatures to distinguish between grades. These findings present interesting hypotheses that, upon further exploration, could produce new insights into pathological mechanisms that give rise to different tumor grades.

In this chapter, I present several other applications and extensions of DIRAC, along with alternative methods to explore gene expression in disease and biology. Importantly, DIRAC is not limited to disease applications, nor is it the only method worth considering for disease classification. As part of another fascinating collaborative project, I used DIRAC to examine the expression of modules in a transcriptional regulatory network model we reconstructed for the honeybee brain [288]. We found that several regulatory modules (i.e., target genes controlled by the same transcription factor) exhibited significant differential expression between behavioral states. These studies highlight some of the insights that can be gleaned through systems analysis of high-throughput data, coupled with deep domain expertise.

The promising results observed with DIRAC also highlight several paths for modification and potential improvement of the method. In particular, the *a priori* defined networks used as inputs to DIRAC might often not be reflective of the active molecular processes in a particular phenotype. Preliminary studies with protein-protein interaction networks have illustrated the potential of adaptively learning context-specific networks with DIRAC. Additionally, efforts have begun to use DIRAC for multi-class

[5] This chapter includes material that was reproduced with permission from the following publications:
(i) Chandrasekaran, S., S.A. Ament, J.A. Eddy, S.L. Rodriguez-Zas, B.R. Schatz, N.D. Price, and G.E. Robinson. 2011. Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states. Proceedings of the National Academy of Sciences USA. 108(44):18020-18025. (**Network Analysis of Transcriptionally Regulated Modules in the Honeybee Brain** section; text written in collaboration with Sriram Chandrasekaran and Seth Ament, analysis contributions stated in section).
(ii) Ko Y., S.A. Ament, J.A. Eddy, J. Caballero, J.C. Earls, L. Hood, and N.D. Price. 2013. Cell type-specific genes show striking and distinct patterns of spatial expression in the mouse brain. Proceedings of the National Academy of Sciences USA. 110(8):3095-3100. (**Spatial Expression Patterns in the Mouse Brain** section; text written in collaboration with Younhee Ko and Seth Ament, analysis contributions stated in section).

classification problems and also to make method tools more widely accessible to the scientific community. Finally, DIRAC does not immediately provide insight into the higher-level interactions between *multiple* networks that may be important in disease. I have helped to develop an alternative approach called GSERA, which quantifies patterns among two or more biological networks.

Importantly, expression analysis that focuses on patterns of individual genes can still be informative, especially when opportunities arise to interface with biology and medical experts, thereby providing valuable context to computational results. For example, a collaborative exploration of mesenchymal-to-endothelial reverting transitions in leiomyosarcomas resulted in mechanistic and prognostic insights, based on the expression of informative marker genes [289]. Applying statistical and clustering approaches, I found that patients classified as more "endothelial-like" (based on up-regulated expression of marker genes) exhibited poorer survival outcomes, relative to the "mesenchymal-like" samples.

Network Analysis of Transcriptionally Regulated Modules in the Honeybee Brain

Both genotype and the environment influence behavior, and these effects lead to massive changes in the brain transcriptome [290]. This has led to the idea that distinct neurogenomic states underlie distinct behaviors, but it is not known how these states are defined and maintained. Using brain transcriptomic profiles from 853 individual honeybees exhibiting different behavioral phenotypes in naturalistic contexts, fellow graduate students Sriram Chandrasekaran and Seth Ament and I investigated whether behavior-specific neurogenomic states can be inferred from the coordinated action of transcription factors (TFs) and their predicted target genes. Our data comprised transcript profiles from 27 integrated pair-wise comparisons that probed hereditary and environmental influences on social behavior and brain gene expression. Each comparison surveyed transcriptome responses to a heritable or environmental factor related to one or more of three ecologically important behavioral categories: aggression, maturation, and foraging. Unsupervised hierarchical clustering of these transcriptomic profiles showed three clusters of profiles that correspond to these categories.

Transcriptional regulatory network construction overview

To explore the mechanisms potentially regulating these behavior-specific neurogenomic states, a brain Transcriptional Regulatory Network (TRN) model was reconstructed by Sriram Chandrasekaran. TRN reconstruction was performed using a new combination of two well-known algorithms: a network of high-confidence putative TF-gene interactions was generated using ARACNE (Accurate Reconstruction of Cellular Networks [291]), and these interactions were leveraged to predict expression in new conditions with LARS (Least Angle Regression [292]). Interactions were inferred from expression profiles and a list

of 236 predicted bee TFs based on orthology with *Drosophila melanogaster*. The goal was to infer relationships between genes and subsequently predict changes in the expression of those genes involved in behavior in the bee brain given the expression values of the TFs.

The TRN model quantitatively predicts gene expression changes of more than 2,000 genes involved in behavior with high accuracy, even for behavioral phenotypes on which it was not trained. Having demonstrated predictive ability and biological relevance, we analyzed the TRN to explore the hypothesis that behaviorally related neurogenomic states arise, in part, from the coordinated action of TFs and their predicted target genes. Most of the top 20 most connected TFs ("hubs") and many of their predicted targets were differentially expressed (FDR < 0.05) across many of the 27 comparisons. Some TFs and their modules of putative target genes were active across all three behavioral categories ("global regulators"), i.e., their predicted target genes were significantly over-represented among differentially expressed genes across all three behavioral categories. Other TFs and their modules were active only for a particular behavioral state within a category. State-specific TFs were significantly more common than predicted by chance (P < 1e-10), indicating that they also are a robust feature of this TRN.

Differential rank conservation analysis of the network

I used Differential Rank Conservation (DIRAC) [293] to investigate changes in relative gene expression within each module, either across individual bees within states, or between states. Specifically, DIRAC quantifies and assesses expression consistency in the context of the rankings of target genes within a selected module: for each microarray, the expression values of the module genes are ordered from highest expression (ranked first) to lowest expression (ranked last). For any state or comparison in which genes of a particular module were not expressed, this module was omitted from DIRAC calculations. Two intriguing dynamic properties of the TRN were detected. First, while modules generally showed consistent gene expression across individuals *within* behavioral states, some modules exhibited significant changes in relative gene expression ordering *between* states.

I first used DIRAC to characterize the consistency of rank ordering within each module for the individual bees within each of the 48 states, as detailed in [293]. A network is considered tightly regulated within a state if the relative expression of module genes is mostly consistent across individuals; a network is considered loosely regulated if the relative expression of genes is greatly varied between individuals of a state. Consistency of relative expression for a module in a selected state can range between 0.5 (relative expression of module genes is completely different in each individual) and 1.0 (relative expression of modules genes is identical in all individuals); the average consistency of all network modules across all states was 0.89±0.06, indicating that modules are tightly regulated in general within states (**Figure 5.1**).

I next used DIRAC to detect changes in relative expression of genes between states for each module. DIRAC identifies variably expressed or "shuffled" modules that, for each comparison, enable statistically significant classification of expression profiles between states [293]. I estimated P-value and FDR for classification accuracies by repeatedly performing calculations for each comparison with 10,000 sets of randomly permuted individuals. That is, DIRAC determines a probability for these "shuffling" events based on the accuracy with which these changes in relative gene expression can be used to classify the two states in the comparison. This shuffling was greatest for hub genes (**Figure 5.2**) such as *ftz-f1*, *Creb*, and *Lag1*, a TF linked to metabolic regulation [294]. Specifically, there was a highly significant association ($R^2 = 0.45$) between the number of target genes in a module and the degree of shuffling. In other biological contexts, gene shuffling is associated with transcriptional "dysregulation" in diseased states [293]; such differences have not yet been detected in network analysis for normal states. There was no such association for random gene sets with the same size distribution ($R^2 = 0.02$), indicating that this is a specific feature to TF-target modules, rather than an artifact of the method.

Module-level expression differences at short to long timescales

Changes in relative gene expression ordering also varied depending on the timescale of the underlying perturbation. We subdivided the 27 comparisons into three categories based on the timescales over which they influence behavior: *long-term differences* (hereditary differences between strains and sub-species that accumulate over evolutionary time); *medium-term differences* (changes occurring over a few weeks, primarily due to environmentally-induced changes); and *short-term differences* (factors that influence behavior for only hours or a few days; e.g., pheromones and the imprints of spatiotemporal memories on foraging). We compared the effects of these timescales on four aspects of network dynamics and regulation (**Figure 5.3**): the number of differentially expressed genes in the TRN (DEGs; A), the number of between-state perturbations in relative gene expression within modules ("shuffling", B), the number of TFs influencing each target gene (in-degree; C); and the number of target genes regulated by each TF (out-degree; D). Statistical significance for differences between timescales was determined using Kruskal-Wallis rank-sum tests. Increased shuffling was more often associated with longer-lasting behavioral states, e.g., those related to behavioral maturation, rather than more dynamic states, e.g., spatiotemporal floral memories ($P = 0.008$). Longer-lasting states also showed a significant increase in the number of TFs predicted to regulate each gene ($P = 0.002$).

In other respects, the network states associated short- vs. long-term responses were quite similar. There were no striking differences either in the number of genes responsive to short- vs. long-term stimuli, or in topological properties such as the number of targets for each TF in a given context (k_{out}); this is consistent

with previous findings that short- and long-term differences in honeybee aggression involve many of the same genes (Alaux 2009. PNAS). Together, these results suggest that both short- and long-acting stimuli can robustly activate genes in the network, but more stable responses frequently also involve perturbations in the relationships among these genes. Such regulatory complexity is considered a hallmark of network stability.

Spatial Expression Patterns in the Mouse Brain

To investigate the extent to which subdivisions in the mammalian brain correspond to regional differences in the expression of genes within major cell types, I worked with postdoctoral fellows Younhee Ko and Seth Ament to analyze highly spatially resolved brain gene expression data from the ABA for panels of cell type-specific genes. The coronal and sagittal sections were broken into small (coronal: 20um x 30um; sagittal: 20um x 40um) "patches" (pixels) reflecting local gene expression. K-means clustering was then performed (with *k* between 2 and 100) to find subsets of patches with similar gene expression across all the neuron-specific, astrocyte-specific, or oligodendrocyte-specific genes; i.e., the k-means clustering grouped patches based on their similarity in the cell type-specific gene expression space. For neurons, the clusters even up to 100 were highly spatially coordinated (**Figure 5.4**), though the size of the added regions became much smaller above ca. $k = 60$, which is the *k* I will use for a number of examples here. The resulting transcriptionally defined clusters were then transformed back into a color-coded brain image to assess their overlap with known brain regions. Our goal was to systematically characterize the transcriptionally distinct, spatially contiguous neuronal and glial subtypes based on their gene expression patterns across the brain.

Investigating spatial contiguity of brain regions defined by neuron-specific expression patterns

Cell type-specific genes were defined as those that were previously shown to be >10-fold enriched in each of these cell types. Cahoy et al. [295] found 320 genes that were >10-fold enriched in neurons, 185 in astrocytes, and 131 in oligodendrocytes, but ABA data were available for only a subset of these genes. All available *in situ* hybridization images downloaded from the ABA showing expression on coronal or sagittal slices through the centermost part of the brain (1-3 images per gene on each plane). The resulting dataset included images for 170 neuron-, 44 oligodendrocyte-, and 50 astrocyte-specific genes for the coronal plane, and for 250 neuron-, 101 oligodendrocyte-, and 154 astrocyte-specific genes for the sagittal plane (where ABA measured more genes).

After systematically quantifying the spatial patterns of neuron-specific using image processing tools in Matlab, I found that clusters derived from neuron-specific genes were highly spatially contiguous; i.e.,

adjacent patches in the brain were very frequently assigned to the same cluster (**Figure 5.4**). For $k = 60$, 59 of the 60 clusters were spatially contiguous with >59% of the patches adjacent to cluster patches belonging to that same cluster. All 59 of these spatially contiguous clusters had significantly greater ($P < 0.001$) contiguity by this measure than I found in a distribution generated from randomly permuting the spatial coordinates for the patches within the mouse brain (while keeping the same number of patches per cluster). Known brain regions are spatially contiguous within the brain. As such, the high levels of spatial contiguity we observed for k-means clusters suggests that the vast majority of these transcriptionally defined regions correspond to functionally relevant brain regions, even with high numbers of clusters.

Characterizing network-level differences between spatial regions

I also used DIRAC [293] to evaluate cluster-specific differences in the relative expression levels of genes within functionally-related categories (defined by GO). Such "shuffled" pathways may be differently regulated in different parts of the brain (**Figure 5.5**). Intriguingly, DIRAC analysis showed that the k-means cluster for striatum differed from the clusters for the pallidum, thalamus, and hypothalamus in the relative expression levels for genes involved in central nervous system development. The clusters corresponding to the pallidum vs. thalamus/hypothalamus could then be further distinguished by changes in the expression of genes related to synaptic transmission. These data are consistent with the idea that neurons in different regions of the brain exhibit distinct states (encoding different functions), specified by a multiplicity of transcripts, and that there are quite a large number of distinct neuronal states. Thus, the relative expression difference amongst a broad range of neuron-specific genes is the main contributor to the observed spatially contiguous, transcriptionally defined clusters.

Extensions of DIRAC and Alternative Network Expression Approaches

In addition to the initial version of DIRAC, I have explored an extension to the method that will add key functionality for dealing with mechanistic reaction networks such as the cancer-specific metabolic reconstruction described in **Chapter 6**. Prior to DIRAC, I also worked on a related project that focused on investigating informative patterns of expression between multiple related networks. Notably, I collaborated with another graduate student, John Earls, to develop a software package—Adaptive Unified Relative Expression Analysis (AUREA)—that includes all published forms of relative expression analysis algorithms for classification of gene expression profiles [246]. AUREA provides a user-friendly interface for employing each algorithm, as well as an adaptive learner that searches for an optimal model and set of parameters for classification. While the current version of AUREA includes the classification aspect of

DIRAC, future versions will also include population-level analyses as well as the extensions and related tools described below.

Adaptive interaction-network based DIRAC

I have developed an extension to DIRAC that offers two potentially advantageous features: (i) instead of *a priori* defined database networks, the method instead operates on specific regions (subnetworks) within protein-protein interaction (PPI) networks generated by high-throughput data; (ii) based on the links between genes in the PPI network, the method is able to adaptively grow and shrink networks to identify more discriminative signatures for classification.

Defining subnetworks using link communities

PPI information was obtained from public, manually curated databases including the Biological General Repository for Interaction Datasets (BioGRID), the Human Protein Reference Database (HPRD), and the Database of Interacting Proteins (DIP), and used to construct PPI networks. These PPI networks serve as inputs to the computational analyses used here. Link Communities is a hierarchical clustering method based on the similarity of links rather than nodes in the graph [296]. In the resulting dendrogram, nodes may occupy multiple agglomerations due to their links; this is important because it allows genes to be grouped into multiple subnetworks.

Adaptive modification of variably expressed subnetworks

The initial pool of networks used as the search space DIRAC may not yield the best signatures for distinguishing between two phenotypes. Networks defined *a priori* in pathway databases—regardless of the level of curation—may not be representative of functioning sets of genes in a particular phenotype, and therefore would be unlikely to serve as accurate classifiers. While interaction networks represent mechanistic links between genes, they may also not be specific to the phenotypes being studied; furthermore, subnetworks defined according to network structure may not necessarily correspond to informative differential expression of genes. I have used a sequential forward floating search algorithm [297] to grow and shrink subnetworks identified as variably expressed, in an effort to improve classification accuracy.

The inputs to the sequential forward floating search (SFFS) algorithm are the top *k*th (as determined by apparent accuracy) link community (LC) sets. Generally, I set k to be 50. While the link communities were defined by utilizing PPI data, it is assumed that there is still useful information content within the

PPI data that has not been captured by the initial link communities. That is, the biomolecular networks defined by the LC sets should be refined to improve biomarker fidelity.

Therefore, the sequential forward floating search seeks to adapt each LC set by iteratively adding and removing genes. The pool of available genes to add during an iteration is the set of genes which share an edge (putative interaction) in the global PPI graph with at least one of the genes already in the seed set. Furthermore, on each iteration, following adding the "best" gene (as determined by apparent accuracy), SFFS begins conditional exclusion of genes already within the set. This is in hopes of constructing the smallest set which still provides the best accuracy. Imagine a set of genes {*a,b,c,d*} to which is added gene *e*. Gene *e* provides all of the information (in the sense of acting as a distinguishing biomarker for phenotypes x versus y) already provided by gene *b*, and then some. Thus, SFFS would be able to safely remove gene *b*. If instead the best gene to remove would be that just added on this iteration, gene *e*, then SFFS removes no genes and goes on to the next iteration. The conditional exclusion step proceeds as long as possible in a given iteration, and thus on a single SFFS iteration, exactly one gene may be added, but zero or more genes may be removed.

Classification results for disease datasets

The adaptive version of DIRAC was used to classify all datasets from the original DIRAC publication. I compared results with and without the adaptive search, using both BioCarta signaling pathways and PPI link communities as starting network inputs. Overall, I found little difference in the performance of DIRAC with any particular set of parameters—some scenarios improved accuracies for particular datasets, but performed less well on other datasets. One possible limitation of this approach could be in the reliability of interactions in the PPI—i.e., any false positive interactions in the network may create additional noise rather than adding biological context. An alternative approach that I plan to explore is to use more carefully curated biochemical networks such as pathways in a metabolic reconstruction as input to the algorithm.

Pathway expression analysis with GSERA

A common feature of existing network analysis methods is a focus on individual networks—independent of other networks in the cell—and how they change between phenotypic states. However, researchers have frequently observed that multiple networks are perturbed in cancer cells [39, 43, 298, 299]. As studying interconnected or related genes on the network level can be highly informative for elucidating functional change in diseased cells, analyzing the combinatorial behavior of multiple related networks may lead to further biological insight. We can easily extend network analysis methods to address this

problem: by defining a measure of overall expression for a network, we can apply existing algorithms to "networks of networks," analogous to the more common networks comprised of participating genes. I previously developed an approach called Gene Set Expression Reversal Analysis (GSERA) to identify molecular signatures comprising multiple related networks. RXA refers to a class of algorithms that assess the ordering among gene expression values, rather than their absolute expression values (reviewed in Eddy et al. [300] and in **Chapter 3**); one searches for characteristic perturbations in this ordering from one phenotype to another, or within phenotypes. In particular, GSERA adapts the top-scoring pair (TSP) algorithm to study network families in disease.

GSERA is a pathway based analysis method that combines gene expression measurement over groups of genes that fall within common pathways for a more effective marker discovery. The pathways used were *a priori* defined and drawn from Gene Ontology [81] and KEGG [80] databases. The approach is based on the top scoring method [243] and extends mRNA expression level comparisons to comparisons of expression levels of pathways or gene sets between different disease classes. It can be applied for disease class prediction as well as identifying the biological signatures actuating disease progression. The method involves pathway aggregates of genes and only relative expression values. GSERA was used to study various cancer and non-cancer datasets including Parkinson's disease, prostate cancer, breast cancer, Marfan syndrome, bipolar disorder, an airway transcriptome dataset, and GIST & LMS tumors.

Calculating Gene set Expression Values

GSERA considers the experiment sets where gene expression profiles of several (on the order of thousands) individual genes are described for samples belonging to different phenotypic classes. These expression profiles are converted from being those for individual genes to the corresponding values for sets of genes or pathways. Thus, GSERA analysis begins with the calculation of gene set expression values from the gene expression values provided by microarray data. Various methods have been previously described for calculation of these gene set values given gene expression data. For ease of calculation and comparison with other such studies, in this study gene set expression values were calculated by taking the average of all gene expression values in a gene set. In addition to mean expression, other summary metrics such as median expression were also used to quantify gene set expression; no metric was optimal in all cases.

The gene set definitions, which describe the constituent "n" genes of each gene set were obtained from public databases. These definitions group genes based on their positions on a chromosome (positional gene sets), on their participation in pathways (biochemical, signaling, or regulatory) as defined in literature (curated gene sets), or as belonging to the same GO term in the Gene Ontology initiative (GO

gene set). The available Matlab implementation of the GSERA program allows use of each of these diverse gene set definitions and additional gene to gene set expression value conversion metrics.

Classification with and Limitations of GSERA

Being a gene-set expression value analysis, GSERA focuses on identifying biologically and statistically significant pathways and processes and exploits the inherent modular structure of biological networks to reveal network changes that accompany a disease process. The GSERA method was tested on different datasets using three different sets of *a priori* defined pathways. The estimated prediction rate of the classifier for each study was also compared to TSP, DIRAC, and SVM classifiers. In general, the method did not lead to significant improvement in classification accuracy compared to other algorithms.

A potential limitation of the method arises because although completely sequenced, the functional annotation of genes in the human genome is still an ongoing process. Due to this, current pathway definitions available are not completely representative of the biological system. Information gathered by microarray experiments is also not completely utilized due to this lack of complete annotation of all probes used on the array. Another difficulty faced by research groups analyzing gene expression data using pathway based methods is the lack of a universal annotation system for genes and pathways. In addition to typically unremarkable gains in classification compared to other tools, the relationships between pairs of pathways found by GSERA are often unclear. Restricting the search to biological processes connected in the gene ontology structure helps provide more tractable results, but both GO relationships and the terms themselves vary greatly in terms of reliability and interpretability.

Finally, and perhaps most importantly, the use of summary statistics such as mean or median to quantify the activity of multiple genes in a single pathway is far from ideal. For example, the expression level of a pathway could be easily skewed by very high or low expression of one or a few genes. The solution to this problem is not immediately obvious, but may involve using measures more closely tied to network-level functioning (such as metabolic fluxes). Still, examining the relationships between multiple gene sets—and how these relationships change in disease—remains a problem of great interest, and the general scheme of GSERA may provide useful in future studies.

Chapter 5 Figures

Figure 5.1. Within-state consistency in the relative expression of target genes within each module. Rows correspond to each TF and columns are the various pairwise comparisons.

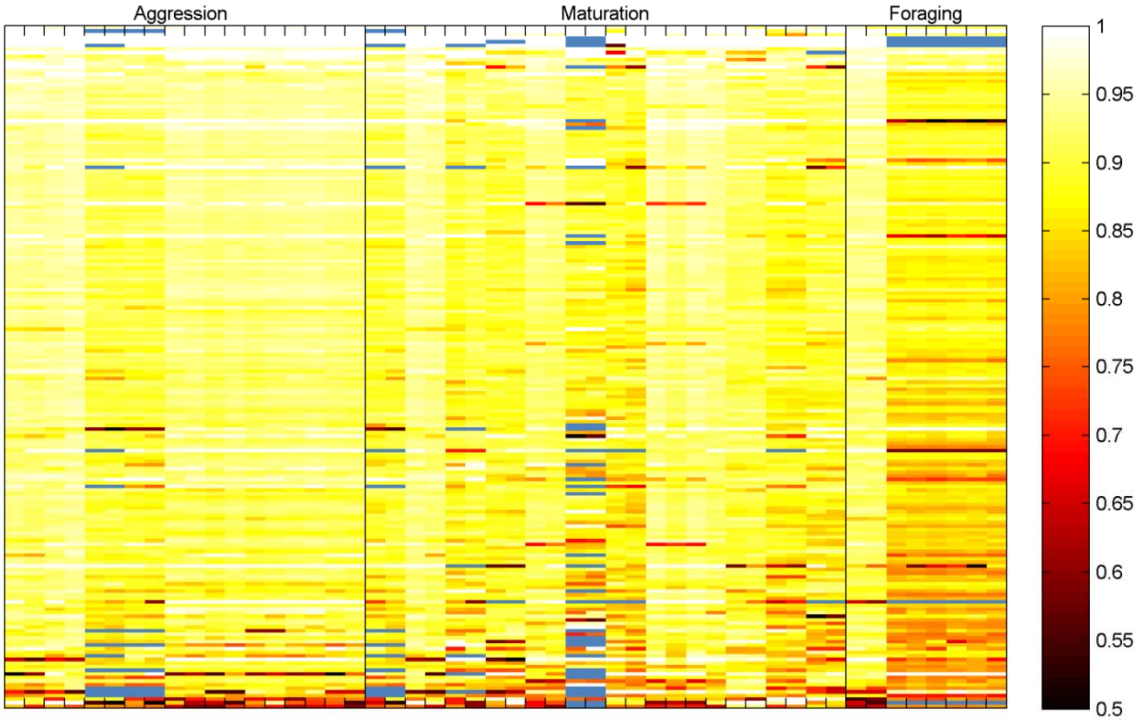


Figure 5.2. Between-state reordering of relative gene expression increases with module size.

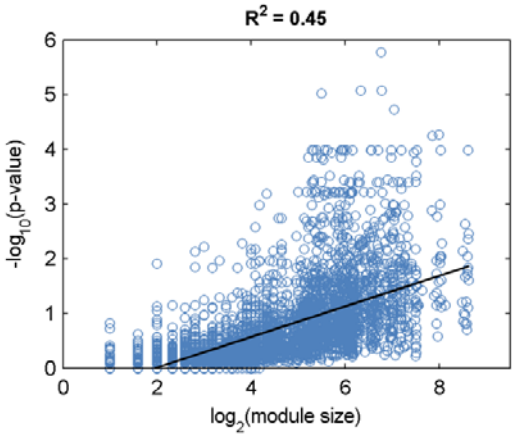


Figure 5.3. Factors that influence behavior over long vs. short timescales influence network states differently.

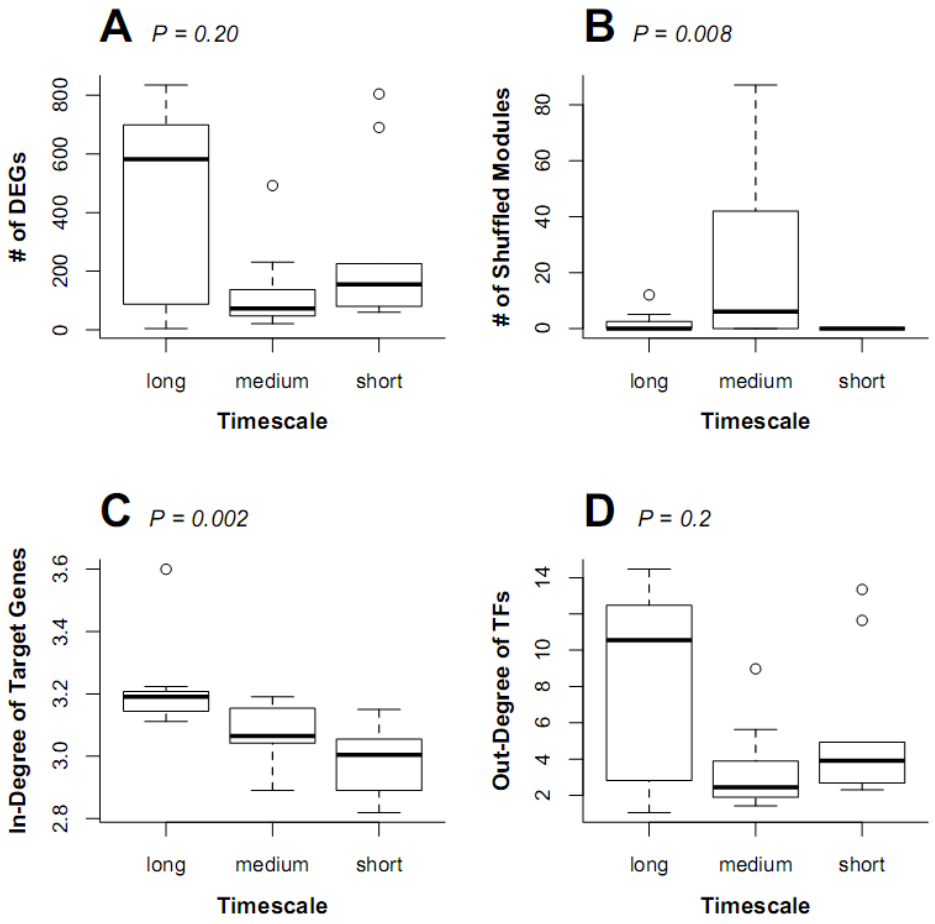


Figure 5.4. Spatial contiguity and size distribution of k-means clusters for neuron-specific genes at large k.

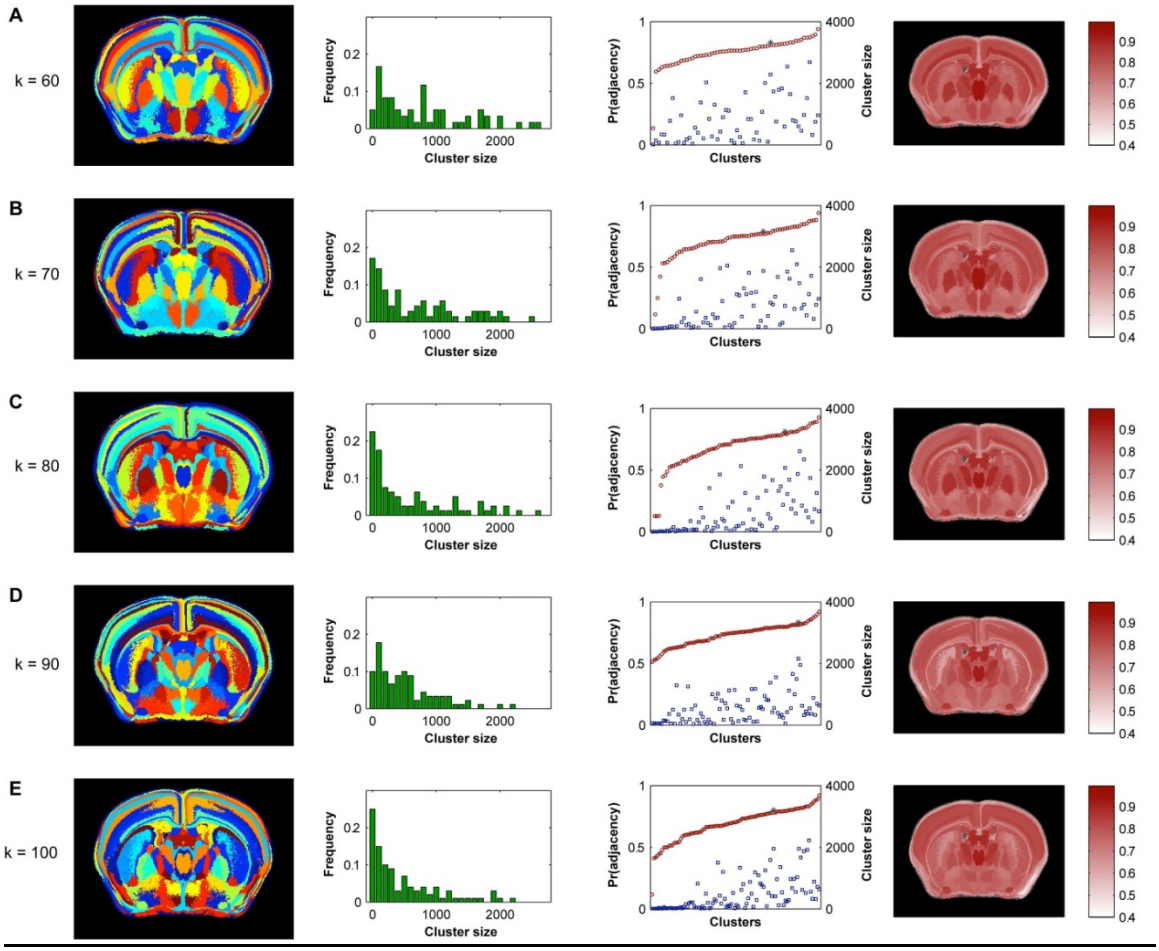
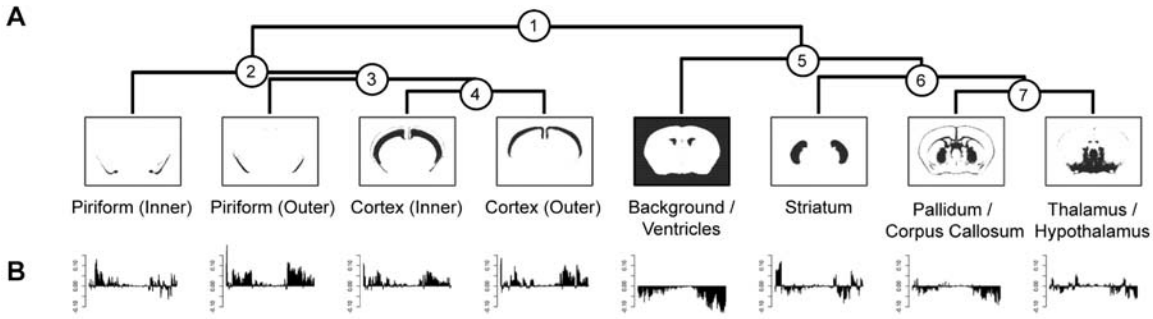


Figure 5.5. Clusters corresponding to major subdivisions in the mammalian brain are characterized by expression differences for individual marker genes and reordering of functionally related gene sets.



C

	Comparison	Representative Genes	Re-ordered gene sets
1	**A. Cortex vs. B. Non-cortex**	*A>B: (many)* *B>A: (none)*	Negative regulation of cell proliferation n = 3 genes*, Accuracy = 80%***
2	**A. Piriform (inner) vs. B. Cortex**	*A>B: Myo5b, Lmo3, Scn3b, Slc6a7* *B>A: Spock1, Stx1a, Satb2, Mef2c*	Membrane organization and biogenesis n = 2 genes*, Accuracy = 80%***
3	**A. Piriform (outer) vs. B. Cortex**	*A>B: L1cam, Nov, Scn3b, Slc7a14* *B>A: Satb2, Arpp21, Foxp1*	Regulation of transport n = 3 genes**, Accuracy = 79%***
4	**A. Cortex (inner) vs. B. Cortex (outer)**	*A>B: Prkcc, Satb2, Elavl2, Myo5b* *B>A: Cacna2d3, Calb1, Camk4*	Glycoprotein metabolic process n = 2 genes**, Accuracy = 76%***
5	**A. Sub-cortical Regions vs. B. Backgroun/Ventricles**	*A>B: (many)* *B>A: (none)*	Synaptic transmission n = 12 genes*, Accuracy = 89%***
6	**A. Striatum vs. B. Pallidum/Thalamus/ Hypothalamus**	*A>B: Cacna2d3, Camkv, Cdh8* *B>A: Ache, Stmn2, Gabra1*	Central nervous system development n = 6 genes, Accuracy = 74%***
7	**A. Pallidum vs. B. Thalamus/Hypothalamus**	*A>B: Snca, Nov, Mef2c, Bcl11a* *B>A: Atp1a3, Calb2, Grin1*	Synaptic transmission n = 12 genes, Accuracy = 64%***

Chapter 6. Evidence-driven reconstruction of a glioblastoma metabolic network: a platform for data integration and *in silico* investigation[6]

Glioblastoma multiforme (GBM) is the most common and aggressive type of primary brain tumor. Classified as grade IV astrocytoma, GBM develops from astrocytes (glial cells) and presents a median survival rate of about 15 months, even with surgery followed by radiation or chemotherapy. Importantly, perturbations to cellular networks in GBM eventually result in the symptoms of disease observed by the patient, including seizure, nausea, vomiting, headache, and a progressive deterioration of memory and personality. These phenotypes arise not from any single mutation, but from the sum effect of complex interactions among multiple aberrant genes and other molecular agents. The combinatorial nature of GBM tumor development therefore mandates a systems-level approach to elucidate underlying mechanisms of the cancer.

Cancer is particularly amenable to systems biology approaches because it is an intrinsically complex and heterogeneous disease [301]. The convolution of genetic effects, changes in gene and protein expression levels, and epigenetic modifications that give rise to malignancies like GBM further illustrates the complex, nonlinear relationship between molecular state and cellular cancer phenotype. Efforts in sequencing the human genome [302, 303] and individual cancers [39, 40] have made it increasingly clear that GBM and other cancers result not only from multiple perturbations, but from differing sets of mutations in every patient [39]. Such a distribution of mutations presents enormous challenges for personalized medicine, because it means that simple mutation to treatment correlations are not likely to be effective. Reconstructing detailed *in silico* models of biochemical reaction networks (e.g., metabolic, signaling, regulatory) at the genome scale establishes a platform on which different genetic perturbations can be related to emergent malignant functions and phenotypes.

Metabolism is arguably the best understood cellular process and is highly perturbed in oncogenesis, where cancer cells have increased metabolic rate to provide the energy needed for increased proliferation [41-43]. Thus, metabolism provides an ideal setting to begin reconstructing detailed biochemical networks at the genome-scale for GBM. Metabolite-based analyses to probe cancer have been performed since the 1980s [44] and have shown that cancer cells display distinct metabolic profiles, which can be characterized to diagnose the type and progression of disease, inform prognosis, and assess efficacy of therapy [45]. Metabolic phenotypes that remain consistent across cancer types, including GBM, reflect

[6] This chapter includes material that was reproduced with permission from the following publication:
Wang, Y., J.A. Eddy, and N.D. Price. 2012. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. BMC Systems Biology. 6:153. (***Metabolic Context-specificity Assessed by Deterministic Reaction Evaluation (mCADRE)*** section; methods, results, and text generated in collaboration with Yuliang Wang).

decreased aerobic respiration activity [46, 47] coupled with increased glycolysis [45, 46] and increased phospholipid production [46, 48]. These observations have led to targeted diagnosis and treatment strategies [45, 49]. In addition to broad spectrum cancer markers, metabolite signatures specific to astrocytomas have been found [304], which can be used to distinguish astrocytomas from other brain cancers.

Utilizing the recently published generic human metabolic reconstruction—a well-curated repository of reaction information—I have computationally integrated gene and protein expression data from multiple sources to define a network specific to GBM. Specifically, to generate a cell-type specific metabolic network for the U87 MG cell line, I used a collection of gene and protein expression evidence with an updated implementation of the mCADRE tool I co-developed. This tool is designed to weigh evidence and network connectivity in order to deterministically rank reactions in a generic (i.e., non-specific) metabolic model, prior to sequentially pruning the lowest ranked reactions from the network. Just as I have used mCADRE to generate the GBM-specific model, mCADRE can be used with other forms of gene or protein expression data to create cell- and tissue-specific metabolic models for other disease models. This can greatly aid in identifying areas of interest for experimental exploration and/or validation.

From this cell line-specific model, one is then able to model metabolic flux through an approach known as constraint-based modeling. Constraint-based modeling with the GBM network allows for simulation of cell growth under physiologically relevant conditions and investigation of mechanisms for tumor development. Curated pathways in the reconstructed network also provide functional context for statistical approaches to study gene expression in disease states. These combined approaches enable the identification and detailed characterization of key perturbed pathways in GBM.

Tissue-specific Metabolic Model Reconstruction Methods

While the need for tissue and cell type (henceforth referred to as context) specific metabolic models is strong, the number of such models is small. Manual curation often takes more than one person-year to produce a functional model [305], limiting the ability to produce a large number of context-specific functional metabolic models. Therefore, computational methods are clearly needed to speed up the model building process. The current state-of-the-art computational method to build context-specific models, the Model Building Algorithm (MBA) has been used to build metabolic models for liver, generic cancer, and two cancer cell lines [306-308]. However, the high computational cost/complexity involved with this method limits the ability to go through iterative rounds of model building and curation, which is typically

done in manual reconstruction of metabolic models [305]. It also limits the ability to build metabolic models for each individual tissue and cell types in an organ as complex as the brain.

Model Building Algorithm (MBA)

I initially explored using MBA, as described in Jerby et al. [306], to systematically prune reactions from the generic *Human Recon 1*, leaving a tissue- or cell-type-specific network. Given expression evidence specific to the cell type of interest, one first defines two core sets of reactions: high confidence reactions in tissue core 1 (TC^1) and moderate confidence reactions in tissue core 2 (TC^2); each core is a subset of the full list of reactions in *Recon 1*. The objective of MBA is to maximize the number of reactions included from TC^2 while minimizing the inclusion of non-core reactions, subject to the requirements that all reactions in TC^1 are consistent (i.e., can carry non-zero flux in at least one condition).

I developed a Matlab implementation of MBA that offers two convenient features: integration with the COBRA toolbox and optional use of fastFVA. The COBRA (Constraint-Based Reconstruction and Analysis) toolbox provides a suite of tools for constructing, curating, simulating, and analyzing metabolic network models in Matlab. To take advantage of the efficient and user-friendly features of the COBRA, I have written this version of MBA to utilize built-in functions for model parsing, organization, and optimization. I have also added an optional feature that significantly increases computational performance relative to a direct translation of the published MBA algorithm to Matlab. Specifically, the original version of MBA uses a heuristic optimization approach to evaluate the consistency of core reactions (see Jerby et al. for a detailed description of the *checkModelConsistency* module), reducing the number of iterations needed for the flux variability analysis (FVA) algorithm, and dramatically speeding up this step. Because FVA is still needed for the majority of calculations in the consistency check, this represents a substantial bottleneck in computational time. To further speed up the *checkModelConsistency* component of MBA, I have included the option to perform all remaining FVA iterations with the fastFVA algorithm; fastFVA was developed by Gudmunsson et al. [309], and has been demonstrated to improve performance of FVA by 30 to 220 fold.

Metabolic Context-specificity Assessed by Deterministic Reaction Evaluation (mCADRE)

Taking advantage of the Matlab implementation of MBA, my fellow graduate student, Yuliang Wang, and I recently developed a new method called *metabolic Context-specificity Assessed by Deterministic Reaction Evaluation (mCADRE)* (**Figure 6.1**), which achieves dramatic computational speed up by deterministically ranking all reactions based on gene expression data and metabolic network topology [310]. This deterministic decision-making in mCADRE, coupled with an automated pipeline of data

collection and processing, enables researchers to efficiently generate accurate and robust initial models from publicly available expression data (**Figure 6.1**). We used mCADRE to build a normal liver model and compared this model with the liver model built with MBA in the original publication [306]. mCADRE achieves dramatic computational speed up compared to MBA, and mCADRE models have better functionality compared to MBA models based on the same input data. The improvements afforded by mCADRE enable both iterative computational model curation as new high throughput data or literature evidence become available, and the reconstruction of metabolic models for a large number of different tissues and cell types.

Curating *Human Recon 2* and Evidence-based GBM Model Construction

To generate a cell-type specific metabolic network for the U87 MG cell line, I used a collection of gene and protein expression evidence with an updated implementation of our mCADRE tool (**Figure 6.2**). This tool is designed to weigh evidence and network connectivity in order to deterministically rank reactions in a generic (i.e., non-specific) metabolic model, prior to sequentially pruning the lowest ranked reactions from the network. The original implementation of mCADRE prohibits removal of any reactions that prevent the production of certain precursor metabolites from a single carbon substrate (i.e., glucose). Here, I assumed that cellular growth is a valid objective for the U87 cancer cells, and therefore imposed a requirement for biomass production given a defined medium. In other words, reactions could not be removed from the generic model if this action prevented cell growth.

Several preliminary steps were necessary to enable this approach, including collection of expression evidence, literature-based definition of the U87 biomass composition, modification of the generic model (i.e., *Human Recon 2*), and updating the mCADRE algorithm. The updated version of *Recon 2* described here represents only minor changes to the overall network composition along with expansion of meta-information such as alternate database identifiers for metabolites and reactions. As such, these incremental changes do not substantiate a new version of the master SBML file, but code is available to automatically recreate all steps for creating *Recon 2.01*.

Curation and priming of Human Recon 2

Requiring the cell-type specific U87 model to be capable of growth at all stages of network pruning mandated that the generic model be *a priori* capable of growth. While the *Human Recon 2* model extracted from the published SBML file can produce biomass "out of the box," the large number of metabolites that are allowed to freely exchange into and out of the cell are not reflective of relevant *in vitro* conditions. When this excessively rich medium was replaced with a more minimal defined medium

(comprising essential amino acids, essential fatty acids, glucose, and several other carbon-free and carbon-containing vitamins and minerals), however, growth could no longer be achieved. I identified specific components of the biomass equation that could not be produced, and traced back along their potential production routes to identify three specific deficiencies in the network:

i. *De novo* production of NAD from tryptophan
ii. Mitochondrial synthesis of NADP from NAD
iii. Synthesis of fatty acyl "R" groups for lumped lipid synthesis pathways

These issues were addressed as follows. (i) I identified the nicotinate-nucleotide diphosphorylase (NNDPR) reaction, previously included in *Human Recon 1*, which catalyzes the conversion of quinolinate and 5-phosphoribosyl diphosphate to nicotinate D-ribonucleotide. This reaction represents an essential step in the pathway from tryptophan (an essential amino acid) to NAD, and was previously annotated with the gene 23474. Without any clear indication of why this reaction had been removed, I added NNDPR to *Recon 2* with the previous meta-information. (ii) I identified a recent study reporting the isolation of a mitochondrial NAD kinase (encoded by the gene 133686). Based on this finding, NADKm was added to *Recon 2* with stoichiometry that mirrors the cytosolic NADK. (iii) Several reactions required for the ultimate synthesis of lipid biomass components from selected shorter-chain fatty acids were originally set to zero upper and lower bounds in *Recon 2*, effectively preventing any flux through these reactions.

Additionally, when pruning the modified version of *Recon 2* to a U87-specific network, I viewed the production of a more relevant U87 biomass composition to be a more suitable requirement than the built-in generic biomass definition. As such, I examined several published experimental studies to modify the overall fractional composition of macromolecules in a unit of biomass (gram of dry weight) as well as the specific distribution of key metabolites within each class of macromolecule. This U87 biomass reaction was added to *Recon 2.01*, along with two additional transport reactions required for the production of glycogen and ganglioside GM1, respectively, prior to pruning with mCADRE.

Collection of publicly available U87-specific expression evidence

Tissue-specific and cell-type specific expression data was obtained primarily from the Human Protein Atlas (HPA) and the Gene Expression Omnibus (GEO). As gene expression is often poorly correlated with protein expression—and by extension, enzyme activity—I have placed stronger evidence on direct protein expression measurements as evidence for the network reconstruction (defined below). The HPA contains protein expression measures for three different GBM cell lines: U138MG, U251MG, and U87

MG, as well as normal and cancer tissue samples from glial cells and malignant gliomas, respectively. Due to the lack of specificity in tissue sample expression data (no distinction between astrocytes, oligodendrocytes, etc.), I have chosen to focus on GBM cell line data thus far to drive the reconstruction of the metabolic network; gene expression data is likewise sparse and unspecific for normal or healthy glial cells.

For the U87 cell line, I generated a union list of proteins, combining those with evidence of strong expression in HPA, and those with evidence of moderate expression. A protein is designated as strongly or moderately expressed in a cell line based on evidence from immunohistochemistry, immunofluorescence, Western blot, or protein array. To control for quality, I kept only those proteins achieving a supportive validation score (as opposed to uncertain or not supportive) for at least one evidence type. As protein lists on HPA are reported as HTML tables for manual inspection only, I developed a Python-based module to enable rapid acquisition of expression evidence for a tissue or cell of interest. Protein expression evidence was mapped to genes in *Recon 2* based on NCBI gene ID. To supplement the protein expression evidence obtained from HPA, I compiled datasets from GEO with gene expression profiles for U87 samples.

Updating mCADRE to account for multiple sources of expression evidence

After *Recon 2* had been effectively primed as the starting point for mCADRE, I used three sources of evidence to rank the context-specificity of model reactions. First, I used raw Affymetrix microarray data for wildtype U87 cell lines from GEO. Using the MAS5 call, I binarized gene expression in this dataset as "expressed" or "not expressed." From the binarized data, I calculated "ubiquity" scores for each gene indicating the prevalence of expression across all samples. These ubiquity scores represent the primary scale by which reactions are ranked in mCADRE. Second, I used RNA-seq gene expression data for wildtype U87 cell lines in multiple conditions. If ubiquity scores for any genes among the RNA-seq data were higher than in the microarray data, the former values was used. Finally, I referenced lists of high and moderately staining proteins in U87 cells from HPA. Given the relatively weak link between gene expression and enzyme function in a metabolic network, I placed greater weight on protein expression when ranking reactions. Any gene for which there was corresponding protein expression evidence in HPA was automatically assigned a perfect ubiquity score; this ensured that the associated reaction would be included in the high-confidence core prior to pruning in mCADRE.

Generation of a genome-scale model of the U87 MG GBM cell line

After running mCADRE, the pruned model contained 3,058 reactions and 2,032 metabolites. Notably, an initial step of processing and ranking of reactions in mCADRE involves removal of all blocked reactions (2,836 total); our gene and protein expression data suggested strong evidence for 919 of these reactions, and these were added back to the pruned model. The final U87-specific metabolic model includes 3,977 genes and 3,274 metabolites corresponding to 1,583 genes. The 996 blocked reactions in the final model highlight gaps in biochemical knowledge—given that these reactions are expected to be present based on strong expression evidence, such gaps would be higher priority targets of future investigation and curation. Importantly, including these reactions and genes enables greater coverage of omics data.

Conclusions & Future Directions

Using the mechanistically detailed model I have built for the U87 GBM cell line, one can begin to interrogate the link between molecular changes in the cancer and key metabolic processes that contribute to tumorigenesis. The model also serves as a powerful tool for studying pathway-level changes in GBM omics measurements. In the future, systems analysis of GBM with the model should continue to provide important biological insights into underlying mechanisms and potential treatment routes for the disease.

Notably, biologists in my lab have begun conducting experiments and generating novel high-throughput measurements for U87 cells under several genetic and environmental conditions. These cells and data will enable multiple additional steps for validating and interrogating the metabolic model. First, using the model, I will attempt to predict qualitative effects (increased or decreased growth, ATP production, oxygen consumption) of numerous inactivated reactions, based on an experimental screen of enzyme inhibitors in U87 cells. Importantly, incorrect predictions point to areas of the network that will need future refinement or that might be subject to additional layers of regulation.

Second, by integrating RNA-seq and targeted metabolomics measurements from U87 cells treated with vehicle or EGF, we will be able to identify metabolic pathways enriched for differential expression of genes or metabolites in these conditions. As discussed above, networks defined *a priori* in pathway databases may not be representative of functioning sets of genes in a particular phenotype. The metabolic network, to be reconstructed based on evidence specific to GBM, will already divided into functional subnetworks—based on known biochemical pathways—that can serve as inputs to network-level analysis tools; these subnetworks can even be grouped into broader categories of metabolism (e.g., energy metabolism) to define network families. When operating on high-confidence pathways in the GBM metabolic network, DIRAC and GSERA should produce more robust and mechanistically relevant

results. In turn, perturbed pathways identified will serve as the focal point for future simulations and model development.

Additionally, GBM was one of the first cancers studied by The Cancer Genome Atlas (TCGA), and remains a major focus of the program. The TCGA has collected data for several hundred GBM tumors—including sequence, copy number, methylation, and expression data, all of which is publically available [299]. The distribution of mutations in GBM has been shown to differ from patient to patient. By examining the frequency of mutations within different biological networks, such as metabolic pathways in the genome-scale model, we may be able to identify functional sets of genes that are consistently altered in the cancer. I will map compiled mutational data onto corresponding enzymes in the GBM metabolic network, and apply statistical tests for enrichment or over-representation to identify key subnetworks that tend to be perturbed in GBM.

Chapter 6 Figures

Figure 6.1. Schematic overview of the mCADRE algorithm for generating tissue-specific metabolic network models based on expression evidence and network connectivity.

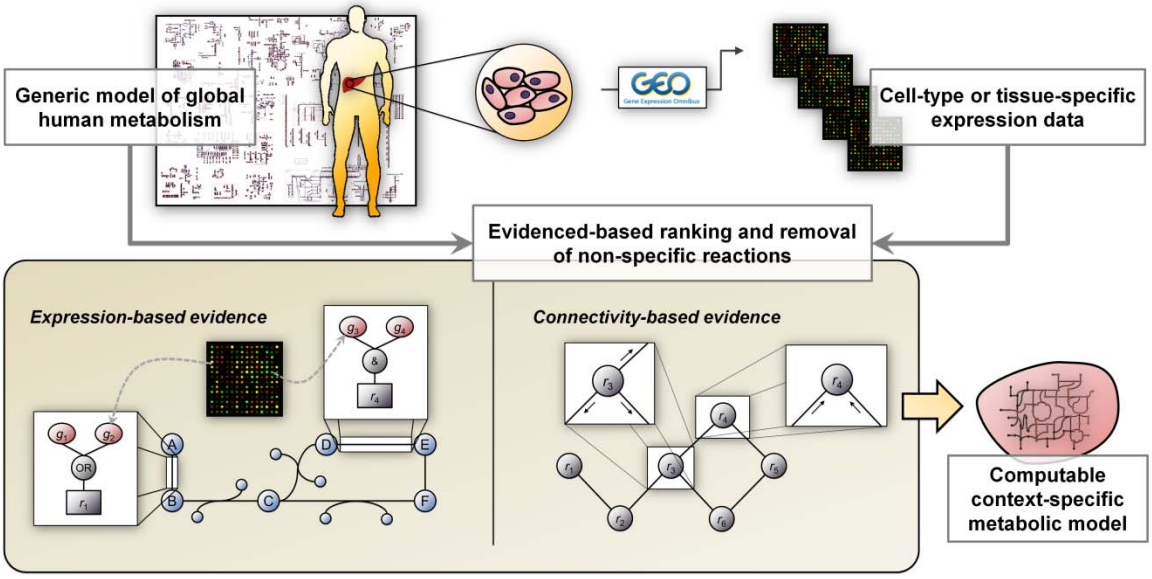
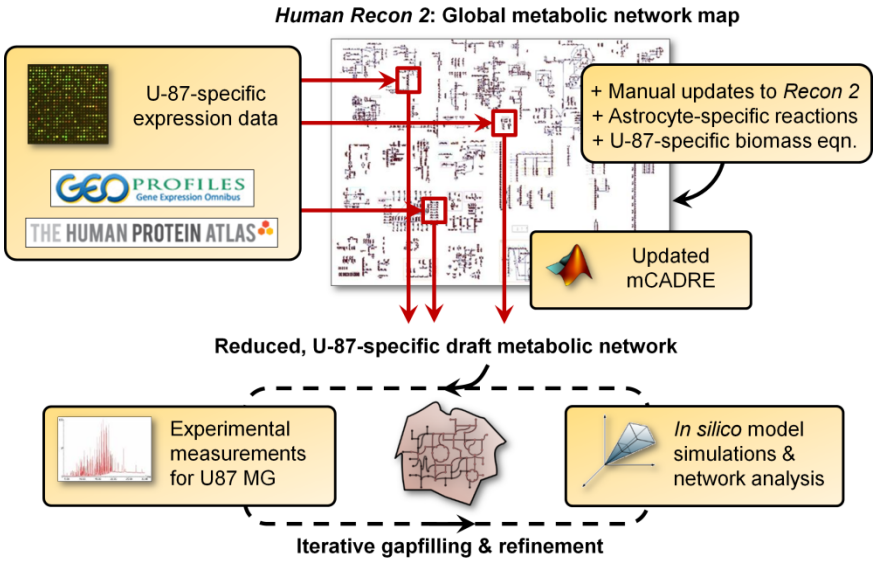


Figure 6.2. Evidence based reconstruction of the U87 glioblastoma metabolic network.



Chapter 7. Conclusion

Systems biology employs an integrative approach to characterize biological systems, in which interactions among all components in a system are described mathematically to establish a computable model that integrates this data into a cohesive whole. These *in silico* models can be simulated to quantitatively study the emergent behavior of a system of interacting components. Integrating heterogeneous dynamic data into quantitative predictive models holds great promise to significantly increase our ability to understand and rationally intervene in disease-perturbed biological systems. This promise—particularly with regards to personalized medicine and medical intervention—has motivated the development of new methods for systems analysis of human biology and disease.

Despite the exciting potential and incremental progress made in systems biology to date, the field as a whole has yet to deliver particularly remarkable or transformational contributions to either health or industry. While this can in some ways be attributed to the monumental complexity in higher organisms like mice and humans, it is clear that there is a great deal we still do not understand—and are thus unable to effectively model—about microbial behavior as well. Ongoing efforts in model and method development are focused on accounting for increasing amounts of cellular complexity with more powerful and sophisticated approaches. On the other hand, the incredible amount of high-throughput data that continues to be generated in almost all areas of biological research demand ever-improving computational tools to integrate and make sense of these measurements. Moving forward, a key goal in modeling efforts will be the integration of multiple network types, which will in turn enable the integration, analysis, and simulation of multiple types of omics data.

Constraint-based modeling is a well-established framework for linking genotype to phenotype, and genome-scale metabolic models have been constructed for an ever-expanding array of organisms, spanning all branches of life. While the potential of such models for both basic and applied science has been demonstrated, extensive opportunity remains for expanding this framework to incorporate and interpret multiple types of omics data. Successive application of a hierarchy of constraints (e.g., transcriptional regulatory, translational, post-translational) within an integrated theoretical framework will enable progress towards a comprehensive understanding of emergent properties of biological systems (**Figure 7.1**). Metabolic network models will serve as a chassis for the integration of other biochemical and statistical network modeling frameworks.

A key aspect of many diseases—especially cancer and viral infections—is the interaction between foreign agents or transformed cells and various players of the host immune system. While methods for molecular network modeling that focus directly on the diseased systems have been highly informative, accounting

for the complex interplay between cancer or viruses and immune cells may reveal important new understanding of underlying mechanisms. Approaches to model such tissue-level interactions can effectively extend intracellular molecular interactions to tissue-level interactions in an unprecedented way. As the interplay between host and disease cells ultimately determines clinical outcome, considering the effects of larger scale interactions on intracellular network behavior will greatly enhance model predictive power.

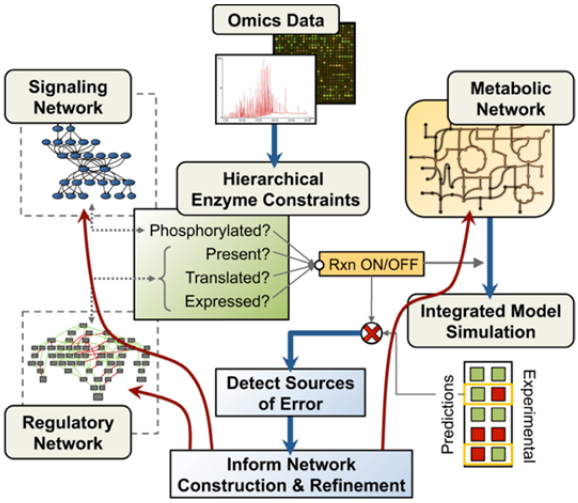
Another important consideration for future efforts in systems biology is not to just account for more complex and more types of data, but to establish a more rigorous understanding of model and method capabilities. *Saccharomyces cerevisiae* (baker's yeast) is not only a model lab microorganism, industrial workhorse, and one of the best-characterized eukaryotes, but the extent to which its metabolic network has been studied, modeled, and curated is second only to that of *E. coli*. Yeast provides an ideal platform to develop and test approaches for integrated modeling and networks because (i) publically available high-quality datasets exist for a range of strains and conditions, and (ii) yeast utilizes regulatory mechanisms that are conserved across many eukaryotes, facilitating application of novel modeling approaches to other organisms. Lessons learned from more carefully examining, curating, and validating model predictions with yeast will facilitate efforts to understand more complex systems in higher organisms.

Finally, it will remain critical to interface computational and modeling efforts with experimental work and domain expertise. The honeybee project (described above) to which DIRAC was applied—a collaborative effort with experts in insect biology and neuroscience—is also an excellent example of the impetus to combine computational expertise with deep biological knowledge of leading scientists. Similarly, I led the computational gene expression analysis component of a larger experimental study of cellular processes in sarcomas being conducted at the University of Texas M.D. Anderson Cancer Center. In both of these studies, modeling approaches provided additional powerful ways to interrogate large amounts of data, but biological expertise was essential for gleaning useful insights from analysis results.

Ultimately, modeling efforts in systems biology should enable predictive, mechanistic interrogation of biomolecular networks in disease, and establish a platform for interpreting large cohorts of patient data with stronger functional context. Integrating heterogeneous data into these quantitative models will not only increase our ability to understand and rationally intervene in disease-perturbed biological systems, but facilitate rapid and powerful hypothesis generation to guide experiments and drive the expansion and refinement of biological network knowledge.

Chapter 7 Figures

Figure 7.1. Iterative process of data incorporation and network refinement through integrated modeling of metabolism and hierarchical regulatory constraints.



Bibliography

1. Hanahan, D. and R.A. Weinberg, *The Hallmarks of Cancer.* Cell, 2000. **100**(1): p. 57-70.
2. Land, H., L.F. Parada, and R.A. Weinberg, *Tumorigenic conversion of primary embryo fibroblasts requires at least two cooperating oncogenes.* Nature, 1983. **304**(5927): p. 596-602.
3. Lloyd, A.C., et al., *Cooperating oncogenes converge to regulate cyclin/cdk complexes.* Genes & Development, 1997. **11**(5): p. 663-677.
4. Fanidi, A., E.A. Harrington, and G.I. Evan, *Cooperative interaction between c-myc and bcl-2 proto-oncogenes.* Nature, 1992. **359**(6395): p. 554-556.
5. Lowe, S.W., E. Cepero, and G. Evan, *Intrinsic tumour suppression.* Nature, 2004. **432**(7015): p. 307-15.
6. McMurray, H.R., et al., *Synergistic response to oncogenic mutations defines gene class critical to cancer phenotype.* Nature, 2008. **453**(7198): p. 1112.
7. Chuang, H.Y., et al., *Network-based classification of breast cancer metastasis.* Molecular Systems Biology, 2007. **3**(140).
8. Liu, E.T. and T. Lemberger, *Higher order structure in the cancer transcriptome and systems medicine.* Molecular Systems Biology, 2007. **3**(94).
9. Auffray, C., *Protein subnetwork markers improve prediction of cancer outcome.* Molecular Systems Biology, 2007. **3**(141).
10. Neely, K.E. and J.L. Workman, *The complexity of chromatin remodeling and its links to cancer.* Biochimica et Biophysica Acta, 2002. **1603**(1): p. 19-29.
11. Seligson, D.B., et al., *Global histone modification patterns predict risk of prostate cancer recurrence.* Nature, 2005. **435**(7046): p. 1262.
12. Esteller, M., *Cancer epigenomics: DNA methylomes and histone-modification maps.* DNA, 2007. **1**(E2): p. E1.
13. Jones, P.A., *DNA methylation and cancer.* Oncogene, 2002. **21**: p. 5358-5360.
14. Esteller, M., et al., *Cancer epigenetics and methylation.* Science, 2002. **297**(5588): p. 1807-8.
15. Laird, P.W., *Cancer epigenetics.* Human Molecular Genetics, 2005. **14**(90001): p. 65-76.
16. Tang, Y., et al., *Extracellular Matrix Metalloproteinase Inducer Stimulates Tumor Angiogenesis by Elevating Vascular Endothelial Cell Growth Factor and Matrix Metalloproteinases.* Cancer Research, 2005. **65**(8): p. 3193-3199.
17. Condeelis, J. and J.W. Pollard, *Macrophages: Obligate Partners for Tumor Cell Migration, Invasion, and Metastasis.* Cell, 2006. **124**(2): p. 263-266.
18. Lewis, C.E. and J.W. Pollard, *Distinct Role of Macrophages in Different Tumor Microenvironments.* Cancer Research, 2006. **66**(2): p. 605-612.
19. Kalluri, R. and M. Zeisberg, *Fibroblasts in cancer.* Nature Reviews Cancer, 2006. **6**(5): p. 392.
20. Bhowmick, N.A., E.G. Neilson, and H.L. Moses, *Stromal fibroblasts in cancer initiation and progression.* Nature, 2004. **432**(7015): p. 332-337.
21. Sheu, B.C., et al., *Cytokine regulation networks in the cancer microenvironment.* Frontiers in Bioscience, 2008. **13**: p. 6255-68.
22. De Luca, A., et al., *The role of the EGFR signaling in tumor microenvironment.* Journal of Cellular Physiology, 2008. **214**(3): p. 559-67.
23. Allinen, M., et al., *Molecular characterization of the tumor microenvironment in breast cancer.* Cancer Cell, 2004. **6**(1): p. 17-32.
24. Smallbone, K., et al., *The role of acidity in solid tumour growth and invasion.* Journal of Theoretical Biology, 2005. **235**(4): p. 476-484.
25. Smallbone, K., et al., *Metabolic changes during carcinogenesis: Potential impact on invasiveness.* Journal of Theoretical Biology, 2007. **244**(4): p. 703-713.
26. Gajewski, T.F., Y. Meng, and H. Harlin, *Immune Suppression in the Tumor Microenvironment.* Journal of Immunotherapy, 2006. **29**(3): p. 233.
27. Gajewski, T.F., et al., *Immune resistance orchestrated by the tumor microenvironment.* Immunological Reviews, 2006. **213**(1): p. 131-145.
28. Croci, D.O., et al., *Dynamic cross-talk between tumor and immune cells in orchestrating the immunosuppressive network at the tumor microenvironment.* Cancer Immunology, Immunotherapy, 2007. **56**(11): p. 1687-1700.

29. Laakkonen, P., et al., *Vascular Endothelial Growth Factor Receptor 3 Is Involved in Tumor Angiogenesis and Growth.* Cancer Research, 2007. **67**(2): p. 593.
30. Su, J.L., et al., *The role of the VEGF-C/VEGFR-3 axis in cancer progression.* British Journal of Cancer, 2007. **96**: p. 541-545.
31. Hirakawa, S., et al., *VEGF-C-induced lymphangiogenesis in sentinel lymph nodes promotes tumor metastasis to distant sites.* Blood, 2007. **109**(3): p. 1010.
32. Taranawski, R., et al., *Repopulation of tumour cells during radiotherapy is doubled during treatment gaps.* Computational and Mathematical Methods in Medicine, 2000. **2**(4): p. 297-305.
33. Kohandel, M., S. Sivaloganathan, and A. Oza, *Mathematical modeling of ovarian cancer treatments: Sequencing of surgery and chemotherapy.* Journal of Theoretical Biology, 2006. **242**(1): p. 62-68.
34. Verschraegen, C., et al., *Modeling the Effect of Tumor Size in Early Breast Cancer.* Annals of Surgery, 2005. **241**(2): p. 309.
35. Ayati, B.P., G.F. Webb, and A.R.A. Anderson, *Computational Methods and Results for Structured Multiscale Models of Tumor Invasion.* Multiscale Modeling & Simulation, 2006. **5**: p. 1.
36. Nordling, C.O., *A new theory on cancer-inducing mechanism.* British Journal of Cancer, 1953. **7**(1): p. 68-72.
37. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* 2001.
38. Venter, J.C., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**: p. 860-921.
39. Parsons, D.W., et al., *An Integrated Genomic Analysis of Human Glioblastoma Multiforme.* Science, 2008. **321**(5897): p. 1807.
40. Sjoblom, T., et al., *The consensus coding sequences of human breast and colorectal cancers.* Science, 2006. **314**(5797): p. 268-74.
41. Hsu, P.P. and D.M. Sabatini, *Cancer Cell Metabolism: Warburg and Beyond.* Cell, 2008. **134**(5): p. 703-707.
42. Cairns, R.A., I.S. Harris, and T.W. Mak, *Regulation of cancer cell metabolism.* Nature Reviews Cancer, 2011. **11**(2): p. 85-95.
43. Kroemer, G. and J. Pouyssegur, *Tumor Cell Metabolism: Cancer's Achilles' Heel.* Cancer Cell, 2008. **13**(6): p. 472-82.
44. Jellum, E., et al., *Classification of human cancer cells by means of capillary gas chromatography and pattern recognition analysis.* Journal of Chromatography A, 1981. **217**: p. 231-7.
45. Spratlin, J.L., N.J. Serkova, and S.G. Eckhardt, *Clinical applications of metabolomics in oncology: a review.* Clinical Cancer Research, 2009. **15**(2): p. 431-40.
46. Roslin, M., et al., *Baseline levels of glucose metabolites, glutamate and glycerol in malignant glioma assessed by stereotactic microdialysis.* Journal of Neuro-Oncology, 2003. **61**(2): p. 151-60.
47. Serkova, N.J. and K. Glunde, *Metabolomics of cancer.* Methods in Molecular Biology, 2009. **520**: p. 273-95.
48. Glunde, K. and N.J. Serkova, *Therapeutic targets and biomarkers identified in cancer choline phospholipid metabolism.* Pharmacogenomics, 2006. **7**(7): p. 1109-23.
49. Pelicano, H., et al., *Glycolysis inhibition for anticancer treatment.* Oncogene, 2006. **25**(34): p. 4633-46.
50. Feist, A.M., et al., *Reconstruction of biochemical networks in microbial organisms.* Nature Reviews Microbiology, 2009. **7**(2): p. 129-143.
51. Francke, C., R.J. Siezen, and B. Teusink, *Reconstructing the metabolic network of a bacterium from its genome.* Trends in Microbiology, 2005. **13**(11): p. 550-558.
52. Reed, J.L., et al., *Towards multidimensional genome annotation.* Nature Reviews Genetics, 2006. **7**(2): p. 130-141.
53. Feist, A.M., et al., *A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.* Molecular Systems Biology, 2007. **3**(121).
54. Feist, A.M., et al., *Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri.* Molecular Systems Biology, 2006. **2**(2006.0004).
55. Herrgard, M.J., et al., *A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology.* Nature Biotechnology, 2008. **26**(10): p. 1155-1160.
56. Chavali, A.K., et al., *Systems analysis of metabolism in the pathogenic trypanosomatid Leishmania major.* Molecular Systems Biology, 2008. **4**(177).
57. Price, N.D., J.L. Reed, and B.O. Palsson, *Genome-scale models of microbial cells: evaluating the consequences of constraints.* Nature Reviews Microbiology, 2004. **2**(11): p. 886-897.

58. Price, N.D., et al., *Genome-scale microbial in silico models: the constraints-based approach.* Trends Biotechnol, 2003. **21**(4): p. 162-9.
59. Milne, C.B., et al., *Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology.* Biotechnology Journal, 2009. **4**(12): p. 1653-70.
60. Oberhardt, M.A., B.O. Palsson, and J.A. Papin, *Applications of genome-scale metabolic reconstructions.* Molecular Systems Biology, 2009. **5**: p. 320.
61. Nevins, J.R. and A. Potti, *Mining gene expression profiles: expression signatures as cancer phenotypes.* Nature reviews. Genetics, 2007. **8**(8): p. 601.
62. Coe, B.P., et al., *Evolving strategies for global gene expression analysis of cancer.* Journal of cellular physiology, 2008. **217**(3): p. 590.
63. Allison, D.B., et al., *Microarray data analysis: from disarray to consolidation and consensus.* Nature Reviews Genetics, 2006. **7**(1): p. 55-65.
64. Mardis, E.R., *The impact of next-generation sequencing technology on genetics.* Trends in Genetics, 2008. **24**(3): p. 133-141.
65. Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nature Biotechnology, 2008. **26**(10): p. 1135-1145.
66. Gingras, A.C., et al., *Analysis of protein complexes using mass spectrometry.* Nature reviews. Molecular cell biology, 2007. **8**(8): p. 645.
67. Cox, J. and M. Mann, *Is proteomics the new genomics?* Cell, 2007. **130**(3): p. 395.
68. Bantscheff, M., et al., *Quantitative mass spectrometry in proteomics: a critical review.* Analytical and bioanalytical chemistry, 2007. **389**(4): p. 1017.
69. Vienna, A., et al., *Transforming omics data into context: Bioinformatics on genomics and proteomics raw data.* Electrophoresis, 2006. **27**: p. 2659-2675.
70. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Research, 2002. **30**(1): p. 207.
71. Sherlock, G., et al., *The Stanford Microarray Database.* Nucleic Acids Research, 2001. **29**(1): p. 152-155.
72. Perier, R.C., T. Junier, and P. Bucher, *The Eukaryotic Promoter Database EPD.* Nucleic Acids Research, 1998. **26**(1): p. 353-357.
73. Zhao, F., et al., *TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies.* Nucleic Acids Research, 2005. **33**: p. D103.
74. Martens, L., et al., *PRIDE: The proteomics identifications database.* Proteomics, 2005. **5**(13): p. 3537–3545.
75. Camon, E., et al., *The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology.* Nucleic Acids Research, 2004. **32**(90001): p. W313-W317.
76. Apweiler, R., et al., *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Research, 2004. **32**: p. D115.
77. Camon, E., et al., *The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro.* Genome Research, 2003. **13**(4): p. 662-672.
78. Flicek, P., et al., *Ensembl 2008.* Nucleic Acids Research, 2008. **36**(Database issue): p. D707.
79. Sterk, P., P.J. Kersey, and R. Apweiler, *Genome reviews: standardizing content and representation of information about complete genomes.* OMICS: A Journal of Integrative Biology, 2006. **10**(2): p. 114-118.
80. Kanehisa, M. and S. Goto, *KEGG: Kyoto encyclopedia of genes and genomes.* Nucleic Acids Research, 2000. **28**(1): p. 27.
81. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nature Genetics, 2000. **25**(1): p. 25-9.
82. Krieger, C.J., et al., *MetaCyc: a multiorganism database of metabolic pathways and enzymes.* Nucleic Acids Research, 2004. **32**(Database Issue): p. D438.
83. Karp, P.D., et al., *Expansion of the BioCyc collection of pathway/genome databases to 160 genomes.* Nucleic Acids Research, 2005. **33**(19): p. 6083.
84. DeJongh, M., et al., *Toward the automated generation of genome-scale metabolic networks in the SEED.* BMC Bioinformatics, 2007. **8**(1): p. 139.
85. Ren, Q., K. Chen, and I.T. Paulsen, *TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels.* Nucleic Acids Research, 2006.
86. Le Novère, N., et al., *BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems.* Nucleic Acids Research, 2006. **34**: p. D689-D691.

87. Sivakumaran, S., et al., *The Database of Quantitative Cellular Signaling: management and analysis of chemical kinetic models of signaling networks.* Bioinformatics, 2003. **19**(3): p. 408-415.
88. Ji, Z.L., et al., *KDBI: Kinetic Data of Bio-molecular Interactions database.* Nucleic Acids Research, 2003. **31**(1): p. 255.
89. Platzer, A., et al., *Characterization of protein-interaction networks in tumors.* BMC Bioinformatics, 2007. **8**: p. 224.
90. Ramani, A.K., et al., *Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome.* Genome Biology, 2005. **6**(5): p. R40.
91. Xenarios, I., et al., *DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.* Nucleic Acids Research, 2002. **30**(1): p. 303.
92. Zanzoni, A., et al., *MINT: a Molecular INTeraction database.* FEBS Letters, 2002. **513**(1): p. 135-140.
93. Pagel, P., et al., *The MIPS mammalian protein-protein interaction database.* Bioinformatics, 2005. **21**(6): p. 832-834.
94. Lloyd, C.M., M.D.B. Halstead, and P.F. Nielsen, *CellML: its future, present and past.* Progress in Biophysics and Molecular Biology, 2004. **85**(2-3): p. 433-450.
95. Le Novère, N., et al., *Minimum information requested in the annotation of biochemical models (MIRIAM).* Nature Biotechnology, 2005. **23**: p. 1509-1515.
96. Hucka, M., et al., *The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.* Bioinformatics, 2003. **19**(4): p. 524-531.
97. Stromback, L. and P. Lambrix, *Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX.* Bioinformatics, 2005. **21**(24): p. 4401-4407.
98. Shannon, P.T., et al., *The Gaggle: An open-source software system for integrating bioinformatics software and data sources.* BMC Bioinformatics, 2006. **7**: p. 176.
99. Shannon, P., et al., *Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.* Genome Research, 2003. **13**(11): p. 2498-2504.
100. Cusick, M.E., et al., *Literature-curated protein interaction datasets.* 2008.
101. Taylor, C.F., et al., *Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project.* Nature Biotechnology, 2008. **26**(8): p. 889-96.
102. Brazma, A., et al., *Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.* Nature Genetics, 2001. **29**(4): p. 365-71.
103. Taylor, C.F., et al., *The minimum information about a proteomics experiment (MIAPE).* Nature Biotechnology, 2007. **25**(8): p. 887-93.
104. Venkatesan, K., et al., *An empirical framework for binary interactome mapping.* Nature Methods, 2008.
105. Braun, P., et al., *An experimentally derived confidence score for binary protein-protein interactions.* Nature methods, 2009. **6**(1): p. 91.
106. Quackenbush, J., *Microarray Analysis and Tumor Classification.* New England Journal of Medicine, 2006. **354**(23): p. 2463.
107. Davis, C.A., et al., *Reliable gene signatures for microarray classification: assessment of stability and performance.* Bioinformatics, 2006. **22**(19): p. 2356.
108. de Souto, M.C.P., et al., *Clustering cancer gene expression data: a comparative study.* BMC Bioinformatics, 2008. **9**(1): p. 497.
109. Alizadeh, A.A., et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature, 2000. **403**: p. 503-511.
110. Bullinger, L., et al., *Use of Gene-Expression Profiling to Identify Prognostic Subclasses in Adult Acute Myeloid Leukemia.* New England Journal of Medicine, 2004. **350**(16): p. 1605-1616.
111. Naderi, A., et al., *A gene-expression signature to predict survival in breast cancer across independent data sets.* Oncogene, 2006. **26**(10): p. 1507.
112. Brown, M.P.S., et al., *Knowledge-based analysis of microarray gene expression data by using support vector machines.* Proceedings of the National Academy of Sciences, 2000. **97**(1): p. 262-267.
113. Khan, J., et al., *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.* Nature Medicine, 2001. **7**: p. 673-679.
114. Herrero, J., A. Valencia, and J. Dopazo, *A hierarchical unsupervised growing neural network for clustering gene expression patterns.* Bioinformatics, 2001. **17**(2): p. 126-136.
115. Geman, D., *Classifying Gene Expression Profiles from Pairwise mRNA Comparisons.* Statistical Applications in Genetics and Molecular Biology, 2004. **3**.

116. Tan, A.C., et al., *Simple decision rules for classifying human cancers from gene expression profiles.* Bioinformatics, 2005. **21**(20): p. 3896-3904.
117. Xu, L., D. Geman, and R.L. Winslow, *Large-scale integration of cancer microarray data identifies a robust common cancer signature.* BMC Bioinformatics, 2007. **8**: p. 275.
118. Price, N.D., et al., *Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas.* Proceedings of the National Academy of Sciences, 2007. **104**(9): p. 3414.
119. Chen, H.Y., et al., *A Five-Gene Signature and Clinical Outcome in Non-Small-Cell Lung Cancer.* New England Journal of Medicine, 2007. **356**(1): p. 11.
120. Michiels, S., S. Koscielny, and C. Hill, *Prediction of cancer outcome with microarrays: a multiple random validation strategy.* Lancet, 2005. **365**(9458): p. 488-492.
121. Yeoh, E.J., et al., *Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling.* Cancer Cell, 2002. **1**(2): p. 133-143.
122. Foekens, J.A., et al., *Multicenter validation of a gene expression-based prognostic signature in lymph node-negative primary breast cancer.* Journal of Clinical Oncology, 2006. **24**(11): p. 1665-71.
123. Ayers, M., et al., *Gene Expression Profiles Predict Complete Pathologic Response to Neoadjuvant Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide Chemotherapy in Breast Cancer.* Journal of Clinical Oncology, 2004. **22**(12): p. 2284.
124. Dressman, H.K., et al., *Gene Expression Profiles of Multiple Breast Cancer Phenotypes and Response to Neoadjuvant Chemotherapy.* Clinical Cancer Research, 2006. **12**(3): p. 819-826.
125. Thuerigen, O., et al., *Gene Expression Signature Predicting Pathologic Complete Response With Gemcitabine, Epirubicin, and Docetaxel in Primary Breast Cancer.* Journal of Clinical Oncology, 2006. **24**(12): p. 1839.
126. Karolchik, D., et al., *The UCSC Genome Browser Database: 2008 update.* Nucleic Acids Research, 2008. **36**(Database issue): p. D773.
127. Kanehisa, M., et al., *The KEGG databases at GenomeNet.* Nucleic Acids Research, 2002. **30**(1): p. 42.
128. Subramanian, A., et al., *GSEA-P: a desktop application for Gene Set Enrichment Analysis.* Bioinformatics, 2007. **23**(23): p. 3251.
129. Kim, S.Y. and D.J. Volsky, *PAGE: Parametric Analysis of Gene Set Enrichment.* BMC Bioinformatics, 2005. **6**: p. 144.
130. Subramanian, A., et al., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences, 2005. **102**(43): p. 15545-15550.
131. Rhodes, D.R. and A.M. Chinnaiyan, *Integrative analysis of the cancer transcriptome.* Nature Genetics, 2005. **37**(6 Suppl): p. S31-S37.
132. Shen, R., A.M. Chinnaiyan, and D. Ghosh, *Pathway analysis reveals functional convergence of gene expression profiles in breast cancer.* BMC Medical Genomics, 2008. **1**: p. 28.
133. Tomlins, S.A., et al., *Integrative molecular concept modeling of prostate cancer progression.* Nature Genetics, 2006. **39**: p. 41-51.
134. Lee, E., et al., *Inferring Pathway Activity toward Precise Disease Classification.* PLoS Computational Biology, 2008. **4**(11).
135. Yu, J.X., et al., *Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer.* BMC Cancer, 2007. **7**: p. 182.
136. Smid, M., et al., *Subtypes of Breast Cancer Show Preferential Site of Relapse.* Cancer Research, 2008. **68**(9): p. 3108.
137. Saal, L.H., et al., *Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity.* Proceedings of the National Academy of Sciences, 2007. **104**(18): p. 7564.
138. Price, N.D., et al., *Systems biology and systems medicine*, in *Genomic and Personalized Medicine*, G.S. Ginsburg and H.F. Willard, Editors. 2009, Elservier: Amsterdam. p. 74-86.
139. Hood, L., et al., *Systems Biology and New Technologies Enable Predictive and Preventative Medicine.* Science, 2004. **306**(5696): p. 640-643.
140. Kitano, H., *Systems biology: a brief overview.* Science, 2002. **295**(5560): p. 1662-4.
141. Lage, K., et al., *A human phenome-interactome network of protein complexes implicated in genetic disorders.* Nature Biotechnology, 2007. **25**: p. 309-316.
142. Price, N.D. and I. Shmulevich, *Biochemical and statistical network models for systems biology.* Current Opinion in Biotechnology, 2007. **18**(4): p. 365-370.

143. Sachs, K., et al., *Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data.* Science, 2005. **308**(5721): p. 523-529.
144. Wang, R.S., et al., *Inferring transcriptional regulatory networks from high-throughput data.* Bioinformatics, 2007. **23**(22): p. 3056.
145. Bansal, M., G.D. Gatta, and D. di Bernardo, *Inference of gene regulatory networks and compound mode of action from time course gene expression profiles.* Bioinformatics, 2006. **22**(7): p. 815-822.
146. Bonneau, R., *Learning biological networks: from modules to dynamics.* Nature Chemical Biology, 2008. **4**(11): p. 658-664.
147. Miller-Jensen, K., et al., *Common effector processing mediates cell-specific responses to stimuli.* Nature, 2007. **448**(7153): p. 604.
148. Janes, K.A., et al., *A Systems Model of Signaling Identifies a Molecular Basis Set for Cytokine-Induced Apoptosis.* Science, 2005. **310**(5754): p. 1646-1653.
149. Gordus, A., et al., *Linear combinations of docking affinities explain quantitative differences in RTK signaling.* Molecular Systems Biology, 2009. **5**: p. 235.
150. Kumar, N., et al., *Multi-Pathway Model Enables Prediction of Kinase Inhibitor Cross-Talk Effects on Migration of Her2-Overexpressing Mammary Epithelial Cells.* Molecular Pharmacology, 2008.
151. Bonneau, R., et al., *The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo.* Genome Biology, 2006. **7**(5): p. R36.
152. Bonneau, R., et al., *A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell.* Cell, 2007. **131**(7): p. 1354-1365.
153. Bansal, M., et al., *How to infer gene networks from expression profiles.* Molecular Systems Biology, 2007. **3**: p. 78.
154. Schafer, J. and K. Strimmer, *An empirical Bayes approach to inferring large-scale gene association networks.* Bioinformatics, 2005. **21**(6): p. 754-764.
155. Basso, K., et al., *Reverse engineering of regulatory networks in human B cells.* Nature Genetics, 2005. **37**: p. 382-390.
156. Hartemink, A.J., *Reverse engineering gene regulatory networks.* Nature Biotechnology, 2005. **23**: p. 554-555.
157. Friedman, N., *Inferring cellular networks using probabilistic graphical models.* Science, 2004. **303**(5659): p. 799-805.
158. Shmulevich, I., et al., *Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks.* Bioinformatics, 2002. **18**(2): p. 261-274.
159. Hashimoto, R.F., et al., *Growing genetic regulatory networks from seed genes.* Bioinformatics, 2004. **20**(8): p. 1241-1247.
160. Pal, R., et al., *Intervention in context-sensitive probabilistic Boolean networks.* Bioinformatics, 2005. **21**(7): p. 1211-1218.
161. Soranzo, N., G. Bianconi, and C. Altafini, *Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data.* Bioinformatics, 2007. **23**(13): p. 1640.
162. Yu, J., et al., *Advances to Bayesian network inference for generating causal networks from observational biological data.* Bioinformatics, 2004. **20**(18): p. 3594-3603.
163. Camacho, D., et al., *Comparison of reverse-engineering methods using an in silico network.* Annals of the New York Academy of Sciences, 2007. **1115**: p. 73-89.
164. Margolin, A.A. and A. Califano, *Theory and limitations of genetic network inference from microarray data.* Annals of the New York Academy of Sciences, 2007. **1115**: p. 51-72.
165. Duarte, N.C., et al., *Global reconstruction of the human metabolic network based on genomic and bibliomic data.* Proceedings of the National Academy of Sciences, 2007. **104**(6): p. 1777.
166. Ma, H., et al., *The Edinburgh human metabolic network reconstruction and its functional analysis.* Molecular Systems Biology, 2007. **3**(135).
167. Hoffmann, A., et al., *The Ikappa B-NF-kappa B Signaling Module: Temporal Control and Selective Gene Activation.* Science, 2002. **298**(5596): p. 1241-1245.
168. Papin, J.A., et al., *Reconstruction of cellular signalling networks and analysis of their properties.* Nature Reviews Molecular Cell Biology, 2005. **6**: p. 99-111.
169. Papin, J.A. and B.O. Palsson, *Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk.* Journal of Theoretical Biology, 2004. **227**(2): p. 283-297.

170. Gianchandani, E.P., et al., *Matrix formalism to describe functional states of transcriptional regulatory systems.* PLoS Computational Biology, 2006. **2**(8): p. e101.
171. Covert, M.W., et al., *Integrating high-throughput and computational data elucidates bacterial networks.* Nature, 2004. **429**(6987): p. 92-96.
172. Vo, T.D., H.J. Greenberg, and B.O. Palsson, *Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data.* Journal of Biological Chemistry, 2004. **279**(38): p. 39532-39540.
173. Thiele, I., et al., *Candidate Metabolic Network States in Human Mitochondria: IMPACT OF DIABETES, ISCHEMIA, AND DIET.* Journal of Biological Chemistry, 2005. **280**(12): p. 11683-11695.
174. Price, N.D., J. Schellenberger, and B.O. Palsson, *Uniform Sampling of Steady-State Flux Spaces: Means to Design Experiments and to Interpret Enzymopathies.* Biophysical Journal, 2004. **87**(4): p. 2172-2186.
175. Papin, J.A. and B.O. Palsson, *The JAK-STAT Signaling Network in the Human B-Cell: An Extreme Signaling Pathway Analysis.* Biophysical Journal, 2004. **87**(1): p. 37-46.
176. Lee, D.S., et al., *The implications of human metabolic network topology for disease comorbidity.* Proceedings of the National Academy of Sciences, 2008.
177. Ma, H. and I. Goryanin, *Human metabolic network reconstruction and its impact on drug discovery and development.* Drug Discovery Today, 2008. **13**(9-10): p. 402-408.
178. Shlomi, T., et al., *Network-based prediction of human tissue-specific metabolism.* Nature Biotechnology, 2008. **26**(9): p. 1003-1010.
179. Robey, I.F., et al., *Hypoxia-Inducible Factor-1a and the Glycolytic Phenotype in Tumors.* Neoplasia, 2005. **7**(4): p. 324.
180. Esteban, M.A. and P.H. Maxwell, *HIF, a missing link between metabolism and cancer.* Nature Medicine, 2005. **11**(10): p. 1047-8.
181. Galmarini, C.M., F. Popowycz, and B. Joseph, *Cytotoxic Nucleoside Analogues: Different Strategies to Improve their Clinical Efficacy.* Current Medicinal Chemistry, 2008. **15**(11): p. 1072-1082.
182. Serkova, N.J., J.L. Spratlin, and S.G. Eckhardt, *NMR-based metabolomics: Translational application and treatment of cancer.* Current Opinion in Molecular Therapeutics, 2007. **9**(6): p. 572-585.
183. Robak, T., et al., *Purine Nucleoside Analogs as Immunosuppressive and Antineoplastic Agents: Mechanism of Action and Clinical Activity.* Current Medicinal Chemistry, 2006. **13**(26): p. 3165-3189.
184. Werner, S.L., D. Barken, and A. Hoffmann, *Stimulus Specificity of Gene Expression Programs Determined by Temporal Control of IKK Activity.* Science, 2005. **309**(5742): p. 1857-1861.
185. Segre, D., et al., *From Annotated Genomes to Metabolic Flux Models and Kinetic Parameter Fitting.* OMICS: A Journal of Integrative Biology, 2003. **7**(3): p. 301-316.
186. Aldridge, B.B., et al., *Physicochemical modelling of cell signalling pathways.* Nature Cell Biology, 2006. **8**(11): p. 1195-203.
187. Kholodenko, B.N., *Cell-signalling dynamics in time and space.* Nature Reviews Molecular Cell Biology, 2006. **7**(3): p. 165-76.
188. Covert, M.W., et al., *Achieving Stability of Lipopolysaccharide-Induced NF-kappaB Activation.* Science, 2005. **309**(5742): p. 1854-1857.
189. Geva-Zatorsky, N., et al., *Oscillations and variability in the p53 system.* Molecular Systems Biology, 2006. **2**: p. 2006 0033.
190. Ma, L., et al., *A plausible model for the digital response of p53 to DNA damage.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(40): p. 14266-71.
191. Hennessy, B.T., et al., *Exploiting the PI3K/AKT pathway for cancer drug discovery.* Nature Reviews. Drug Discovery, 2005. **4**(12): p. 988-1004.
192. Asthagiri, A.R. and D.A. Lauffenburger, *A computational study of feedback effects on signal dynamics in a mitogen-activated protein kinase (MAPK) pathway model.* Biotechnology Progress, 2001. **17**(2): p. 227-39.
193. Brightman, F.A. and D.A. Fell, *Differential feedback regulation of the MAPK cascade underlies the quantitative differences in EGF and NGF signalling in PC12 cells.* FEBS Letters, 2000. **482**(3): p. 169-74.
194. Schoeberl, B., et al., *Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors.* Nature Biotechnology, 2002. **20**(4): p. 370-5.
195. Christopher, R., et al., *Data-Driven Computer Simulation of Human Cancer Cell.* Annals of the New York Academy of Sciences, 2004. **1020**(1): p. 132-153.
196. Li, H., X. Fan, and J. Houghton, *Tumor microenvironment: the role of the tumor stroma in cancer.* Journal of Cellular Biochemistry, 2007. **101**(4): p. 805-15.

197. Bellomo, N. and L. Preziosi, *Modelling and Mathematical Problems Related to Tumor Evolution and Its Interaction with the Immune System.* Mathematical and Computer Modelling, 2000. **32**(3): p. 413-452.
198. Araujo, R.P. and D.L. McElwain, *A history of the study of solid tumour growth: the contribution of mathematical modelling.* Bulletin of Mathematical Biology, 2004. **66**(5): p. 1039-1091.
199. Araujo, R.P. and D.L.S. McElwain, *A history of the study of solid tumour growth: The contribution of mathematical modelling.* Bulletin of Mathematical Biology, 2004. **66**(5): p. 1039-1091.
200. Quaranta, V., et al., *Invasion emerges from cancer cell adaptation to competitive microenvironments: Quantitative predictions from multiscale mathematical models.* Seminars in Cancer Biology, 2008.
201. Alarcón, T., H.M. Byrne, and P.K. Maini, *A mathematical model of the effects of hypoxia on the cell-cycle of normal and cancer cells.* Journal of Theoretical Biology, 2004. **229**(3): p. 395-411.
202. Ambrosi, D. and F. Mollica, *The role of stress in the growth of a multicell spheroid.* Journal of Mathematical Biology, 2004. **48**(5): p. 477-499.
203. Anderson, A.R.A., *A hybrid mathematical model of solid tumour invasion: the importance of cell adhesion.* Mathematical Medicine and Biology, 2005. **22**(2): p. 163-186.
204. Johnston, M.D., et al., *Mathematical modeling of cell population dynamics in the colonic crypt and in colorectal cancer.* Proceedings of the National Academy of Sciences, 2007. **104**(10): p. 4008.
205. d'Onofrio, A. and I.P.M. Tomlinson, *A nonlinear mathematical model of cell turnover, differentiation and tumorigenesis in the intestinal crypt.* Journal of Theoretical Biology, 2007. **244**(3): p. 367-374.
206. Byrne, H. and L. Preziosi, *Modelling solid tumour growth using the theory of mixtures.* Mathematical Medicine and Biology, 2003. **20**(4): p. 341-366.
207. Breward, C.J.W., H.M. Byrne, and C.E. Lewis, *The role of cell-cell interactions in a two-phase model for avascular tumour growth.* Journal of Mathematical Biology, 2002. **45**(2): p. 125-152.
208. Macklin, P. and J. Lowengrub, *Nonlinear simulation of the effect of microenvironment on tumor growth.* Journal of Theoretical Biology, 2007. **245**(4): p. 677-704.
209. Alarcon, T., H.M. Byrne, and P.K. Maini, *A multiple scale model for tumor growth.* Multiscale Modeling & Simulation, 2005. **3**(2): p. 440–475.
210. Spencer, S.L., et al., *An ordinary differential equation model for the multistep transformation to cancer.* Journal of Theoretical Biology, 2004. **231**(4): p. 515-524.
211. Baum, M., et al., *Does breast cancer exist in a state of chaos?* European Journal of Cancer, 1999. **35**(6): p. 886-891.
212. Rew, D.A., *Tumour biology, chaos and non-linear dynamics.* European Journal of Surgical Oncology, 1999. **25**(1): p. 86-89.
213. Deisboeck, T.S., J. Yoon, and J. Costa, *In silico cancer modeling: is it ready for prime time?* Nature Clinical Practice Oncology, 2008. **6**(1): p. 34-42.
214. Mansury, Y., et al., *Emerging patterns in tumor systems: simulating the dynamics of multicellular clusters with an agent-based spatial agglomeration model.* Journal of Theoretical Biology, 2002. **219**(3): p. 343-70.
215. Little, M.P., P. Vineis, and G. Li, *A stochastic carcinogenesis model incorporating multiple types of genomic instability fitted to colon cancer data.* Journal of Theoretical Biology, 2008.
216. Dingli, D., A. Traulsen, and J.M. Pacheco, *Stochastic dynamics of hematopoietic tumor stem cells.* Cell Cycle, 2007. **6**(4): p. 461-6.
217. Ostby, I., L. Oyehaug, and H.B. Steen, *A stochastic model of cancer initiation including a bystander effect.* Journal of Theoretical Biology, 2006. **241**(4): p. 751-64.
218. Kansal, A.R., et al., *Simulated brain tumor growth dynamics using a three-dimensional cellular automaton.* Journal of Theoretical Biology, 2000. **203**(4): p. 367-82.
219. Mallet, D.G. and L.G. De Pillis, *A cellular automata model of tumor–immune system interactions.* Journal of Theoretical Biology, 2006. **239**(3): p. 334-350.
220. Alarcón, T., H.M. Byrne, and P.K. Maini, *A cellular automaton model for tumour growth in inhomogeneous environment.* Journal of Theoretical Biology, 2003. **225**(2): p. 257-274.
221. Bauer, A.L., T.L. Jackson, and Y. Jiang, *A Cell-Based Model Exhibiting Branching and Anastomosis during Tumor-Induced Angiogenesis.* Biophysical Journal, 2007. **92**(9): p. 3105.
222. Patel, A.A., et al., *A Cellular Automaton Model of Early Tumor Growth and Invasion: The Effects of Native Tissue Vascularity and Increased Anaerobic Tumor Metabolism.* Journal of Theoretical Biology, 2001. **213**(3): p. 315-331.
223. Mansury, Y. and T.S. Deisboeck, *The impact of "search precision" in an agent-based tumor model.* Journal of Theoretical Biology, 2003. **224**(3): p. 325-337.

224. McDougall, S.R., A.R.A. Anderson, and M.A.J. Chaplain, *Mathematical modelling of dynamic adaptive tumour-induced angiogenesis: Clinical implications and therapeutic targeting strategies.* Journal of Theoretical Biology, 2006. **241**(3): p. 564-589.
225. Anderson, A.R., *A hybrid mathematical model of solid tumour invasion: the importance of cell adhesion.* Mathematical Medicine and Biology, 2005. **22**(2): p. 163-86.
226. Zhang, L., C.A. Athale, and T.S. Deisboeck, *Development of a three-dimensional multiscale agent-based tumor model: simulating gene-protein interaction profiles, cell phenotypes and multicellular patterns in brain cancer.* Journal of Theoretical Biology, 2007. **244**(1): p. 96-107.
227. Anderson, A.R., et al., *Tumor morphology and phenotypic evolution driven by selective pressure from the microenvironment.* Cell, 2006. **127**(5): p. 905.
228. Ohno, R. and Y. Nakamura. *Prediction of response to imatinib by cDNA microarray analysis*. in *Seminars in Hematology*. 2003: Elsevier.
229. Maglott, D., et al., *Entrez Gene: gene-centered information at NCBI.* Nucleic Acids Research, 2006.
230. Bicciato, S., et al., *Pattern identification and classification in gene expression data using an autoassociative neural network model.* Biotechnology and Bioengineering, 2003. **81**(5): p. 594-606.
231. Bloom, G., et al., *Multi-platform, multi-site, microarray-based human tumor classification.* American Journal of Pathology, 2004. **164**(1): p. 9-16.
232. Khan, J., et al., *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.* Nature Medicine, 2001. **7**(6): p. 673-679.
233. Boulesteix, A.L., G. Tutz, and K. Strimmer, *A CART-based approach to discover emerging patterns in microarray data.* Bioinformatics, 2003. **19**(18): p. 2465-2472.
234. Dettling, M. and P. Buhlmann, *Boosting for tumor classification with gene expression data.* Bioinformatics, 2003. **19**(9): p. 1061-1069.
235. Zhang, H., C.Y. Yu, and B. Singer, *Cell and tumor classification using gene expression data: construction of forests.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(7): p. 4168-4172.
236. Qu, Y., et al., *Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients.* Clinical Chemistry, 2002. **48**(10): p. 1835-1843.
237. Peng, S., et al., *Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines.* FEBS Letters, 2003. **555**(2): p. 358-362.
238. Yeang, C.H., et al., *Molecular classification of multiple tumor types.* Bioinformatics, 2001. **17** p. S316-22.
239. Subramanian, A., et al., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(43): p. 15545-15550.
240. Chuang, H.Y., et al., *Network-based classification of breast cancer metastasis.* Molecular Systems Biology, 2007. **3**: p. 140.
241. Lee, E., et al., *Inferring Pathway Activity toward Precise Disease Classification.* PLoS Computational Biology, 2008. **4**(11): p. e1000217.
242. Auffray, C., *Protein subnetwork markers improve prediction of cancer outcome.* Molecular Systems Biology, 2007. **3**: p. 141.
243. Geman, D., et al., *Classifying gene expression profiles from pairwise mRNA comparisons.* Statistical Applications in Genetics and Molecular Biology, 2004. **3**(1): p. Article 19.
244. Lin, X., et al., *The ordering of expression among a few genes can provide simple cancer biomarkers and signal BRCA1 mutations.* BMC Bioinformatics, 2009. **10**: p. 256.
245. Leek, J.T., *The tspair package for finding top scoring pair classifiers in R.* Bioinformatics, 2009. **25**(9): p. 1203-1204.
246. Earls, J.C., et al., *AUREA: an open-source software system for accurate and user-friendly identification of relative expression molecular signatures.* BMC Bioinformatics, 2013. **14**: p. 78.
247. Xu, L., et al., *Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data.* Bioinformatics, 2005. **21**(20): p. 3905-3911.
248. Braga-Neto, U.M. and E.R. Dougherty, *Is cross-validation valid for small-sample microarray classification?* Bioinformatics, 2004. **20**(3): p. 374-380.
249. Gordon, G.J., et al., *Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma.* Cancer Research, 2002. **62**(17): p. 4963-4967.

250. Ma, X.J., et al., *A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen.* Cancer Cell, 2004. **5**(6): p. 607-616.
251. Clarke, R., et al., *Antiestrogen resistance in breast cancer and the role of estrogen receptor signaling.* Oncogene, 2003. **22**(47): p. 7316-7339.
252. Jordan, C., *Historical perspective on hormonal therapy of advanced breast cancer.* Clinical Therapeutics, 2002. **24**: p. 3-16.
253. Clarke, M.J., *Tamoxifen for early breast cancer.* Cochrane Database of Systematic Reviews (Online), 2008(4): p. CD000486.
254. Nicholson, R.I., et al., *The biology of antihormone failure in breast cancer.* Breast Cancer Research and Treatment, 2003. **80**: p. 29-34.
255. Dudoit, S., et al., *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.* Statistica Sinica, 2002. **12**(1): p. 111-140.
256. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-537.
257. Singh, D., et al., *Gene expression correlates of clinical prostate cancer behavior.* Cancer Cell, 2002. **1**(2): p. 203-209.
258. Price, N.D., et al., *Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(9): p. 3414-3419.
259. Xu, L., et al., *Merging microarray data from separate breast cancer studies provides a robust prognostic test.* BMC Bioinformatics, 2008. **9**: p. 125.
260. Rajeshkumar, N.V., et al., *Antitumor effects and biomarkers of activity of AZD0530, a Src inhibitor, in pancreatic cancer.* Clinical Cancer Research, 2009. **15**(12): p. 4138-4146.
261. Raponi, M., et al., *A 2-gene classifier for predicting response to the farnesyltransferase inhibitor tipifarnib in acute myeloid leukemia.* Blood, 2008. **111**(5): p. 2589-2596.
262. Weichselbaum, R.R., et al., *An interferon-related gene signature for DNA damage resistance is a predictive marker for chemotherapy and radiation for breast cancer.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(47): p. 18490-18495.
263. Edelman, L.B., et al., *Two-transcript gene expression classifiers in the diagnosis and prognosis of human diseases.* BMC Genomics, 2009. **10**: p. 583.
264. Michiels, S., S. Koscielny, and C. Hill, *Prediction of cancer outcome with microarrays: a multiple random validation strategy.* The Lancet, 2005. **365**(9458): p. 488-492.
265. Nam, D. and S.Y. Kim, *Gene-set approach for expression pattern analysis.* Briefings in Bioinformatics, 2008. **9**(3): p. 189-97.
266. Shaw, R.J., *Glucose metabolism and cancer.* Current Opinion in Cell Biology, 2006. **18**(6): p. 598-608.
267. Mellinghoff, I.K., et al., *Molecular Determinants of the Response of Glioblastomas to EGFR Kinase Inhibitors.* New England Journal of Medicine, 2005. **353**(19): p. 2012-2024.
268. Shaw, R.J. and L.C. Cantley, *Ras, PI(3)K and mTOR signalling controls tumour cell growth.* Nature, 2006. **441**(7092): p. 424-430.
269. Karin, M., *Nuclear factor-[kappa]B in cancer development and progression.* Nature, 2006. **441**(7092): p. 431-436.
270. Parsons, D.W., et al., *An Integrated Genomic Analysis of Human Glioblastoma Multiforme.* Science, 2008. **321**(5897): p. 1807-1812.
271. McLendon, R., et al., *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-1068.
272. Jones, S., et al., *Core signaling pathways in human pancreatic cancers revealed by global genomic analyses.* Science, 2008. **321**(5897): p. 1801-6.
273. Price, N.D., et al., *Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(9): p. 3414.
274. Geman, D., et al., *Classifying gene expression profiles from pairwise mRNA comparisons.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article19.
275. Yu, Y.P., et al., *Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy.* Journal of Clinical Oncology, 2004. **22**(14): p. 2790-9.
276. Trapani, J.A. and M.J. Smyth, *Functional significance of the perforin/granzyme cell death pathway.* Nature Reviews Immunology, 2002. **2**(10): p. 735-47.

277. Tagnon, H.J., W.F. Whitmore, Jr., and N.R. Shulman, *Fibrinolysis in metastatic cancer of the prostate.* Cancer, 1952. **5**(1): p. 9-12.
278. Chandran, U.R., et al., *Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process.* BMC Cancer, 2007. **7**: p. 64.
279. Shimada, K., et al., *Molecular roles of MAP kinases and FADD phosphorylation in prostate cancer.* Histology and Histopathology, 2006. **21**(4): p. 415-22.
280. Vapnik, V.N., *The nature of statistical learning theory*. 2000, New York, NY: Springer Verlag.
281. Joachims, T., *Making Large-Scale SVM Learning Practical*, in *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf, C.J.C. Burges, and A.J. Smola, Editors. 1998, MIT Press: Cambridge, MA.
282. Moreno, C.S., et al., *Evidence that p53-mediated cell-cycle-arrest inhibits chemotherapeutic treatment of ovarian carcinomas.* PLoS One, 2007. **2**(5): p. e441.
283. Yao, Z., et al., *A Marfan syndrome gene expression phenotype in cultured skin fibroblasts.* BMC Genomics, 2007. **8**: p. 319.
284. Kuriakose, M.A., et al., *Selection and validation of differentially expressed genes in head and neck cancer.* Cellular and Molecular Life Sciences, 2004. **61**(11): p. 1372-83.
285. Ryan, M.M., et al., *Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes.* Molecular Psychiatry, 2006. **11**(10): p. 965-78.
286. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-7.
287. Armstrong, S.A., et al., *MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia.* Nature Genetics, 2002. **30**(1): p. 41-7.
288. Chandrasekaran, S., et al., *Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(44): p. 18020-5.
289. Yang, J., et al., *Integrated proteomics and genomics analysis reveals a novel mesenchymal to epithelial reverting transition in leiomyosarcoma through regulation of slug.* Molecular and Cellular Proteomics, 2010. **9**(11): p. 2405-13.
290. Robinson, G.E., *Genomics. Beyond nature and nurture.* Science, 2004. **304**(5669): p. 397-9.
291. Margolin, A.A., et al., *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* BMC Bioinformatics, 2006. **7 Suppl 1**: p. S7.
292. Efron, B., *Least angle regression*. 2002, Stanford, Calif.: Stanford University, Department of Biostatistics. 41 p.
293. Eddy, J.A., et al., *Identifying tightly regulated and variably expressed networks by Differential Rank Conservation (DIRAC).* PLoS Comput Biol, 2010. **6**(5): p. e1000792.
294. Bauer, R., et al., *Schlank, a member of the ceramide synthase family controls growth and body fat in Drosophila.* EMBO J, 2009. **28**(23): p. 3706-16.
295. Cahoy, J.D., et al., *A transcriptome database for astrocytes, neurons, and oligodendrocytes: A new resource for understanding brain development and function.* Journal of Neuroscience, 2008. **28**(1): p. 264-278.
296. Ahn, Y.Y., J.P. Bagrow, and S. Lehmann, *Link communities reveal multiscale complexity in networks.* Nature, 2010. **466**(7307): p. 761-4.
297. Pudil, P., J. Novovicová, and J. Kittler, *Floating search methods in feature selection.* Pattern recognition letters, 1994. **15**(11): p. 1119-1125.
298. Shaw, R.J. and L.C. Cantley, *Ras, PI(3)K and mTOR signalling controls tumour cell growth.* Nature, 2006. **441**(7092): p. 424-30.
299. TCGA, *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-8.
300. Eddy, J.A., et al., *Relative expression analysis for molecular cancer diagnosis and prognosis.* Technology in Cancer Research and Treatment, 2010. **9**(2): p. 149-59.
301. Edelman, L.B., J.A. Eddy, and N.D. Price, *In silico models of cancer.* Wiley Interdisciplinary Reviews. Systems Biology and Medicine, 2010. **2**(4): p. 438-59.
302. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
303. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
304. Griffin, J.L. and R.A. Kauppinen, *A metabolomics perspective of human brain tumours.* FEBS J, 2007. **274**(5): p. 1132-9.

305. Thiele, I. and B.O. Palsson, *A protocol for generating a high-quality genome-scale metabolic reconstruction.* Nat Protoc, 2010. **5**(1): p. 93-121.
306. Jerby, L., T. Shlomi, and E. Ruppin, *Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism.* Mol Syst Biol, 2010. **6**: p. 401.
307. Folger, O., et al., *Predicting selective drug targets in cancer through metabolic networks.* Mol Syst Biol, 2011. **7**: p. 501.
308. Frezza, C., et al., *Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase.* Nature, 2011. **477**(7363): p. 225-8.
309. Gudmundsson, S. and I. Thiele, *Computationally efficient flux variability analysis.* BMC Bioinformatics, 2010. **11**: p. 489.
310. Wang, Y., J.A. Eddy, and N.D. Price, *Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE.* BMC Syst Biol, 2012. **6**: p. 153.