Identifying and Modeling Perturbed Networks in Cancer through Statistical and Constraint-based Analysis

by

James Allan Eddy

Thesis Proposal for Preliminary Examination

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, 2011

Committee:

Nathan D. Price, Assistant Professor (chair)
Gene Robinson, Professor
Michael Insana, Professor
Sheng Zhong, Assistant Professor
Jian Ma, Assistant Professor

Table of Contents

A. Overview and Specific Aims

Perturbations to biomolecular networks in glioblastoma multiforme (GBM)—the most common and devastating form of brain cancer—eventually result in the symptoms of disease observed by the patient. Malignant phenotypes in GBM 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. Reconstructing detailed *in silico* models of biochemical reaction networks (e.g., metabolic, signaling, regulatory) at the genome scale establishes a platform on which genetic perturbations can be related to emergent malignant functions and phenotypes. I am reconstructing the first genome-scale network of GBM metabolism, as altered function in several metabolic pathways has been shown to be critical in the development of this and other cancers. Constraint-based modeling with the network enables the simulation of cell growth under physiologically relevant conditions and provides a means to investigate mechanisms for tumor development. Curated pathways in the reconstructed network also provide functional context for studying changes in gene expression between disease states—a promising application for methods I am developing to assess changes in networks within and between phenotypes. These combined approaches enable the identification and detailed characterization of key perturbed pathways in GBM. The specific aims of my proposed work as well as my expected timeline (**Figure 1**) are outlined below.

Specific Aim 1. Develop tools for network-based analysis of high throughput expression data in cancer and other diseases: I am working to develop a suite of tools that assess cellular regulation of networks in the context of *relative expression*, either among genes within a network, or between multiple networks. In addition to an approach I previously developed to investigate differential patterns of network rankings within and between phenotypes, I will complete an interaction-network based extension to the method that will be advantageous for studying mechanistic networks such as metabolism or signaling. I will also build off of this approach to complete a method for comparative analysis of related networks in "network families." These methods will be developed and tested using data for a number of cancer types and other diseases.

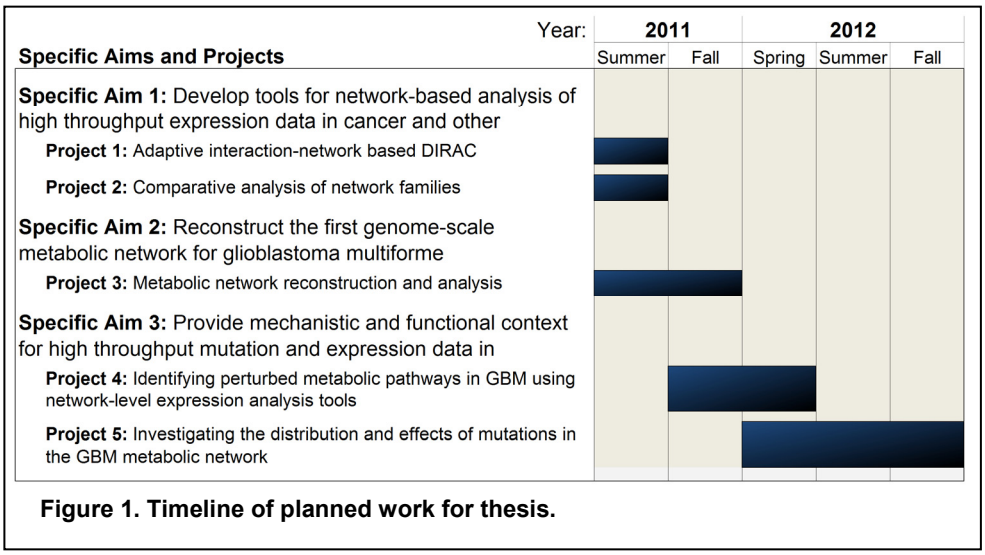

Figure 1. Timeline of planned work for thesis.

Specific Aim 2. Reconstruct the first genome-scale metabolic network for glioblastoma multiforme: To reconstruct a metabolic network specific to GBM, I am using the previously reconstructed genome-scale human metabolic network combined with gene and protein expression data to predict the subset of metabolism active in the cancer. Metabolomic profiling of GBM cell lines will complement the model-building process, providing information for gap filling and validation. I will use constraint-based modeling to interrogate network properties and simulate cell growth under physiologically relevant conditions. I will take advantage of ongoing experimental studies with the U87 GBM cell line in the Price lab to assess predictive accuracy and validate the model.

Specific Aim 3. Provide mechanistic and functional context for high throughput mutation and expression data in glioblastoma using the metabolic model: After evaluating the predictive accuracy of the GBM metabolic network from **Specific Aim 2** by comparing *in silico* simulations with experimentally observed behavior, I will use the model as a platform for systems-level analysis of high throughput data related to the cancer. I will first employ the tools developed in **Specific Aim 1** to study patterns of gene expression within the high-confidence pathways of the metabolic network, focusing on comparisons between GBM and healthy astrocytes. I will also use the metabolic network to contextualize the abundance of mutational data available for GBM, inspecting the distribution and frequency of mutations within metabolic pathways, and modeling the effect of specific mutations that have been described for the cancer.

B. Background and Significance

Glioblastoma multiforme: Glioblastoma multiforme (GBM) is the most common and aggressive type of primary brain tumor, arising from astrocyte cells in the central nervous system (CNS). A recent study of patients with newly-diagnosed GBM showed survival rates of 42% at 6 months, 18% at 1 year, and only 3.3% at 2 years, despite access to state-of-the-art surgery, imaging, radiotherapy, and chemotherapy [3]. Primary CNS tumors such as GBM are diagnosed in over 22,000 patients annually in the United States, representing 1.4% of all diagnosed cancers and 2.3% of all cancer related deaths; as of 2000, there were an estimated 128,000 people living with a diagnosis of a primary CNS tumor in this country [4]. Gliomas account for 80% of adult primary brain tumors [5], and arise from glial cells in the brain—i.e., oligodendrocytes, astrocytes, ependymal cells, Schwann cells, microglia, and satellite cells. The three main types of gliomas—determined by the cell of origin—are astrocytomas, ependymomas, and oligodendrogliomas. Astrocytomas are the most common type of glioma, and can be classified into four grades: *Grade 1*, pilocytic astrocytoma; *Grade 2*, low-grade astrocytoma; *Grade 3*, anaplastic astrocytoma; and *Grade 4*, glioblastoma multiforme (GBM), which accounts for just over half of these gliomas. Notably, there is evidence that the incidence of malignant gliomas is increasing, particularly in the elderly [6, 7]. While clinical prognosis for those with high grade gliomas is bleak, there are early signs and cautious optimism that the field may be poised for significant advancement [8]. Importantly, perturbations to *biomolecular 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.

Systems approaches to cancer: The systems biology approach to study, diagnose, and treat human cancer has become a critical and necessary method to combat the disease's complex nature. High-throughput experimental technologies have enabled the identification of biological components at unprecedented scale, from cells and tissues to genes and proteins. Collectively, these technologies provide a "parts list" for biological systems (e.g., biochemical pathways, larger interaction networks). Systems biology employs an integrative approach to characterize biological systems, in which interactions among all these components in a system are described mathematically to establish a computable model that integrates this data into a cohesive whole. These *in silico* models—which complement traditional *in vivo* and *in vitro* 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.

Molecular and physiological complexity in cancer: Cancer is particularly amenable to systems biology approaches because it is an intrinsically complex and heterogeneous disease [9]. 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 [10]. At the molecular level, genetic lesions interact synergistically to evade tumor suppression pathways, with no single mutation typically sufficient to cause transformation [11-15]. The convolution of genetic effects, changes in gene and protein expression levels [16-18], and epigenetic modifications [19-24] further illustrates the complex, nonlinear relationship between molecular state and cellular cancer phenotype, emphasizing the need for heterogeneous data integration through *in silico* models. Today, important efforts in sequencing the human genome [25, 26] and now individual cancers [27, 28] mean that malignant genetic transformations can be studied in the context of the entire genome. As a result, it is becoming increasingly clear that GBM and other cancers result not only from multiple perturbations, but from differing sets of mutations in every patient [28]. 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 in oncogenesis and in the brain: 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 [29-31]. Thus, metabolism provides an ideal setting to begin reconstructing detailed biochemical networks at the genome-scale for GBM. Notably, while the brain accounts for only 2% of the body by weight, it consumes 20% of the oxygen and 25% of the glucose in the body, making it the most metabolically active

organ. As glial cells make up 90% of the brain, their metabolism represents a dominant aspect of human metabolic activity; astrocytes are the most abundant type of glial cell. Metabolite-based analyses to probe cancer have been performed since the 1980s [32] 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 [33]. Metabolic phenotypes that remain consistent across cancer types, including GBM, reflect decreased aerobic respiration activity [34, 35] coupled with increased glycolysis [33, 34] and increased phospholipid production [34, 36]. These observations have led to targeted diagnosis and treatment strategies [33, 37]. In addition to broad spectrum cancer markers, metabolite signatures specific to astrocytomas have been found, including elevated levels of *N*acetylaspartate and myo-inositol [38], which can be used to distinguish astrocytomas from other brain cancers. Based on past advances resulting from metabolite analysis in cancer cells, accurate profiling of the GBM cell metabolome—and integrating this data into a cell-scale metabolic model—is of high interest. Doing so will not only further elucidate the pathophysiolgy of GBM, but has potential to improve methods of diagnosis and treatment.

Stoichiometric models of biochemical reaction 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. 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. 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.

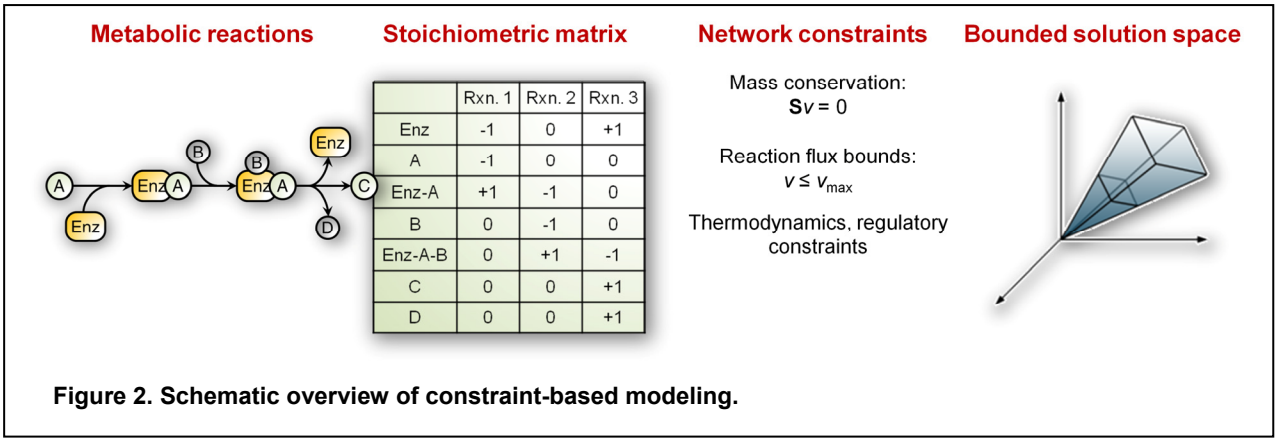

Figure 2. Schematic overview of constraint-based modeling.

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 [39-41], and genome-scale models of metabolism have been completed for a diverse range of organisms, including the prokaryote *E. coli* [42], archaea [43], and eukaryotes such as *S. cerevisiae* [44] and the protozoan *L. major* [45]. 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 2**) are then used to interrogate the properties of the reconstructed network *in silico*, and to facilitate model-driven validation and refinement [46]. 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 [47, 48]. 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 [49, 50].

C. Preliminary Studies and Results

Network-level expression analysis: Identifying tightly regulated and variably expressed networks by Differential Rank Conservation (DIRAC)

Methods for analyzing high-throughput microarray data are shifting towards an increased focus on biologically meaningful pathways or networks—we consider all pathways to in fact be part of interconnected biological networks, and henceforth use the term network rather than pathway. Cellular regulation of a network can be generally thought of as the manner in which gene expression is controlled in response to specific conditions. Existing network-based expression analysis tools commonly investigate informative patterns of up- or down-regulation of grouped genes in different disease states. For example, the widely-used gene set enrichment analysis (GSEA) platform identifies networks that are significantly enriched for over- or under-expressed genes [51, 52]. Other methods employ a single statistic to represent the collective activity of a network (e.g., mean or median gene expression) [17, 53]. 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 [16, 17].

I have developed a new method called Differential Rank Conservation (DIRAC) [54] which considers the "combinatorial comparisons" of network genes as opposed to increases or decreases alone, and provides quantitative measures of how network expression differs within and between phenotypes (**Figure 3**). 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. 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. 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). Extensions and improvements to DIRAC will be the subject of my work to complete **Specific Aim 1**.

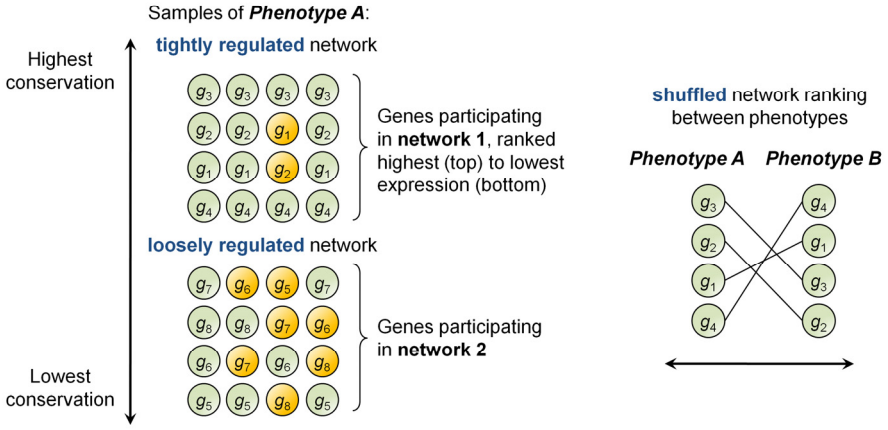

Figure 3. Prototypical scenarios observed for networks in DIRAC.

Population-level DIRAC: 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 two scenarios (**Figure 3A**). In the first, networks are ranked similarly in all samples (high rank conservation); in the second case, the ordering of network genes is highly varied (low conservation). I examined several disease phenotypes including cancer subtypes and neurological disorders and identified networks (defined according to the BioCarta database for human signaling pathways) that are tightly regulated, as defined by high conservation of transcript ordering. Whereas studying up- or down-regulation of a network provides a specific measure of how a process changes, identifying networks that are tightly or loosely regulated indicates the level of control across samples in a population. Since 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.

Certain networks may be tightly regulated in one phenotype, but not in another; such cases represent the *deregulation* of a network in one phenotype relative to another. I used the absolute difference in rank conservation between phenotypes to identify the most deregulated networks for different binary phenotype comparisons.

Interestingly, I found that network rankings tend to be more varied (less tightly regulated) in the more pathological state (**Figure 4**). Additionally, averaging rank conservation over all networks provided a measure of global regulation in different phenotypes. The trend of lower conservation in more malignant phenotypes seems to persist even from a coarser, global perspective. This global pattern of increased disorder with malignancy highlights the utility of studying gene ordering within networks, and also reveals a striking phenomenon that will drive future investigation and may lead to new understandings of gene expression in cancer and other diseases.

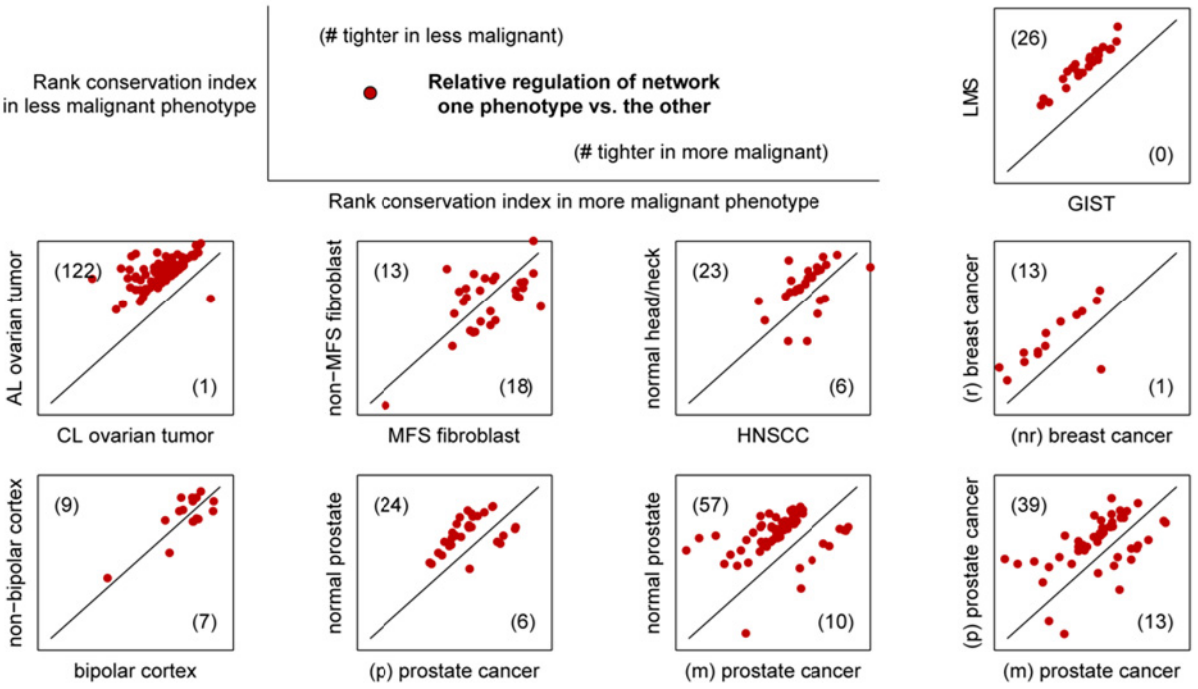

Figure 4. Deregulation of networks in disease. Rank conservation index values in the less malignant phenotype (y-axis) are plotted against indices in the more malignant phenotype (x-axis). Higher rank conservation in the less or more malignant phenotypes is indicated by points above or below the diagonal line, respectively.

Sample-level DIRAC: DIRAC can also be applied at the sample level to identify conservation differences (i.e., shuffling) between phenotypes for a specified network (**Figure 3B**). 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.

Previous efforts in modeling human metabolism and reconstructing cell- and tissue-specific models

While the generation of stoichiometric models and analysis of their properties is a well-established process in microbes, these methods are now being adapted for modeling human systems in greater detail. The models and methods described below, while previously developed by other groups, will be critical for the completion of **Specific Aim 2**. Additionally, I have worked directly on the reconstruction of two genome-scale models—for the protozoan parasite *Leishmania major* [45] and the butanogenic bacteria *Clostridium beijerinckii* [Milne, Eddy, et al., *submitted*]—providing me with valuable expertise for building and analyzing metabolic networks in human.

Modeling human metabolism in human disease: 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 [55], linear programming and random sampling were applied to identify candidate steady states of the network under normal, diabetic, ischemic, and dietetic conditions [56]. In a related study, Monte Carlo sampling of flux spaces was used to study the effects of enzymopathies on the human erythrocyte metabolic network [57]. The recent completion of a global reconstruction of the human metabolic network [58] represents a significant milestone

in human systems biology; *Human Recon 1* comprises 1496 genes and 3748 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 [59]. An independently reconstructed human metabolic network [60] was also used to demonstrate the potential of systems modeling in human metabolism to aid in drug discovery [61].

Reconstructing cell- and tissue-specific metabolic models: 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 [2] (**Table 1**). 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.

Tissue	Active (of 1165)	Inactive	Undetermined
Liver	12.40%	19.20%	68.40%
Brain	7.00%	32.10%	60.90%
Heart	4.20%	22.00%	73.80%
Kidney	12.90%	21.50%	65.70%
Lung	3.90%	33.60%	62.50%
Pancreas	3.80%	35.50%	60.80%
Prostate	2.80%	30.00%	67.10%
Spleen	1.10%	38.50%	60.40%
Thymus	1.00%	37.30%	61.60%
Skeletal Muscle	4.60%	28.00%	67.40%

Table 1. Organ-specific metabolism (from [2]).

Recent efforts have focused on modeling the human metabolic network in the context of specific cell and tissue types. Shlomi et al. developed an MILP based approach to predict the active flux distribution within the global model (*Human Recon 1*) based on tissue-specific gene expression data [2]. More recently, two groups have reconstructed computable genome-scale metabolic networks specific to the liver [1, 62]. While the work by Gille et al. was a more intensive manual effort, effectively building a liver network from the ground up, the model built by Jerby et al. is based on a more automated and algorithmic approach. As such, the latter is more appealing as a tool for researchers interested in reconstructing tissue-specific metabolic networks. The model building algorithm (MBA) used employs tissue-specific reaction evidence (literature, -omics) to systematically prune down the generic (non-tissue specific) model and establish a reduced tissue-specific model for the liver (**Figure 5**). Specifically, the available liver-specific evidence is used to divide reactions in the generic model into three subsets: a high probability core set, a moderate probability core, and the remaining reactions. The MBA uses an optimization-based approach to maximize the number of core reactions included and minimize the number of extra reactions added to construct the most parsimonious and consistent model. I have already implemented a Matlab version of the MBA algorithm which I will use to reconstruct the metabolic network of GBM for **Specific Aim 2** (see **Research Design and Methods**). Using *Human Recon 1* as a starting point, with gene and protein evidence acquired automatically from online databases, I have used MBA to generate an initial draft of the GBM network; this draft will be the subject of further curation and refinement, prior to simulation and validation.

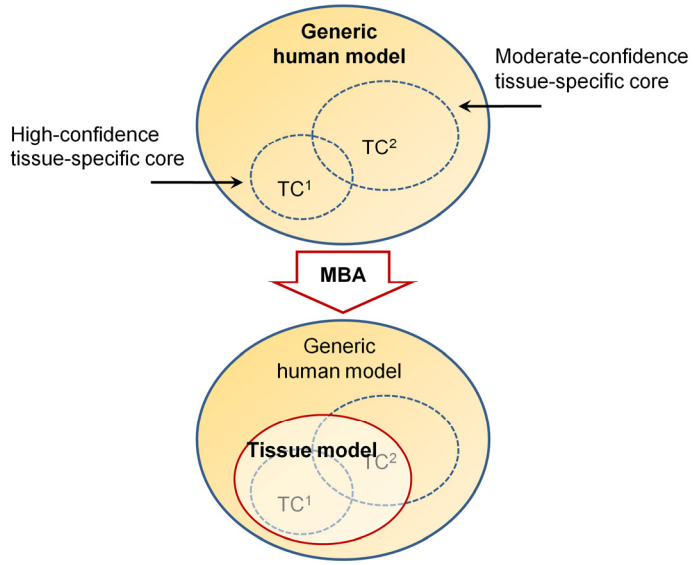

Figure 5. Model building algorithm (MBA; adapted from [1]).

D. Research Design and Methods

Specific Aim 1. Develop tools for network-based analysis of high throughput expression data in cancer and other diseases.

In addition to the initial version of DIRAC, I am working on an extension to the method that will add key functionality for dealing with mechanistic reaction networks such as the GBM metabolic reconstruction. At the same time, I am completing a related project that focuses on investigating informative patterns of expression between multiple related networks. Importantly, I am collaborating with another graduate student in my group, 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. 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 [Earls, Eddy, et al., *in preparation*]. 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.

Project 1. Adaptive interaction-network based DIRAC

I am currently working with an undergraduate student to develop an extension to DIRAC that offers two 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 [Hadidi, Eddy, et al., *in preparation*].

Defining subnetworks using link communities: Protein-protein interaction (PPI) information is 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 [63]. 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. We are using a sequential forward floating search algorithm [64] to grow and shrink subnetworks identified as variably expressed, in an effort to improve classification accuracy.

Project 2. Comparative analysis of network families

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 [28, 31, 65, 66]. 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 have developed an approach called Network Family Relative Expression Analysis (NF-RXA) to identify molecular signatures comprising multiple related networks (denoted as “families” based on Gene Ontology terminology). RXA refers to a class of algorithms that assess the ordering among gene expression values, rather than their absolute expression values (reviewed by Eddy et al. [67]); one searches for characteristic perturbations in this ordering from one phenotype to another, or within phenotypes. In particular, the NF-RXA adapts the top-scoring pair (TSP) and DIRAC algorithms to study network families in disease.

Defining network families using Gene Ontology: The Gene Ontology (GO) database is organized as a directed acyclic graph, with each GO term representing a vertex that is a part of one or more directed paths. In keeping with

GO structure and vocabulary (i.e., the designation of "descendants" and "ancestors"), NF-RXA defines all sink vertices (those with no outgoing edges) as "children", and vertices one height level up in the directed path from the sink vertices as the "parents." These children terms comprise the most specialized terms in the ontology, and are related to one or more parent terms via the *is_a* and *part_of* relationships defined in the GO database. Given these definitions, the term "family" thus becomes a natural label for the sets of related networks (GO terms) derived from GO. Prior to application of the algorithms described below, the expression level of each child network (i.e., activity level) is calculated as the mean expression of participating genes.

Top scoring pairs of networks – NF-TSP: The TSP algorithm (described here [67]) identifies gene pairs exhibiting a characteristic "relative expression reversal" between the phenotypes or classes of interest. Briefly, one selects the pair of genes g_i and g_j for which the difference $|\text{Prob}(X_i < X_j \mid \textit{class A}) - \text{Prob}(X_i < X_j \mid \textit{class B})|$ is maximized, where X denotes the expression profile of the corresponding gene g. Operating on network expression levels instead of genes, NF-TSP selects the pair of networks m_i and m_j that satisfies the same rule. Network expression reversals between different phenotypes are exclusively assessed within network families. By restricting pair-wise comparisons to children of the same parent GO term, I expect a stronger and easier to interpret functional relationship between networks in the classifier pair, while still maintaining high predictive accuracy.

Differential rank conservation of network families – NF-DIRAC: As with NF-TSP, NF-DIRAC uses the previously published algorithm [54], but treats whole networks rather than individual as the core variables for analysis. In this case, the expression levels of children within a network family are ordered from highest to lowest expression to define the family ranking—this ranking represents the relative activity levels of related networks. I can then use NF-DIRAC to assess how network family rankings change within phenotypes—to identify tightly and loosely regulated network families—or between phenotypes, signifying variably expressed network families.

Specific Aim 2. Reconstruct the first genome-scale metabolic network for glioblastoma multiforme.

Project 3. Metabolic network reconstruction and analysis

To reconstruct a metabolic network specific to GBM, I am using gene and protein expression data to predict the subset of metabolism active in the cancer. Specifically, I am using the Model Building Algorithm (MBA) [1] to systematically prune reactions from the well-curated, but generic *Human Recon 1* [68] and generate a draft network for GBM (**Figure 6**). Experimental studies with the U-87 GBM cell line—including metabolomics, siRNA knockdowns, and high-throughput growth phenotyping—conducted in my lab, in conjunction with *in silico* simulations, will be used to refine and validate the model. In this manner, I will be able to build the first computable, genome-scale model of metabolism for GBM.

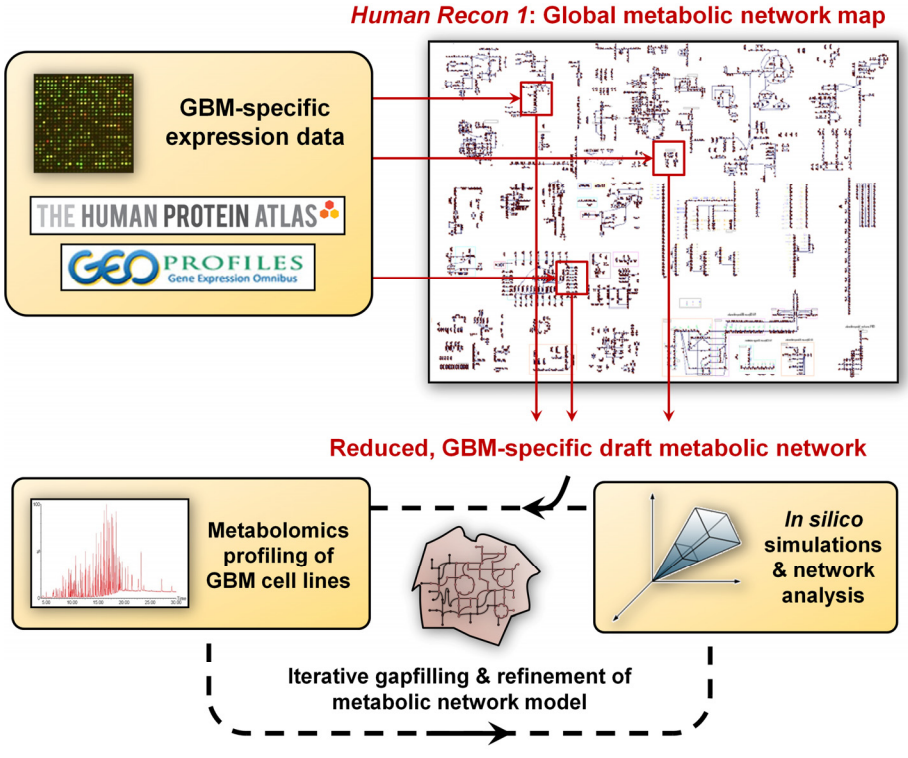

Figure 6. Overview of GBM metabolic network reconstruction.

Cell-specific expression data collection: Tissue-specific and cell-type specific expression data has been obtained primarily from the Human Protein Atlas (HPA) [69] and the Gene Expression Omnibus (GEO) [70]. 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. The HPA contains protein expression measures for three different GBM cell lines: U-138MG, U-251MG, and U-87 MG—I am basing the metabolic model

primarily on evidence for the U-87 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. The specificity of each antibody for expressed proteins is quantified 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 two evidence types. 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 cell line of interest. Protein expression evidence was mapped to genes in *Human Recon 1* based on NCBI gene ID. Protein expression evidence constitutes the *high confidence core* set of genes that are inputted into the Model Building Algorithm (MBA; described below).

To supplement the protein expression evidence obtained from HPA, I have compiled multiple datasets from GEO with gene expression profiles for GBM samples, including U-87 cell lines as well as biopsied tissue from glioma patients. After processing raw microarray files to determine whether probes are present or absent in each sample, the corresponding genes are included only if they are detected in all samples of a dataset. The union list of expressed genes across all GEO datasets comprises the *moderate confidence core* for MBA.

Model pruning with the Model Building Algorithm: I am using the Model Building Algorithm (MBA) described in Jerby et al. [1] to systematically prune reactions from the generic *Human Recon 1*, leaving GBM-specific network. The algorithm will iteratively remove reactions from *Recon 1*, conditional upon the requirement that all reactions encoded by the high confidence core genes remain unblocked; the objective of MBA is to maximize the number of moderate confidence core reactions while minimizing non-core reactions. I have developed a Matlab implementation that offers two convenient features: integration with the COBRA toolbox [71] and optional use of *fastFVA* [72]. The COBRA (Constraint-Based Reconstruction and Analysis) toolbox provides a suite of tools for constructing, curating, simulating, and analyzing metabolic network models in Matlab [71]. To take advantage of the efficient and user-friendly features of COBRA, I have written our 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. Specifically, the original version of MBA uses a heuristic optimization approach to evaluate the consistency of core reactions (see Jerby et al. [1] 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. Since 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 [72]; *fastFVA* was recently developed by Gudmunsson et al., and has been demonstrated to improve performance of FVA by 30 to 220 fold.

Network refinement with metabolomics profiling: Advances in technology for mass-spectrometry-based profiling of small molecules have led to an increased focus on characterizing the human metabolome, especially in relation to disease [73, 74]. High-throughput metabolomics data has a number of applications in medicine and systems biology, including biomarker discovery and elucidating function among the underlying biochemical reactions [33, 75]. At the minimum, global profiling of the metabolome in GBM cell lines may reveal missing components and functionality in the reconstructed network. In the first phase of the metabolomics study, my group will perform shotgun, qualitative metabolomics; the primary objective here is the discovery of small molecules present in the cell lines. I will use this metabolomics data to identify compounds not presently found in the metabolic reconstruction of GBM to aid in the gap filling process. Specifically, any metabolites detected that are not present in the current version of the reconstruction will direct us towards missing functionality (i.e., missing reactions or pathways) that need to be added. With the help of two post-doctoral fellows in my lab—one an experimental biologist and the other an expert in mass spectrometry—I plan to develop a protocol to perform metabolomics on GBM cell lines. With the equipment and expertise of the Roy J. Carver Biotechnology Center available at the University of Illinois, most of our work in the lab should involve determining optimal procedures for quenching and lysing U-87 cells before submitting samples to the Center for analysis. Ultimately, our aim will be to collect metabolomics profiles for different media conditions, a number of specific mutations, and if possible, varying time points in the cell cycle of U-87 lines.

Simulation and experimental validation: Constraint-based analysis [48] restricts the functional states of a network according to physico-chemical laws, known stoichiometric relationships, and other constraints. Linear algebra and optimization tools are then used to determine a steady-state flux distribution, given a defined cellular objective (e.g., optimal growth). Using the COBRA toolbox in Matlab [71], I will simulate GBM cell growth in a variety of conditions—genetic and environmental—to assess the predictive accuracy of the model; maximum growth is assumed to be a reasonable objective for our model, as this is a characteristic phenotype of most cancers [29-31]. I will initially use a

generic cancer biomass equation developed by Shlomi et al. to investigate the Warburg Effect using metabolic network modeling [76]. Experimentally measured values that will serve as criteria for evaluating accuracy and highlight discrepancies in behavior will include (i) growth rates; (ii) gene knock-out effects, including altered growth rates and lethality; (iii) secretion rates of selected by-products; and (iv) differential growth behavior on varied media.

Specific Aim 3. Provide mechanistic and functional context for high throughput mutation and expression data in glioblastoma using the metabolic model.

Project 4. Identifying perturbed metabolic pathways in GBM using network-level expression analysis tools

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. Additionally, the adaptive network modification approach being developed as an extension to DIRAC will be used to fine-tune metabolic subnetworks to identify the optimal signatures for distinguishing GBM from normal astrocytes. When operating on high-confidence pathways in the GBM metabolic network, DIRAC and NF-RXA 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.

Project 5. Investigating the distribution and effects of mutations in the GBM metabolic network

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 [66]. In addition to this data, which will primarily be used to investigate the effects of mutations on GBM metabolism, our lab is also performing RNA-seq experiments on several U-87 GBM cell lines with known mutations; while this cell line data cannot be used to study distributions of mutations, it can provide insights into the causes of specific mutations relative to the human genome.

Mapping mutation frequencies to the metabolic network: 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.

Simulating phenotypic effects of mutations in metabolic genes: Mutations in several key metabolic and signaling genes are almost universally implicated in the development of cancers, including GBM. While the altered function and downstream effect of these genes may be hypothesized, simulations with the genome-scale model will allow us to efficiently evaluate whole-cell metabolic outcomes in response to these aberrations. The simplest types of aberration that I can examine are loss of function mutations in metabolic enzymes; by blocking flux through the corresponding reactions, I can easily simulate the effect of these mutations on the entire network. For mutations leading to altered or novel function in enzymes, I can test a variety of alternative substrates and products—restricting the search to those already produced or consumed elsewhere in the network—to predict which reaction best explains experimentally-observed phenotypes; if specific altered functions are already hypothesized for an enzyme (e.g., *IDH1* and *IDH2* [30]), I can examine these directly. Furthermore, I can impose constraints to measure the qualitative effects of perturbations in signaling or regulatory pathways.

Conclusions

The systems-level methods and tools developed in this work will be immensely valuable for studying and understanding molecular perturbations in the highly complex glioblastoma multiforme. Using the mechanistically-detailed model I am building for GBM, I will begin to interrogate the link between mutations in the cancer and key metabolic processes that contribute to tumorigenesis. The model will also serves as a powerful tool for studying network-level gene expression changes in GBM. In the future, systems-level analysis of GBM with the model should continue to provide important biological insights into underlying mechanisms and potential treatment routes for the disease.

E. References

1. Jerby, L., T. Shlomi, and E. Ruppin, *Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism.* Molecular Systems Biology, 2010. **6**: p. 401.
2. Shlomi, T., et al., *Network-based prediction of human tissue-specific metabolism.* Nature Biotechnology, 2008. **26**(9): p. 1003-10.
3. Ohgaki, H., et al., *Genetic pathways to glioblastoma: a population-based study.* Cancer Res, 2004. **64**(19): p. 6892-9.
4. Howlander, N., et al., *SEER Cancer Statistics Review, 1975-2008*. 2011, National Cancer Institute: Bethesda, MD.
5. Kleihues, P., W.K. Cavenee, and International Agency for Research on Cancer, *Pathology and Genetics of Tumours of the Nervous System*. World Health Organization Classification of Tumours. 2000, Lyon: IARC Press. 314.
6. Christensen, H.C., M. Kosteljanetz, and C. Johansen, *Incidences of gliomas and meningiomas in Denmark, 1943 to 1997.* Neurosurgery, 2003. **52**(6): p. 1327-33; discussion 1333-4.
7. Hess, K.R., K.R. Broglio, and M.L. Bondy, *Adult glioma incidence trends in the United States, 1977-2000.* Cancer, 2004. **101**(10): p. 2293-9.
8. Reardon, D.A., et al., *Recent advances in the treatment of malignant astrocytoma.* J Clin Oncol, 2006. **24**(8): p. 1253-65.
9. 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.
10. Hanahan, D. and R.A. Weinberg, *The Hallmarks of Cancer.* Cell, 2000. **100**(1): p. 57-70.
11. 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.
12. 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.
13. Lloyd, A.C., et al., *Cooperating oncogenes converge to regulate cyclin/cdk complexes.* Genes & Development, 1997. **11**(5): p. 663-677.
14. Lowe, S.W., E. Cepero, and G. Evan, *Intrinsic tumour suppression.* Nature, 2004. **432**(7015): p. 307-15.
15. McMurray, H.R., et al., *Synergistic response to oncogenic mutations defines gene class critical to cancer phenotype.* Nature, 2008. **453**(7198): p. 1112.
16. Auffray, C., *Protein subnetwork markers improve prediction of cancer outcome.* Molecular Systems Biology, 2007. **3**(141).
17. Chuang, H.Y., et al., *Network-based classification of breast cancer metastasis.* Molecular Systems Biology, 2007. **3**(140).
18. Liu, E.T. and T. Lemberger, *Higher order structure in the cancer transcriptome and systems medicine.* Molecular Systems Biology, 2007. **3**(94).
19. Esteller, M., *Cancer epigenomics: DNA methylomes and histone-modification maps.* DNA, 2007. **1**(E2): p. E1.
20. Esteller, M., et al., *Cancer epigenetics and methylation.* Science, 2002. **297**(5588): p. 1807-8.
21. Jones, P.A., *DNA methylation and cancer.* Oncogene, 2002. **21**: p. 5358-5360.
22. 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.
23. Laird, P.W., *Cancer epigenetics.* Human Molecular Genetics, 2005. **14**(90001): p. 65-76.
24. Seligson, D.B., et al., *Global histone modification patterns predict risk of prostate cancer recurrence.* Nature, 2005. **435**(7046): p. 1262.
25. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
26. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
27. Sjoblom, T., et al., *The consensus coding sequences of human breast and colorectal cancers.* Science, 2006. **314**(5797): p. 268-74.
28. Parsons, D.W., et al., *An Integrated Genomic Analysis of Human Glioblastoma Multiforme.* Science, 2008. **321**(5897): p. 1807.
29. Hsu, P.P. and D.M. Sabatini, *Cancer Cell Metabolism: Warburg and Beyond.* Cell, 2008. **134**(5): p. 703-707.
30. Cairns, R.A., I.S. Harris, and T.W. Mak, *Regulation of cancer cell metabolism.* Nature Reviews Cancer, 2011. **11**(2): p. 85-95.
31. Kroemer, G. and J. Pouyssegur, *Tumor Cell Metabolism: Cancer's Achilles' Heel.* Cancer Cell, 2008. **13**(6): p. 472-82.
32. 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.
33. 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.

34. 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.
35. Serkova, N.J. and K. Glunde, *Metabolomics of cancer.* Methods in Molecular Biology, 2009. **520**: p. 273-95.
36. Glunde, K. and N.J. Serkova, *Therapeutic targets and biomarkers identified in cancer choline phospholipid metabolism.* Pharmacogenomics, 2006. **7**(7): p. 1109-23.
37. Pelicano, H., et al., *Glycolysis inhibition for anticancer treatment.* Oncogene, 2006. **25**(34): p. 4633-46.
38. Griffin, J.L. and R.A. Kauppinen, *A metabolomics perspective of human brain tumours.* FEBS J, 2007. **274**(5): p. 1132-9.
39. Feist, A.M., et al., *Reconstruction of biochemical networks in microbial organisms.* Nature Reviews Microbiology, 2009. **7**(2): p. 129-143.
40. 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.
41. Reed, J.L., et al., *Towards multidimensional genome annotation.* Nature Reviews Genetics, 2006. **7**(2): p. 130-141.
42. 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).
43. Feist, A.M., et al., *Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri.* Molecular Systems Biology, 2006. **2**(2006.0004).
44. 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.
45. Chavali, A.K., et al., *Systems analysis of metabolism in the pathogenic trypanosomatid Leishmania major.* Molecular Systems Biology, 2008. **4**(177).
46. 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.
47. Price, N.D., et al., *Genome-scale microbial in silico models: the constraints-based approach.* Trends Biotechnol, 2003. **21**(4): p. 162-9.
48. 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. 886-97.
49. 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.
50. Oberhardt, M.A., B.O. Palsson, and J.A. Papin, *Applications of genome-scale metabolic reconstructions.* Molecular Systems Biology, 2009. **5**: p. 320.
51. 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.
52. Subramanian, A., et al., *GSEA-P: a desktop application for Gene Set Enrichment Analysis.* Bioinformatics, 2007. **23**(23): p. 3251.
53. Lee, E., et al., *Inferring Pathway Activity toward Precise Disease Classification.* PLoS Computational Biology, 2008. **4**(11).
54. Eddy, J.A., et al., *Identifying tightly regulated and variably expressed networks by Differential Rank Conservation (DIRAC).* PLoS Computational Biology, 2010. **6**(5): p. e1000792.
55. 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.
56. 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.
57. 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.
58. 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.
59. Lee, D.S., et al., *The implications of human metabolic network topology for disease comorbidity.* Proceedings of the National Academy of Sciences, 2008.
60. Ma, H., et al., *The Edinburgh human metabolic network reconstruction and its functional analysis.* Molecular Systems Biology, 2007. **3**(135).
61. 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.
62. Gille, C., et al., *HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology.* Molecular Systems Biology, 2010. **6**: p. 411.
63. Ahn, Y.Y., J.P. Bagrow, and S. Lehmann, *Link communities reveal multiscale complexity in networks.* Nature, 2010. **466**(7307): p. 761-4.

64. Pudil, P., J. Novovicová, and J. Kittler, *Floating search methods in feature selection.* Pattern recognition letters, 1994. **15**(11): p. 1119-1125.
65. 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.
66. TCGA, *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-8.
67. 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.
68. 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 of the United States of America, 2007. **104**(6): p. 1777.
69. Berglund, L., et al., *A genecentric Human Protein Atlas for expression profiles based on antibodies.* Molecular and Cellular Proteomics, 2008. **7**(10): p. 2019-27.
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. 207-10.
71. Becker, S.A., et al., *Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox.* Nature Protocols, 2007. **2**(3): p. 727-38.
72. Gudmundsson, S. and I. Thiele, *Computationally efficient flux variability analysis.* BMC Bioinformatics, 2010. **11**: p. 489.
73. Pearson, H., *Meet the human metabolome.* Nature, 2007. **446**(7131): p. 8.
74. Wishart, D.S., et al., *HMDB: a knowledgebase for the human metabolome.* Nucleic Acids Research, 2009. **37**(Database issue): p. D603-10.
75. Weckwerth, W., *Metabolomics: an integral technique in systems biology.* Bioanalysis, 2010. **2**(4): p. 829-36.
76. Shlomi, T., et al., *Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect.* PLoS Computational Biology, 2011. **7**(3): p. e1002018.