IDENTIFICATION OF GENES AND THEIR REGULATION THAT DETERMINE A PHENOTYPE: A SYSTEMATIC APPROACH

By

Ming Wu

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Computer Science

2012

ABSTRACT

IDENTIFICATION OF GENES AND THEIR REGULATION THAT DETERMINE A PHENOTYPE: A SYSTEMATIC APPROACH

By

Ming Wu

Research in computational systems biology focuses on establishing the complex relationship and interactions between genes and how they work together to render a particular phenotype. This involves the development and application of systematic approaches to study the biological regulation in the context of a network in which genes are regulating each other. Our research aim to develop novel approaches to identify genes and their regulation that determine a phenotype, which involves the reverse engineering of regulatory mechanisms through identification of condition specific genes and interactions, as well as the systematical modeling and simulation to reconstruct context-dependent regulatory networks.

Chapter 1 introduces the fundamental approaches in systems biology. Data mining techniques have been developed to identify genes and interactions from gene expression data, while systems modeling integrate current knowledge to develop a functional context to address the complexity that arises in biological systems. We provide examples to demonstrate the practical aspects and biological relevance of the methodologies. Chapter 2 introduces and discusses the multi-layer approach that is able to reconstruct condition-specific genes and their regulation through an integrative analysis of large scale information of gene expression, protein interaction and transcriptional regulation. In Chapter 3 we explore a dynamic feature of gene network— the switch-like behavior, wherein we show that gene switches have specific pattern of gene expression which can be uncover by mining microarray data. This study demonstrates that one can capitalize on genome-wide expression profiling to capture

dynamic properties of a complex network, thereby predicting gene switches that could be important for a phenotype and can participate in cell fate decision. In Chapter 4 the cancer phenotype is studied using systems modeling of the human metabolic network. We develop a novel approach to simulate context dependent metabolic states that upon perturbation of the gene(s) that modulate metabolic functions, can determine whether the gene is involved in confering a phenotype. The approach is then applied to predict therapeutic microRNAs for human hepatocellular cancer. Chapter 5 provides a brief summary of the implications of the research towards a systematic understanding of gene network as well as a future perspective of the field.

ACKNOWLEDGMENT

First I would like to thank my advisor Dr.Christina Chan, and other members in the research group. I would like to thank members in my guidance committee: Dr. Rong Jin, Dr. Yuehua Cui, and Dr. C. Titus Brown, and my collaborates Dr. S. Patrick Walton, Li Liu, Amanda Portis Malefyt, HyunJu Cho, Dr. Xuerui Yang, Dr. Linxia Zhang, and Xiao Pan. Finally I am very grateful for the great support from my wife and my Mom during the PhD study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

A central question in biological science is the gene(genotype)-phenotype relationship, i.e. how do the genes and their regulation determine a phenotype.

The *genotype* used to define by a single genetic trait in Mendel's experiments. This concept has been extended to a large variety of genetic endowment of living organisms at the molecular level, involving gene and gene network upon genetic and epi-genetic regulations. In genetics, the genotype could be mutants (e.g. SNPs), insertion/deletions, or duplications in a genomic region, or some modifications on the chromatin. In cell biology, as we focus on throughout this dissertation, a particular genotype could be a signaling molecule, a transcription factor, a pathway, or a gene network involving molecules that can affect a cellular process.

The *phenotype* represents the biological observations to be investigated. Depending on the research question and the experimental design, a phenotype could be a particular cell type, a decision of cell fate, a biological process, or some physiological or disease states.

Research in molecular biology focuses on establishing the relationships between genes, interactions of genes and their functions in regulating the processes by which cells respond

to external or internal signals to determine a phenotype, i.e. to answer the question: what is the underlying molecular mechanism that governs a particular state of a biological system?

The question has been studied previously by the identifying and analyzing the function of individual genes and proteins. However, the recent development of high-throughput techniques has driven the need to find global changes at the "omics" level, in order to elucidate how the genes interact to regulate biological processes. The ability to routinely study thousands of genes' expression has shifted the paradigm in the biological research community, suggesting that genes and their interactions should be evaluated in the context of the whole network. To gain a better understanding of the complexity of biological regulation has raised new challenges in analyzing the data, designing the experiments to validate these analysis and the processing of the information. Thus, computational and systems biology approaches are rapidly being developed to analyze and integrate omics data to obtain a coherent and mechanistic snapshot of cellular regulation.

1.1 The paradigm shift and computational systems biology

Since the advent of molecular cell biology, researchers have studied biological phenomenon mainly by analyzing the function(s) of individual genes and proteins, and the change(s) they exhibit in diseased states. This reductionist approach helped discover many of the underlying biological principles [1]. However, researchers subsequently found that the relationship between genotype and phenotype is more complicated then can be ascribed to a change in a single gene [2], and the behavior of the different components in the biological system

cannot be captured in isolation. These observations, together with the recent availability of 'omics' data, have revolutionized the previous view of single gene-phenotype correlation by demonstrating the importance of the inter-relationships between genes, as Linus Pauling said "Life is a relationship among molecules and not a property of any molecule". This has intensified the investigation of protein function in the context of complex biological systems, and initiated a more systematic perspective of biological processes [3].

In the last decades biologists have become increasingly aware of the importance of functional context, the community has been enthusiastic about the paradigm shift from cataloguing molecular characterizations to understanding the functional activities of genes and proteins specific to a phenotype. More and more experiments are guided by models that serve as a basis for generating hypothesis and interpreting results (e.g. regulatory schemes or pathway diagrams). Nevertheless, biological processes consist of many interacting components, exceeding the human capacity to systematically analyze them, thus requiring computational methods to reduce the complexity and thereby enhance their accessibility [4]. Thus, a central idea in computational systems biology is to construct mathematical formulation of data-driven or hypothesis-generating models to help reveal the function of biological systems [2].

Computational systems biology gathers gene expression, interaction, or perturbation data to build a specific network that regulates a biological process, and studies how the design features of the network specify biological decisions. One of the main focuses has been predominantly on analyzing expression data to understand the regulation of gene expression and its functional activity for a phenotype, wherein a "systems modeling" or "systematic approach" is to address the following two questions:

- How does one identify the genes and interactions that are specific for a phenotype/condition, given the large amount of gene expression changes that can be measured in high-throughput techniques?

- How do genes interact with each other to determine a phenotype?

The two questions are attempting to systematically establish the relationship between gene expression and phenotype from two different aspects, i.e., learning from data to identify important genes, and modeling based on mechanism to understand a phenotype. Therefore, there are two distinct fields of computational modeling in systems biology that are being developed to address these two questions: data mining and *in-silico* simulations. Data mining, or "top-down" modeling, aims at discovering patterns from large amount of high-throughput biological data to help generate hypotheses and make predictions, by learning data driven models to establish a statistical relationships between genotype and phenotype. Alternatively, simulation-based analysis, or "bottom-up" systems modeling tests hypotheses with *in-silico* experiments and predict network behavior by integrating current knowledge of molecular mechanisms.

In this chapter I introduce these two different computational modeling approaches in systems biology and explain the importance of identifying the genes and interactions that are specific to a phenotype.

1.2 Learning from data

Data mining, or "top-down" modeling, learns data driven models to establish statistical relationships between genotype and phenotype. Data mining techniques have been widely

applied on gene expression data to identify genes and interactions that are important to a phenotype.

1.2.1 Current approaches to identify genes/biomarkers based on expression data

Identification of important genes based on gene expression microarray data, also known as "biomarker discovery", "analysis of differentially expressed genes", is essentially a *feature selection* problem in data mining. The gene expression data is collected from samples of different phenotypes, or, different "annotations". The expression levels across all samples of each gene are investigated to discover genes that are relevant to particular phenotype/conditions, or other target annotations. These genes can be used as biomarkers that could be valuable in disease prediction, drug target discovery, or further analysis to reconstruct the regulatory network that underlies the manifestation of the phenotype.

Problem statement: Given a gene expression matrix M obtained from a series of preprocessing steps involving transformation and normalization of the raw data, in which there are m genes (features) across n samples, $m >> n$ in most cases, and x_{ij} is the expression level of gene i and sample j. The problem becomes how to identify the genes (features) that govern any one of the phenotypes in the samples.

In the field of data mining and machine learning, feature selection usually aims to find the features that provide *better classification*, thus the conventional categorization of feature selection techniques is based on the relationships between the selection approach and the classification model: [5, 6]

- Filter methods assess the relevance of features by analyzing the data matrix, which is

a preprocessing step, independent of the classifier, that will be used in the subsequent learning problem.

- Wrapper methods assess the features with classifiers, thereby the features are evaluated based on a particular classifier, e.g. SVM. The methods attempt to select the most discriminant subset of features by minimizing the prediction error of the classifier chosen.

- Embedded methods search in the combined space of feature subsets and the model hypothesis set in the classifiers.

However, in the application of feature selection to biology to identify specific genes, the biological question asked may not be necessarily based on the classification accuracy of the phenotypes, but in most cases it is to understand the phenotypes, e.g.:

- What is the most important difference between two phenotypes in terms of gene expression?

- What are the characteristic genes for cancer?

Thus filter approaches are the most widely applied in the literature to identify differentially expressed genes [7]. Univariate approaches dominate the field since the output is a ranking of the genes that are intuitive and easy to interpret and validate. Multivariate approaches have been developed to account for potential interactions between genes' expression. Filter approaches in feature selection can be separate into two classes [7]:

- Ranking approaches: define a certain scoring function to estimate the relevance indices (scores) for each gene and select genes that are ranked at the top.

- Space searching approaches: features are selected by optimization of a predefined cost function.

Ranking approaches Ranking approach first define a scoring function $S(x)$ to quantify the difference in gene expression between different groups, score the genes and rank them based on the scores estimated. Usually a higher score indicates an important gene that is expressed differentially. The choices for the scoring functions include parametric and non-parametric models.

The approaches that are mostly widely used are parametric models which assumes a Gaussian distribution. One could use t-statistics, fold change or ANOVA to quantify the divergence or distance between the sample distributions from different groups (phenotypes/conditions) for a given gene. For example, the commonly used scoring function for gene expression [8] that applies Welch t-statistics is:

$$S = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma^2_{\bar{x}_1}/n_1 + \sigma^2_{\bar{x}_2}/n_2}} \tag{1.1}$$

in which x_1, x_2 are the gene expression in the samples from two groups/phenotypes, and n_1,n_2 are the number of samples in these two groups. There are modifications of the standard t-test to address the potential problem of the small sample size and noise in the array data, either by adjusting the estimation of variances or applying prior degrees of freedom and variances in a Bayesian framework [9, 10].

Non-parametric approaches do not assume a target distribution but compares and combines the ranking of the gene expression level across different samples. In these approaches, a "ranking" is assigned to samples in different groups based on one gene's expression level.

For each gene, the score is computed by taking the sum or average of the ranking of those samples in a particular group. The scoring functions for the Wilcoxon rank sum [11] and Rank product [12] are shown as follows:

$$S = \sum_{j=1}^{k} R_j,\ k = \min(n_1, n_2) \tag{1.2}$$

$$S = (\prod_{n}^{j=1} {R_j}^{1/n}) \tag{1.3}$$

in which R_j is the ranking of sample j based on the expression level of the gene being investigated.

There are other types of scoring functions based on information theory to account for more complex relationships (e.g. non-linear, non-monotonous) between gene expression and phenotypes , e.g.:

$$S = \sum_{x_i} \sum_{c} P(x_i, c) \log \frac{P(x_i, c)}{P(x_i)P(c)} \tag{1.4}$$

in which $P(x_i, c)$ is the joint probability gene x_i and group/phenotype c.

After scoring and ranking the genes, in many applications researchers might set a threshold to select genes. A threshold could be a pre-defined fold-change difference between gene expression in the different groups (2-fold change is usually used as a threshold in earlier studies), or based on the selection of genes that minimize the training sample mis-classification rate if the goal is to classify the phenotypes, or the statistical significance of the score. The statistical significance tests estimate the possibility that a particular score would have been obtained by chance. P-value is commonly used, which represents the probability of achieving a test statistics that is at least as extreme as the one observed in the data. Although there

are standard asymptotic methods based on a pre-defined distribution of the test statistics, gene expression data might not follow the distribution and the sample size of the data is usually not sufficiently large for such analysis. Therefore p-values are estimated empirically by applying permutation test on the data, i.e. running tests (computing scores) which are identical to the original scoring approaches but the target feature (e.g. the grouping of samples) is permuted differently for each test, whereby a null distribution of the scores can be generated to estimate p-values. Another issue of concerned in these tests is the *multiple hypothesis testing problem* [13] due to the large number of genes that are simultaneously tested. The threshold should be adjusted to control the error rate (e.g. type I error: the false positive rate). P-value with Bonferroni correlation or False Discovery Rate are commonly applied to address the problem [14].

Space search approaches Instead of scoring and ranking individual genes, space search approaches aim to find the subgroup of features (genes) which optimize a pre-defined cost function. Cost functions in data mining are usually defined to maximize the information content of the features while minimizing the redundancy, e.g.:

$$MIQ = \max_{i \in \Omega_m} \frac{I(x_i, c)}{(1/|\Omega_s|) \sum_j I(x_i, x_j)} \tag{1.5}$$

The function computes the Mutual information quotient [15], in which the $I(x_i, c)$ is the mutual information between gene x_i and class label/phenotype c, and $I(x_i, x_j)$ is the mutual information between the gene candidate $x_i \in \Omega_m$ to be added into the feature subset and the genes $x_j \in \Omega_s$ that have been included in the subset. Ω_m is the entire gene set and Ω_s is the current selection of a subset of features/genes.

Although these feature selection approaches are widely applied in the identification of genes that are specific for a phenotype, there are issues in the assumptions implicit in these approaches that are inconsistent with the biological processes. Many ranking approaches assume features are independent, but the genes that regulate each other and their expression should not all be independent. Space search approaches account for potential relationships between features but biologically the "redundant" genes should not be removed if they are also specific or relevant to a phenotype. We therefore suggest applying the Relief method that optimizes a distance function to seperate samples from different groups and results in a ranking of genes, which can address some of these problems. The details of our approach are discussed in Chapter 2.

1.2.2 Current approaches to identify condition-dependent gene interactions

To uncover the complex gene regulatory network that renders a phenotype or disease state of a biological system, not only does one need to identify specific genes but also identify the interactions that are specifically functioning to render the phenotype. These interactions can be discovered based on gene expression data to suggest some specific association, regulatory relationship or functional relationship between genes. This step could be consider another *feature selection* problem, i.e. given all possible interactions, e.g. the complete graph connecting all gene pairs, or based on some known biological network, how does one select the interactions that are specific to a phenotype?

Similar to the filter approaches introduced in the last section to identify specific genes, the methods to identify specific interactions can either define a scoring function to rank

interactions, or search the space of possible network structures that can best explain the (in)dependencies within the expression profiles.

A scoring function that could be used to quantify the potential association between two gene expression profiles is Pearson correlation:

$$R(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} \tag{1.6}$$

for gene X and gene Y: Nevertheless the dependencies between two genes' expression could be more complex than a linear relationship, thus researchers have used mutual information to account for all different types of dependencies:

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{1.7}$$

The probabilities in the equation are meant to be estimated from frequency counts of discrete variables, however gene expression measurements are usually continuous variables. One may discretize the variables or approximate the probabilistic densities with non-parametric approaches such as Parzen windows [16].

After calculating all the pair-wise scores one needs to select the features/interactions that are "important". Algorithms have been developed to reduce the number of candidates and thereby identify important gene interactions, including:

- MRNET (maximum relevance/minimum redundancy, MRMR) [17] removes redundant genes connecting to a target gene. For example, gene X_1 and X_2 both have high mutual information with gene Y, thus one could connect $X_1 - Y$ and $X_2 - Y$. Nevertheless, if X_1 and X_2 are correlated, then they are "redundant" with respect to their connections

to Y, thus the algorithm will delete one of the connections to Y. When there are many other genes that can be connected to Y, the algorithm selects from the least redundant set of genes the ones that have the largest mutual information with gene Y.

- CLR(Context Likelihood of Relatedness) [18] suggests that the connections between different genes should apply different thresholds. CLR computes mutual information $MI(X, Y)$ between genes X and Y, and generates a distribution of mutual information to compute z-scores to evaluate the $MI(X, Y)$, i.e. to estimate the $MI(X, Y)$ relative to all possible MIs involving gene X or Y. In other words, is $MI(X, Y)$ much higher than most of the other potential "background interactions" to gene X and gene Y ($MI(X, *)$ and $MI(Y, *)$ in which $*$ represents any other gene that could interact with X or Y. Finally, whether or not gene X and gene Y should be connected is determined by a hypothesis test of Z_{XY} which determines whether the mutual information between X and Y is significantly higher than the background.

- ARACNE(Algorithm for the Reconstruction of Accurate Cellular Networks) [19] attempts to mathematically define and separate the direct from the indirect effects. If one knows the expression of genes X and Y are correlated, genes Y and Z are correlated as well, one would expect that X and Z could be correlated. ARACNE thus compares the triplets: $MI(X, Y)$, $MI(Y, Z)$, $MI(X, Z)$ in order to remove indirect effects. Based on intuition that an indirect effect could be smaller than the direct effects, ARACNE defines a "Data Processing Inequality" to test all gene triplets such that for each triplet (X, Y, Z), the connection corresponding to the lowest mutual information $MI = \min\{MI(X, Y), MI(Y, Z), MI(X, Z)\}$ is elimiated if it is lower than

a threshold:

$$MI \leq MI_2(1 - \epsilon) \tag{1.8}$$

where MI_2 is the second lowest mutual information for the triplet and the ϵ is a pre-defined parameter.

Although any of these methods can effectively reduce the number of candidates, the assumption of MRNET is questionable with respect to biology, since there could be "redundant" interactions that are effectively functional under a condition to provide a robust design of a biological function (e.g. different isozymes). CLR removes "false correlations" in the network by eliminating promiscuous cases wherein a gene weakly co-varies with a large number of other genes. This is more reasonable since such promiscuity can arise when the assayed conditions are inadequately or unevenly sampled, which could be real in biological experiments [18]. From another perspective, the many "indirect effects" is a major problem in biological studies of regulation of gene expression. ARACNE provides a heuristic definition help focus on the direct interactions, thus has raised more interest and had some success in the community.

1.2.3 The reconstruction of transcription regulatory network

Overall, the basic idea of "reconstruction of gene regulatory network" is the same as inferring gene interactions — to identify pair-wise relationship between the genes, or more specifically, to determine whether a gene (or its product) directly controls the expression of another. By learning the dependencies contained within the expression profiles, researchers are attempting to reconstruct a network that depicts the global regulatory network of genes. The development and application of the theory and tools for network inference, or so-called

"reverse engineering" tasks, have been predominantly based on data mining/learning techniques. Many approaches mentioned above have been used in network reconstruction. There are other approaches that are not built upon the scoring of every single interactions but searching in the space of possible network structures to identify the best structure to explain the data. One example is the Bayesian Network (BN) [20], which aims to determine a directed acyclic graph that can represent and simplify the joint probability of the expression of all the genes investigated, in which descendants are independent to each other given their parent nodes so that the joint probability are separable based on such independencies learned from data.

There are several in-depth reviews in the literature that introduce and compare these different modeling schemes and more recent developments for learning [21–23], including a primer on regression methods [24]. In the review by Margolin et al. [25] they examine the theoretical underpinnings behind current reverse-engineering algorithms that are based on systems control theory (e.g. linear or non-linear regression model), probabilistic graphic learning, and information theory.

However, these reviews and literature in this area have generally been concerned about the practical aspect of data mining, and few have paid attention to the relevance of the methodology to the biological problem. The focus on the difficulties in network reconstruction has been from the computational perspective. A major discussion point has been the limited sample size available (in most cases less than a hundred samples) for identifying pair-wise relationships between the genes (which could be hundreds of thousands pairs). This results in an under-determined system requiring methods for dimensionality reduction, i.e. clustering or module/pathway analysis [21, 26, 27]. Another challenge is the large search

space of possible regulatory schemes, which requires either advanced optimization strategies or a priori information to reduce the computation time [22, 28, 29]. Nevertheless, in addition to previous improvements to the "predictive power"or "computational efficiency", it is important to understand how much biological information can be appropriately extracted from expression data to deduce the rules of gene expression/transcriptional control.

From a biological perspective, instead of studying the physical network, many of these reverse engineering methods are actually learning the *influence network* and trying to interpret the influence with gene expression control (the transcriptional regulation in most cases), which results a *transcriptional regulatory network* but with a interwoven mixture of true and promiscuous, direct or indirect effects [25], thereby creating a considerable divide between the influence network that is constructed and the real biological regulatory mechanisms. This is due largely to the limitations of the dataset itself, and the presence of multiple, unobserved levels of regulation leading to difficulties in the biological interpretations and undermining the biological significance of these influence networks and their further applications.

The reconstruction of gene network (inferring gene interactions) is widely applied to identify transcriptional regulatory network and transcriptional regulatory relationship between transcription factor and genes, thus we attempt to dissect the information content in the expression data from a biological perspective, and scrutinize the biological foundations of the computational models, and critically analyze the underlying assumptions of most *in silico* learning approaches applied to expression data that confounds the interpretation of the results [25], which are:

1. Statistical dependencies exist between the transcription factors and their target genes with respect to both their expression levels.

2. Measurements of the relative amount of mRNA level in the microarray data are predictive of the activity of the regulatory molecules. This assumption can be further sub-divided into three sub-types, as follows:

    - Type 1 model assumes the expression level of a transcription factor correlates with the activity of the transcription factor. (e.g. [30, 31])
    - Type 2 model estimates the activity of a transcription factor based on the behavior of its target genes. (e.g. [27, 32, 33])
    - Type 3 model assumes co-expression implies co-regulation by the same transcription factor, and estimates the existence or activity of an uncharacterized cis-motif by clustering analysis. (e.g. [34])

Using the yeast microarray data as an example, we combine information on the yeast transcriptional regulatory network and different data-sets and -types to examine each of these assumptions.

The association between transcription factor and their targets To illustrate the first assumption, we used yeast data to characterize the information in the expression data that is used to infer interactions at the transcriptional level. The yeast dataset contains 255 conditions from environmental stress [35] and cell cycle [36] microarray experiments. These datasets have been widely applied in previous studies to develop novel reverse-engineering methods [34]. Actually the yeast environmental stress response dataset has been cited 2,260 times so far, and among the 1,000 of these citations that are related to computational studies, there are 256 papers discussion about network reconstruction (citation data is provide by googleScholar, `http://scholar.google.com`), which suggests the popularity of the dataset.

Many of the yeast transcription factors have been studied and their *cis*-regulatory modules on gene promoters across the genome have been identified and are now available in public databases such as YEASTRACT [37, 38], thus enabling the attainment of a putative transcriptional regulatory network based on known motifs on the gene promoters collated in YEASTRACT.

Since the fundamental assumption is that the expression level of a gene depends on its regulators, we calculate the correlation between the expression level of the transcription factors and that of their target genes, where the target genes of a transcription factor in the regulatory network are identified by corresponding *cis*-motifs on their promoters. As shown in Table 1.1 on page 18, the average absolute correlation coefficient of the expression data, taking the absolute value since both positive and negative correlation represents perceptible dependencies, is about 0.08 between the transcription factors and their target genes. This appears to be negligible, even smaller than the background with a correlation of 0.19 between any gene pair, suggesting that it is difficult to directly identify TF-gene (a transcription factor and its target gene) pairs based on the dependencies of their expression.

Characteristic	Avg. Corr (variance)
Background (all gene pairs)	0.19 (0.02)
Overall (TF–target gene)	0.08 (0.02)
1-to-1 (gene–its only known TF)	0.16 (0.02)
1-to-1 (gene–its only known TF) (TF highly expressed)	0.18 (0.02)

Table 1.1: The average Spearman correlation between the expression level of the transcription factors and the expression level of their target genes, considering the conditions when the transcriptions factors are highly expressed (higher than their own mean level for all available conditions). "1-to-1" considers genes with only one known TF that can bind to their promoters.

Next, we place a "1-to-1" constraint that considers only the TF-gene pairs in which the target gene has no other type of known effectors besides the transcription factor paired to it. That is in contrast to an "n-to-1" relationship in which many transcription factors regulate the expression of one gene. We identify 596 "1-to-1" TF-gene pairs in the yeast network. This constraint assumes that the target genes of these 596 pairs are not regulated by multiple transcription factors. We found that the average correlation of the expression data of these pairs (0.16), albeit still lower than background noise (0.19), is about two times higher than the overall TF-gene pairs (TABLE 1),. This suggests that the combinatorial regulation of multiple transcription factors on a target gene plays an important role in the transcriptional regulation, thereby complicating the TF-gene relationship and the use of correlation of their expression profiles for inferring regulatory networks. Moreover, the correlation between a TF-gene pair could be increased slightly when the samples containing lowly expressing transcription factors are removed. A rationale for doing this is that low expression level may suggest reduced control by the regulator [39].

Overall our results demonstrate a weak correlation in the expression exists between the transcription factors and their target genes, thus making it difficult to uncover transcriptional regulation due to the high background. The high background could possibly be due to both direct and indirect associations between the genes in the network. The result is consistent with previous observations [40] that only a very small proportion of transcription factors' mRNAs are significantly correlated with the expression level of its target genes. Our results also indicate that combinatorial regulation contributes to the reduced correlation between the TFs and their target genes.

Besides the combinatorial effect, there are many complex features on the binding sites

or DNA-TF interactions that could impair the association between transcription factors and their target genes [41]. A transcription factor may not bind to all its targets with the same binding affinity, indeed many specifically interact with only a few targets depending on other genomic features, i.e. DNA modifications, or due to stochasticity of the binding events [42]. Differences in the sequence of and around the binding site will also affect the binding affinity [43]. Thus, to determine bindings that are functional remains a challenge [44–46]. Moreover, there are other levels of regulation that cannot be obtained from the expression data and transcriptional regulatory network. For example, the protein-DNA interactions may depend on other protein co-factors in order to become functional, and sometimes the binding itself requires adaptors [41, 47]. The mRNA level of the target genes may also be regulated at the post-transcriptional level, through the coordination of different rates of mRNA decay [48].

Incorporation of information from multiple sources It is increasingly evident that the network reconstruction by microarray data alone is imperfect, in part due to the limited sample size use to infer a very complex network, but more importantly because of the limitation of the information content of microarray, in which only the mRNA expression level is measured. Incorporating biological knowledge on different levels of regulation (e.g. protein interactions) could improve the results. Many recent studies on reverse engineering have attempted to integrate these information but focus on the technical utility of multi-source information in providing priors to limit the search space, and enable further validation or promote more intrinsic learning models [30, 49–51]. Here we show from a biological perspective, in an intuitive but quantitative manner, an enhancement in the TF-gene correlation by incorporating other sources of information with the yeast expression data.

We consider the regulation of possible protein co-factors on the transcription factors,

which have been suggested to impact the binding and activation of transcription factors. We impose an experimentally confirmed yeast protein-protein interaction (PPI) network on our transcription regulatory network. The physical interaction data is downloaded from SGD `http://www.yeastgenome.org/`, and we compute for each transcription factor the number of its potential interacting proteins—- given that the PPI is static and context independent. We found that among all our 1-to-1 TF-target gene pairs (correlation of 0.16), those transcription factors with a low number of possible co-factors exhibits significantly better TF-gene correlation ($p < 0.05$, average correlation coefficient as high as 0.25), as shown in Figure 1.1 on page 22.

In Figure 1.1 on page 22, part A, 596 1-to-1 TF-target gene pairs are considered. The data is categorized into three groups according to the number of interaction partners of each transcription factor in the yeast protein-protein interaction network. The group "Top 10%" consists of TF-target pairs in which the transcription factor has many interactions (more than 97 possible cofactors). The group "Bottom 10%" consists of pairs where the transcription factor has few protein interactions (no more than 2 cofactors). The others (80%) are in the "Middle" group. The group "Bottom 10%" exhibits significantly better average TF-gene correlation ($p < 0.05$).

Since the regulatory mechanisms depend on the dynamics of TF-gene binding, incorporating more binding information as a constraint should better reveal the TF-gene relationship. We searched the ChIP-Chip data and found a conditional ChIP-chip dataset [52] containing 23 transcription factors under H_2O_2 ($0.4nM$) treatment of yeast for which we also have microarray expression profiles. We therefore can define conditional TF-gene pairs using actual binding profiles (i.e. in the CHIP-chip data a significant p value indicates TF binding

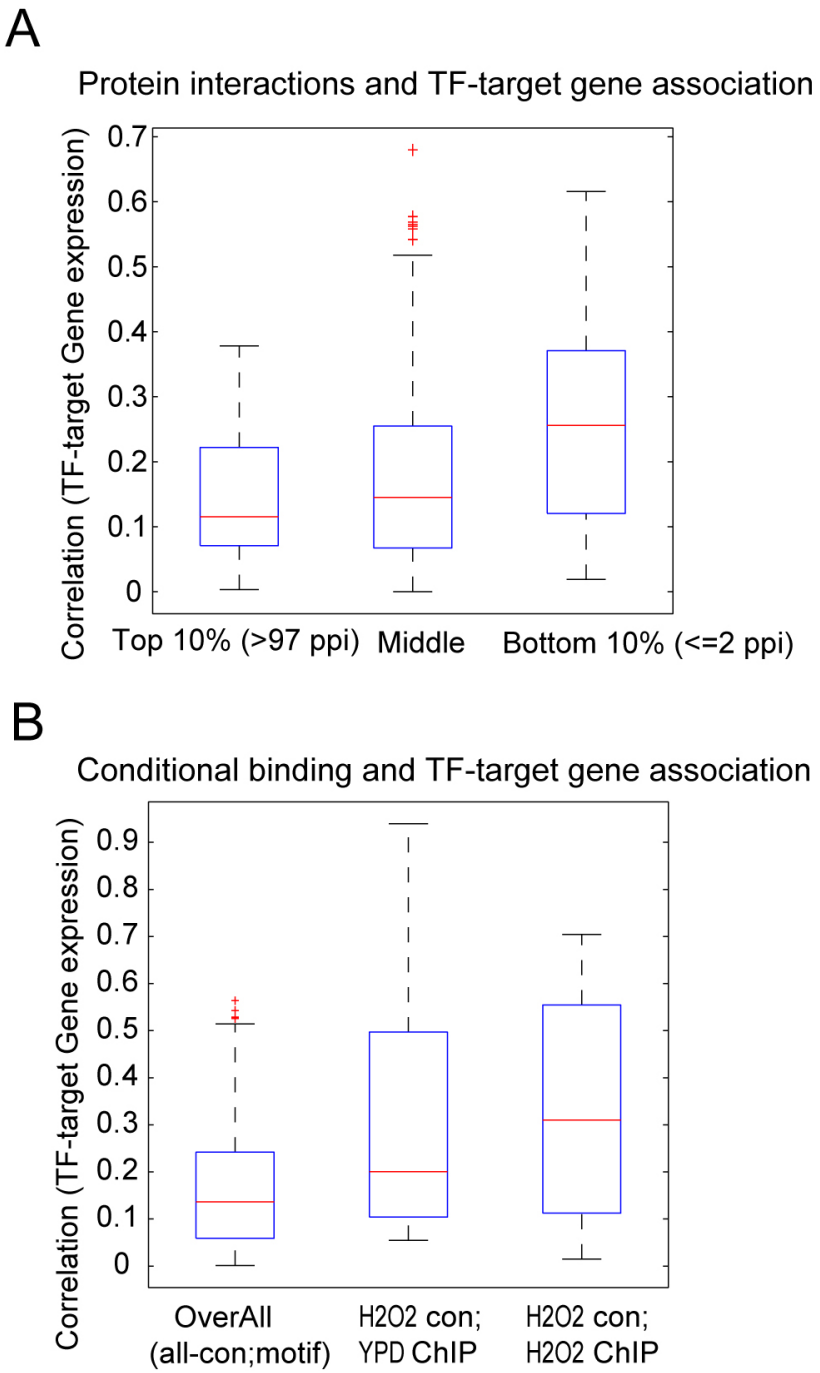

Figure 1.1: Incorporation of information from multiple sources. **A:** Box plots of the average correlation between the expression level of transcription factors and the expression of their target genes. **B:** Box plots of the average correlation between the expression level of transcription factors and the expression of the target genes. For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

on the gene promoters), rather than the sequence-level motif analysis which indicates only the capability of TF binding instead of actual binding events. We compare the correlation coefficient of TF-gene pairs defined by *cis*-motifs, or by non-conditional CHIP-chip (under normal growth condition in YPD media), with that of gene pairs defined by conditional binding information, the result is shown in part B of the Figure 1.1 on page 22. since CHIP-chip data is not available for every transcription factor, only 99 pairs have binding information and 20 of them are (condition-specific) conditional binding. The overall gene-TF (box on the left) describe the average correlation of all 596 pairs under all conditions. Under the H2O2 condition, the box in the middle (YDP-CHIP) use the non-conditional binding (CHIP-chip results under normal condition, 99 pairs) to determine the average correlation of the Gene-TF pairs that actually bind, and the box on the right use conditional binding CHIP-chip data (20 pairs). The average TF-gene correlation increased significantly when conditional binding data is available (box on the right, $p < 0.01$ compared with the box on the left). The average correlation increased significantly (coefficient > 0.3) when conditional binding data is available, confirming the importance of context-specific information.

Estimation of the transcription factors' activity Besides the TF-target gene association, learning transcriptional regulation from expression data relies on a second assumption, that is, the mRNA measurements in the microarray data are predictive of the activity of the actual regulatory molecules. As in the aforementioned correlation analysis, one simply uses the expression level of a transcription factor as the identifier for its activity, which we defined as Type 1 estimation. Since many transcription factors are largely reported as being regulated by post-transcriptional modifications, simply equating mRNA level and protein activity has been criticized [39]. Therefore, a different type of estimation has been suggested

to represent the activity of a transcription factor, which is based on the behavior of its target genes, which we call "Type 2" estimation. Instead of the mRNA level, Type 2 model uses the expression level of the target genes to represent the transcription factor activity. Figure 1.2 on page 25 demonstrates the basic ideas in these different type of models: Type 1 models rely on the expression level of the transcription factors. Type 2 models compare the genes with the transcription factor binding site (target genes) and genes without the binding site (background) and use the differences between expression of the target genes and the background genes to represent the activity of the transcription factor. Type 3 models assume that target genes expression are better correlated if the transcription factor is activated, and use the target gene correlation to represent the activation of the transcription factor.

We apply the Type 2 estimation on the yeast expression data. We use the difference in expression between a TF's target genes and non-target genes as its activity level for a given condition. We then compare the 1-to-1 TF-gene correlations with the results obtained using the Type 1 estimation. As shown in Figure 1.3 on page 27, there is a significant increase in the average correlation coefficient when using the Type 2 estimation. Genes with fewer TF binding sites on their promoters has less uncertainty of their regulatory mechanism, thereby these genes may contribute more to approximating the activity of its regulators. The TF activity inferred from target genes can then be weighted by "1/n" where n is the total number of TFs that are able to control a particular gene. Such weighted summation version of Type-2 model provides better correlation (as high as 0.28, see Figure 1.3 on page 27). The weighting reduces the contribution of genes with many different *cis*-modules on their promoters; presumably they are subjected to combinatorial regulatory effects. Figure 1.4 on page 28 shows that by applying the Type 2 estimation and incorporating context-specific

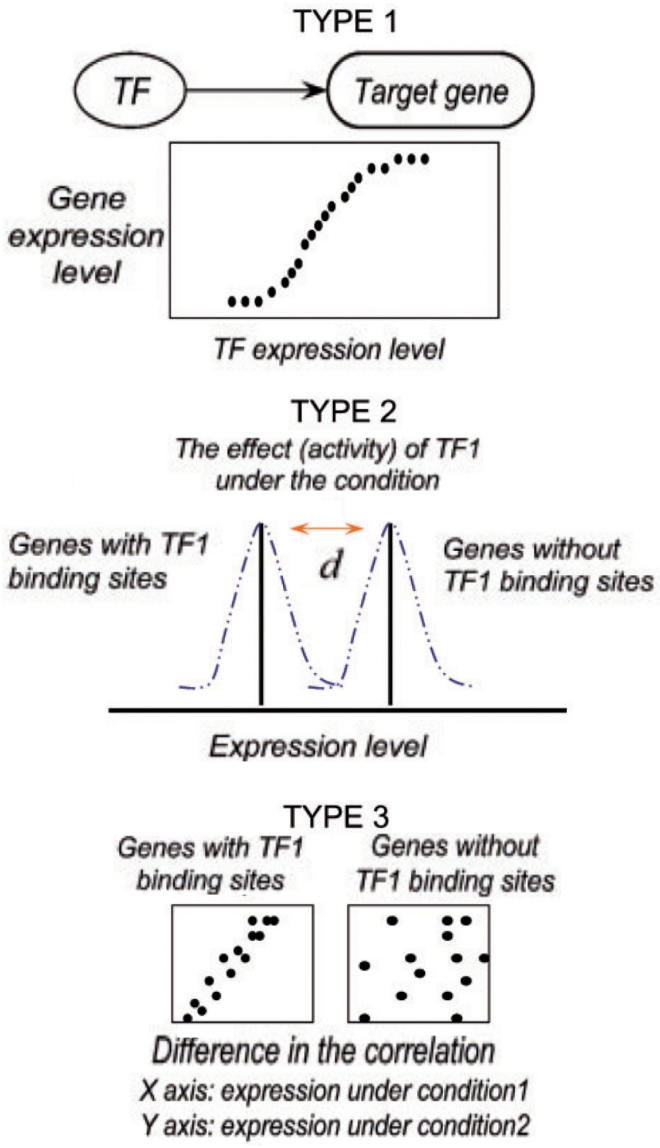

Figure 1.2: Schematic representations of different models to estimate activity of transcription factors.

information one could reach an average TF-gene correlation coefficient of 0.5.

The Type-2 method provides better estimation of TF activity without assuming mRNA level represents TF activity, however such a good character of Type-2, does not benefit much to the network reconstruction. To demonstrate this we simulate a kinetic model $d[Gene]/dt = (1 + [TFactivity]/K) - 1$ (with arbitrarily defined effective kinetic constant K) for the transcription regulation of a single TF on its target gene, shown in the first case of Figure 1.5 on page 30. If the mRNA level would represent TF activity, which may not be the case in many real biological systems, the mechanism by which a gene is regulated by a single TF would be uncovered using Type-1 model (in the single inhibition case, as shown in the responsive curve of the middle plot in Figure 1.5 on page 30). Type-2 model may have better estimation of TF activity but could be difficult to identify the regulatory mechanisms, because the information used to estimate TF activity, in Type 2 model that is the TF-target hypothetical relationship and the target gene expression, and the information required to predict TF-target gene relationship, are overlapped in some way, although the two pieces of information are not all the same—the TF activity is estimated from the overall effect of its potential targets, but a TF-gene relationship is established by testing whether or not the change of the TF activity can explain the change of the expression of the particular gene. In addition, the Type 2 estimation assumes the influence of a transcription factor on all its targets is equal, which does not account for the variability in binding and the function of a transcription factor on its various targets due to adapters and co-factors [53]. Another limitation is that the Type 2 and Type 1 estimations need a priori knowledge about the transcriptional regulatory network, and such information is less often available in other model organisms. Thus, the Type 3 estimation, which assumes co-expression genes

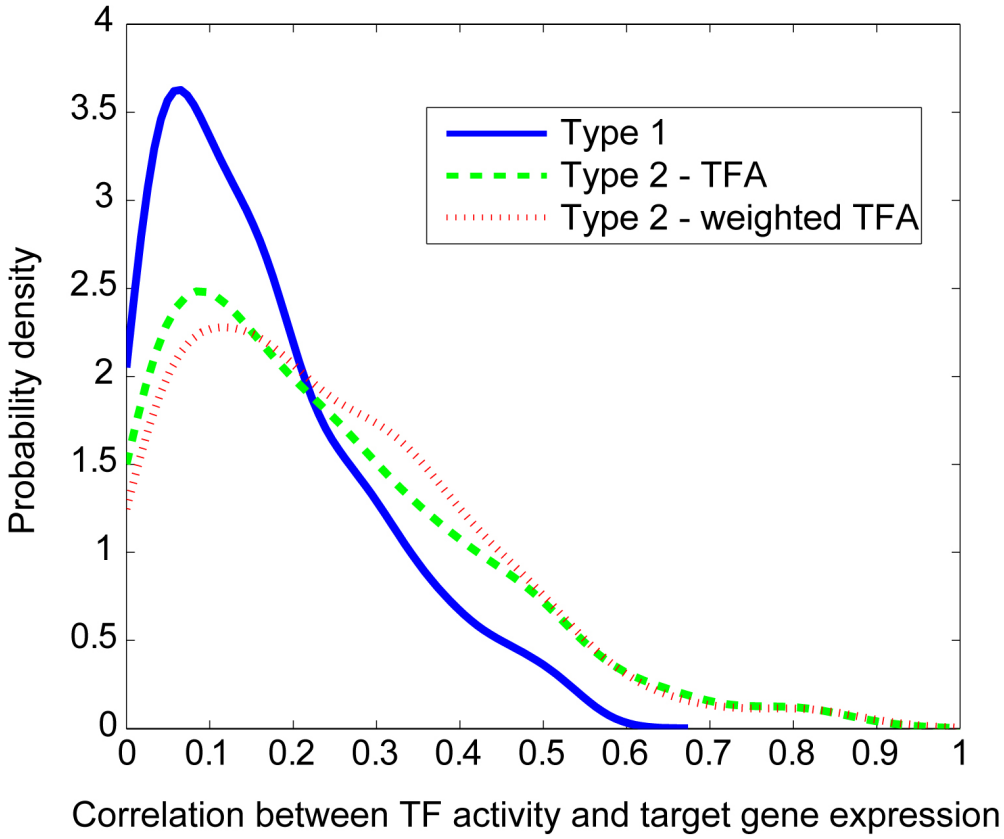

Figure 1.3: Distributions of the correlation achieved by three different TF activity measures. only "1-to-1" cases are considered. Although the average correlation does not improve much in the Type 2 model, there are more genes whose expression level is better correlated with their TFs. Type 1: The TF activity is represented by its expression level; Type 2-TFA: The TF activity is represented by the difference between the mean expression value of its target genes and the mean expression value of the (other) unrelated genes; Type 2-weighted TFA: Genes with fewer TF binding sites on their promoters contribute more to approximating the TF activity, TFA is then weighted by "1/n" where n is the total number of TFs that are able to control a particular gene. The probability density is calculated using the kernel smoothing density estimate function (ksdensity) in Matlab, with a Guassian kernel. A histogram describing the frequencies of gene expression in different categorical bins should show a less continuous but similar distribution.

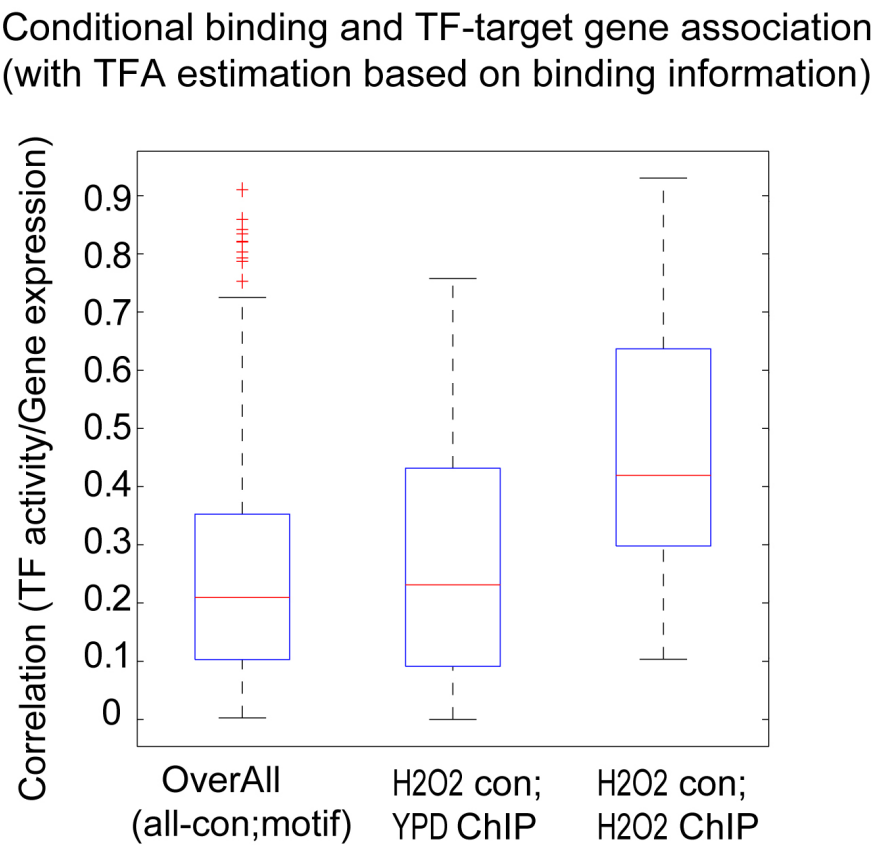

Figure 1.4: Box plots of the average correlation between activity of the transcription factors and the expression of the target genes.596 1-to-1 TF-target gene pairs are considered. The activity of transcription factors are estimated by Type 2 model. CHIP-chip data is not available for every transcription factor, so only 99 pairs have binding information and 20 of them are conditional-specific binding. The overall gene-TF (box on the left) describes the average correlation of all 596 pairs under all conditions. To estimate the activity of the transcription factors, target genes of each transcription factor are determined by yeast transcriptional regulatory network, where the TF-gene interactions are based on motif/binding site. Under the H2O2 condition, the box in the middle (YDP-CHIP) uses the unconditional binding (CHIP-chip results under normal condition) to determine both the activity of transcription factors and the average correlation of the Gene-TF pairs that actually bind, and the box on the right uses conditional binding CHIP-chip data. The average TF-gene correlation increased significantly when conditional binding data is available (box on the right, $p < 0.05$ compared with the box on the left).

are co-regulated by the same transcription factor, has been widely implemented to predict *cis*-motifs or to estimate the activity of *cis*-motifs by co-expression analysis.

With the Type 3 model, one could use clustering analysis first to identify the co-expressed genes, followed by enrichment analysis on the promoters of genes within a same cluster, to identify functional *cis* elements. However, the assumption that co-expression indicates co-regulation is imperfect. Since co-expression may occur in situations other than TF-gene associations [25], such as in a signaling cascade. Thus, one should be cognizant of the assumptions that lead to the results, and further experiments are required to validate the cis-elements uncovered in this manner.

Combinatorial regulation of transcription factors Most of aforementioned analysis focus on the "1-to-1" pairs whereby the genes are likely regulated by only one transcription factor, nevertheless there are many more genes ($> 80\%$) with multiple promoter regions that bind different transcription factors, thereby complicating the regulatory mechanism through TF cooperation. Models of TF binding and gene transcription have been extensively studied experimentally on prokaryotes with small-scale quantitative measurements of numerous perturbations on a subset of the regulatory circuits [54]. A detailed thermodynamic binding and control model has been established [55] and successfully applied to many, highly specific regulatory models in *E.coli* and Drosophila [31, 56], providing a quantitative framework for studying the combinatorial regulation of transcription factors. Researchers are now attempting to generalize the dynamic model to automate the procedure of learning the detailed mechanisms from high-throughput data. Questions have arisen on whether or not the information in large-scale expression data is adequate to support these detailed mechanistic models. Since current knowledge of combinatorial regulation in real eukaryotic

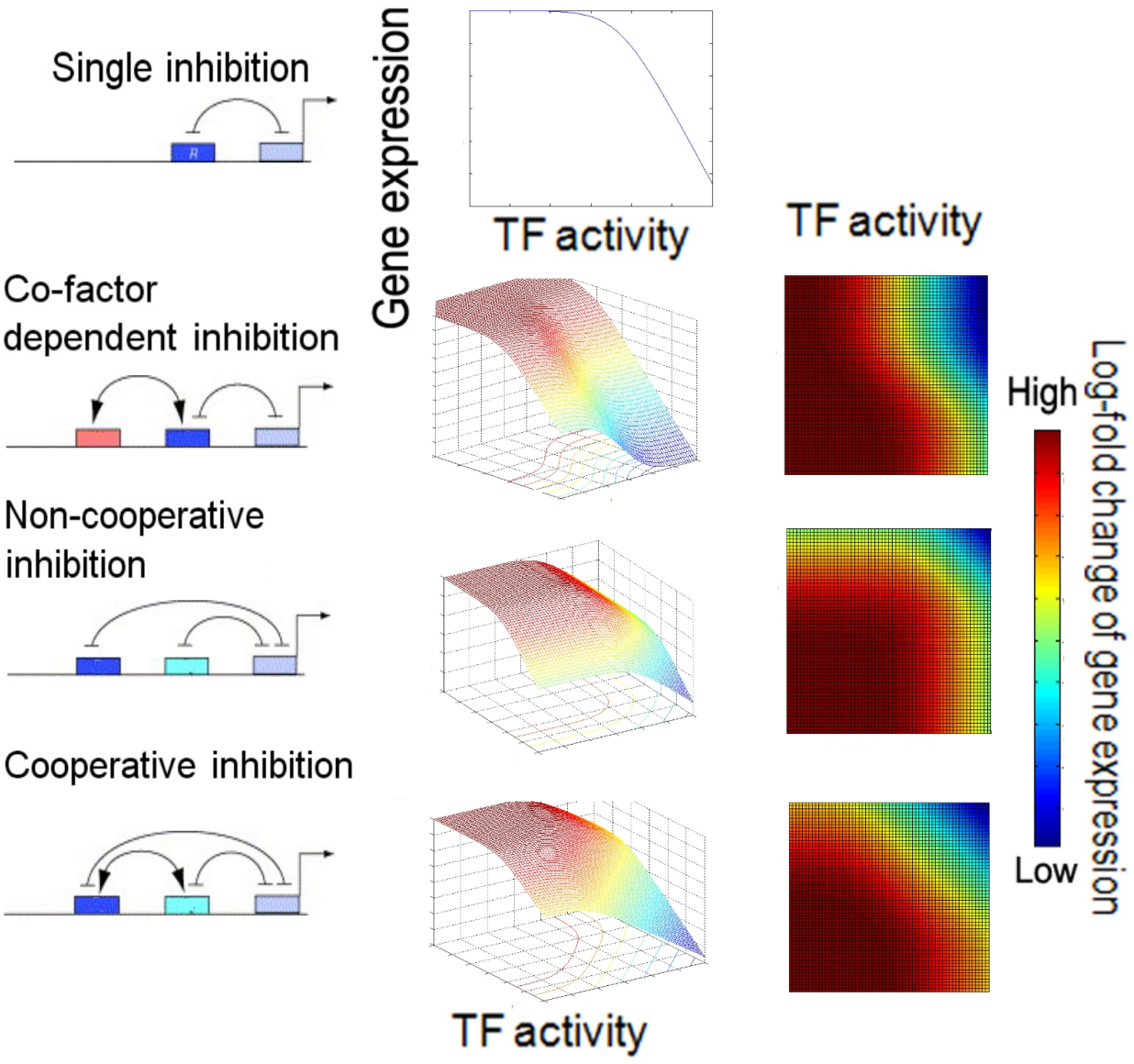

Figure 1.5: Simulations of the combinatorial transcription regulation. The interactions between transcription factors and between transcription factor and the initiation of the target gene transcription are modeled with kinetic equations. Response curves are simulated and plotted. For the regulation by single transcription factor, the x-axis is the activity of the transcription factor and the y-axis is the expression level of the target gene. Both 3-D plots and 2-D color maps are provided for the combinatorial regulation of two transcription factors, where the x- and y-axis represent the activity of the two transcription factors and the z-axis/color represents the expression level of the target gene. Microarray data does not directly provide activities of transcription factors, so the rightmost plots for each regulatory scenario instead uses estimated activity (Type-2 model) of the transcription factors, showing what profiles that could be obtained from microarray analysis. Numerical simulation of kinetic model is performed using the Runge Kutta method in Matlab and the plots are generated with customized code in Matlab.

cell systems is still very limited, it is hard to assess predictions of combinatorial regulation. We try to use simulation study to show some cases of combinatorial control as well as to what extent different mechanisms can be revealed by analyzing expression data. To address the combinatorial control of two transcription factors we explore the mechanistic cooperation schemes and acquire kinetic equations from previous studies of theoretical modeling (thermodynamic binding model) [55], then generate putative gene expression profiles. The simulations show different expression profiles depending on the cooperation mechanisms (Figure 1.5 on page 30). These profiles demonstrate that the differences in the different cooperation schemes are so subtle that only a few specific perturbations/conditions would capture the distinctive features, as depicted by the narrow transition (regions with significant changes of colors representing target gene expression in the 2-D response surface color-maps) in the response surface curves in Figure 1.5 on page 30.

In Figure 1.6 on page 33 we show the clearest pattern that we have seen in the yeast data with interpolation on the combinatorial regulation of two TFs on one gene. Compared with Figure 1.5 on page 30, we show that without independent measurements of the actual activity of the transcription factors, as well as enough number of perturbations to cover all possible combinations of the two TF's activity level, the subtle features in the different cooperation schemes are not sufficiently distinctive, and will be further exacerbated by the noise and limited array data measurements. Therefore, the "top-down" approaches have serious limitations in their ability to learn these detailed mechanisms. Unlike the "bottom-up" quantitative experiments performed for small systems, the high-throughput data involves many layers of interconnected regulations making it difficult to segregate the contributions of each TF on the expression of its target genes. As observed by Gitter et al. [57] in the

systematic knock-out experiments [58], the overwhelming majority of its target genes would not be affected even if a transcription factor is knocked out.

In this section, we introduce briefly the current approaches to identify genes and interactions that are specific to a phenotype. To identify genes, a problem is the selection of genes that can distinguish a phenotype, thus many filter methods in feature selection can be applied to address the problem. To identify interactions, or "reconstruct a network", the general idea is to determine whether or not a gene controls the expression of another. From a biological perspective, we used a set of yeast microarray data as a working example to evaluate the fundamental assumptions implicit in learning the transcriptional regulation from gene expression data. We show that the detailed transcriptional mechanism is overly complex for expression data in the conventional setting (samples of phenotypes and controls) alone to reveal. A proper incorporation of multi-source biological knowledge, especially context-specific information, is beneficial for network reconstruction. The idea has been implemented by integrating PPI (e.g. MATISSE [59]) or P-DNA data (e.g. MINDy, MARINA [19]) to provide a "reference" biological network to identify the interactions wherein are specifically functioning in a phenotype. Furthermore, researchers have attempted to combine the systematic design of high-throughput experiments with these computational approaches to achieve better reconstruction of regulatory networks. A theoretic simulation study by Yu et al. [20] shows that Dynamic Bayesian Network is both accurate and efficient in recovering the simulated molecular network when there is a large time-series dataset with many (25-100) time points (although it is still difficult to recover regulatory relationships when a gene have multiple regulators, as shown by our yeast analysis). Sachs et al. [60]performed systematic perturbation experiments on a small network of protein signaling, with

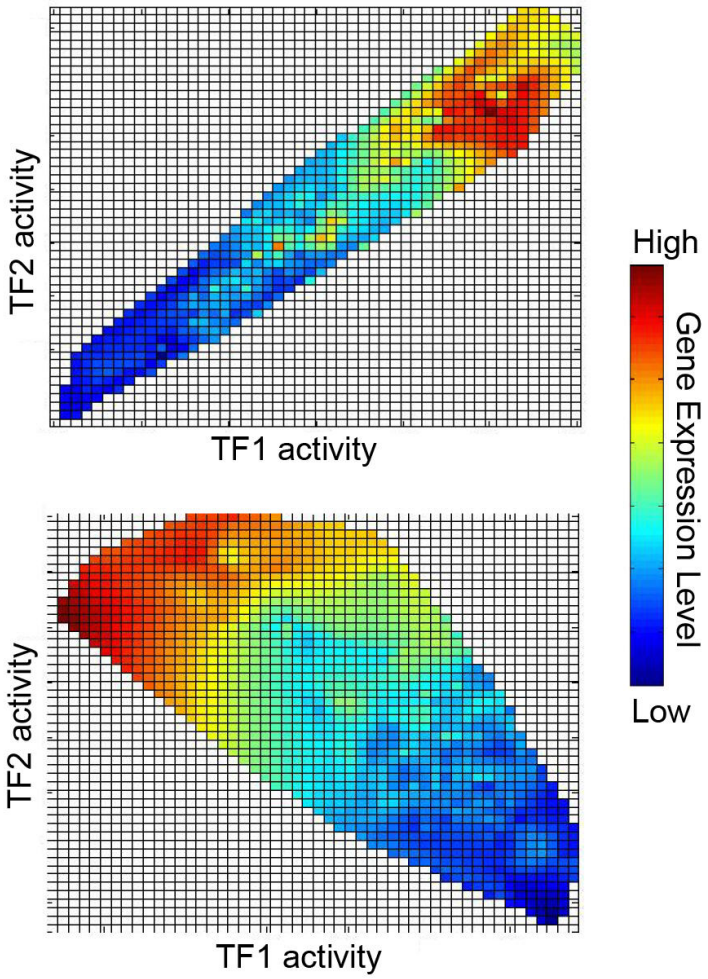

Figure 1.6: Examples of combinatorial transcription regulation in gene expression data. In these two examples, the genes have only two potential TF regulators based on TRN. The space is interpolated based on the 255 datapoints.

flow cytometry measurements of stimulatory/inhibitory interventional conditions on most of the proteins in the network. Based on these perturbation data, they applied Bayesian Network and successfully reconstruct a detailed regulatory model for the RAS/MAPK signaling in human primary immune cells. Therefore, combining systematic design of high-throughput experiments with machine learning approaches can identify the causal/regulatory network underlying a particular biological process.

Overall we suggest a conscientious inspection of both the biological assumptions underlying the mathematical formulations of the models, and the information contents in the data in support of the statistical learning processes, which we believe is required in order to achieve learning results with lasting biological significance. This would help accelerate fruitful capitalization and continuation of computation in promoting our understanding of the biological regulatory system. Unless coupled with interaction data, systematic perturbations or time-series experiments, the "top-down" approaches have limitations in their ability to learn detailed mechanisms and causal relationships. Thus in the next section I will introduce the "bottom up" approach in systems biology that builds models based on the knowledge of the regulatory mechanism of a gene network to provide specific hypothesis on why and how a gene is involved in determining a phenotype.

1.3 Modeling based on mechanism

Systems biology integrates experimental and modeling approaches to develop a detailed dynamic understanding of the functions and behaviors that are specific to a biological system. The approach of systems modeling is different from mining approaches that are attempting to identify and characterize genes and interactions to distinguish phenotypes based on omics

data. The modeling approaches combines systems theory and biological information to achieve a functional understanding of the system as dynamic processes [61].

The modeling study in systems biology integrates experimental results to develop a functional context to address the following complexity that arises in biological systems as dynamic interacting networks:

- The large number of different molecular components in cells.
- The enormous, heterogenous, direct or indirect interactions between the molecules.
- The complex functionality of molecular components. For example, the same component could have very different function in different biological processes (multi-functional) or organisms. A particular function can be accomplished by different sets of components (redundant).
- The dynamic behavior of biological systems. In many cases the dynamics are seldom linear, and contain feedback circuits that are very complex. The diversity of biological systems.

1.3.1 Modeling of gene network: An overview of systems theory

Modeling is a process which associates mathematical concepts with natural system(Wolkenhauer, 2001). Particularly, in gene regulation or cell signaling, one aims to connect the interaction of the molecular components and the observable biological phenomenon with mathematical variables and relationships so as to generate predictions and hypothesis. The basic modeling approach can be categorized broadly into deterministic kinetic modeling, stochastic modeling and qualitative discrete dynamic modeling. These are mathematical tools that can integrate

network information and formalize the causal structure of the signaling system with variables and functions [62].

Simulation is a modeling experiment that can provide a temporal profile representing the dynamics of the system. For example, in a kinetic model, a simulation may apply numerical integration to find a solution of the differential equations. In a stochastic model, a simulation is a single realization of the stochastic process in which the probability of the resultant state of each variable corresponds to the distribution constrained by the stochastic equations.

The basic framework of modeling, which could be used to generalize many different models in systems biology, can be established with systems theory. The following derivations in systems theory can be found in the review [63] . In systems theory, a *system* is defined by a set of objects and their relationships:

$$S \subset O_1 \times O2 \times \ldots \tag{1.9}$$

where the O_i are objects and $\times$ denotes the Cartesian product (a combination of objects in the object set) that represents the relationship between objects. In a signaling system where one observes that some stimulations cause some responses could be described as

$$S \subseteq \Omega \times \Gamma \tag{1.10}$$

where the Ω and Γ represents the stimulus and response, respectively. For example, the stimulus could be a set of different ligands and the responses are the expression of specific genes. There should be causal relationships between stimuli and responses, which can be

mapped:

$$\begin{aligned} \sigma :&\Omega \to \Gamma \\ &\omega \to \gamma \end{aligned} \tag{1.11}$$

where $\Omega = \{\omega : I \to U\}$ and $\Gamma = \{\gamma : I \to Y\}$ defines each stimulus/response a temporal sequence of events, that is, at any time $t \in I$, a stimulus $u(t)$ of the system S results a $y(t)$ at time t. In practice, a time-course plot of the stimulation and response can visualize one $\omega \to \gamma$ mapping.

Thus we now have a "phenomenological" model from the stimulation to the response, in the sense that we treat the internal system as a black-box and only consider the input and output. There are many phenomenological models that have been applied to describe physiological changes, such as different bacterial growth curve under different conditions involving temperature, pH, nutrition and other factors in the medium. Unsatisfied with a phenomenological description, systems biology usually looks for deeper *mechanistic* models, which consider the transduction process within the cell at the molecular scale to reveal the mechanistic design of the cell system. Thus the model should be extended with a state-space X: $\Omega \to X, X \to \Gamma$. The behavior of the system is encoded by the states $x \in X$ and the temporal evolution of $x(t) = \varphi(t; t_0, x, \omega)$ depends on the state-transition mapping:

$$\varphi : I \times I \times X \times \Omega \to X \tag{1.12}$$

where the state of the system at a particular time depends on the initial state $x(t_0)$, the stimulation Ω , and probably the states at some other time.

A dynamic system S is 'continuous-time' if I is a set of real numbers: $I = \mathcal{R}_+$. If I are

integers: $I = \mathcal{Z}_+$, the system is a 'discrete-time' system. A dynamic model that determines precise values $x(t)$ are 'deterministic' and if the $x(t)$ are random variables the model is 'stochastic'. S is finite dimensional if X is a finite-dimentisional space (which means $x(t)$ are discrete variables with finite states). The $X \to \Gamma$ mapping could be used to model the noise in the measurement of the system.

Therefore:

- The kinetic modeling approach which applies a set of differential equations to describe the concentration changes of molecules essentially defines a deterministic continuous-time system S, where the mapping φ is represented as equations:

$$dx/dt = f(x(t), u(t)) \tag{1.13}$$

- The Boolean Network modeling, is a method of discrete-dynamic modeling, defines a discrete-time finite-dimensional system where the variables $x(t)$ has finite states (e.g. two states on/off only) and the mapping φ includes some state-transition rules. The state of the system in the next time point is determined by the previous ones through logic relationships between the variables.

$$x(t+1) = F(x(t)) + G(u(t)) \tag{1.14}$$

  in which F represents the state-transition rules and the G defines the possible effect of the stimulation.

- The stochastic modeling approach defines a stochastic discrete-time finite-dimensional system where the state change is discrete-time (similar to equation.6) but variables

represent random processes.

The mathematical concepts and techniques of modeling and simulation in systems dynamic theory are generic [64]. The systems approaches to manage complexity arising from dynamic interactions are well developed, the general framework of modeling, different type of models and the simulation methods are not novel and have already been successfully applied to a wide range of processes in other areas, such as engineering, physics and chemistry. For the systems modeling of gene networks, the major challenge is to build a model that is consistent with our understandings of the biological mechanisms. In the next section I will use one of my studies as an example to show how a gene network model can be developed based on biological assumptions.

1.3.2 Example: discrete dynamic modeling of insulin signaling in liver cells

The approach to build a systems model involving genes and interactions that determines a phenotype can be decomposed to address the following two questions:

- How to model a phenotype?
- How to define the perturbations that determines or changes a phenotype?

To model a phenotype , we need to collect information how the phenotype is regulated. As an example, the insulin signaling is a well-studied and complicated signaling network in mammalian cells. It is composed of branched downstream signaling pathways and various feedback mechanisms, which could benefit from modeling. In a separate study, our group identified the involvement of a novel player, PKR, in the insulin signaling network of

HepG2 cells [65]. Research in our group has shown that as one of the downstream target of insulin signaling, PKR is intricately involved in regulating the insulin signaling process through a feedback mechanism [65]. Therefore, we model the phenotype of insulin signaling in liver cells as a dynamic signaling process involving PKR, in which potential interactions and components are collated from literature and the patterns of their dynamic activation under insulin treatment defines a "phenotype". Since the literature information and our experimental approaches provides mostly qualitative information, we choose to use discrete variables to represent the activation of components (the signaling molecules involved) and specify the state transition rules to describe interactions between the components that determine a phenotype.

The signaling network can be formalized in terms of an oriented graph, where the vertices represent the elementary components involved in the process and the arcs describe the regulatory interactions between those components. In the signaling network, each directed arc reflects the direction of information flow from the source vertex to its target in the signal transduction, and is labeled with a positive or negative sign which defines activation or inhibition, respectively.

We associate each vertex in the signaling network with a discrete variable, which has three states representing the activity of the protein —- 0: lower than control, 1: the control state, 2: higher than control. Thus by definition every component starts at the "control" state in the absence of insulin stimulation.

We define transition rules based on the activation/inhibition attributes on the arcs in the signaling network. Two operations: shift up $(+1)$ and shift down (-1) adapted from the triple logic are applied in the model. If an activator is in a state "higher than control", (e.g.

the kinase phosphorylation is increased), the state of its target will be shifted upwards. In contrast, if the state of an inhibitor is higher than control, its target will be shifted downwards in the next updating event. For some components, there may be multiple regulators that are active in one updating event and the combinatory effect is determined by comparing the number of activating to inhibitory factors. The target is shift-up if there are more activating factors, and vice versa [66, 67]. If the number of activating and inhibiting factors equal, we assume the target remains at the control or current state. Finally, the state of a component will decay if its regulators can no longer maintained their active state. This way we construct a systems model for the insulin signaling in liver cells, whereby we can simulate the evolving of the activity of the signaling proteins along with time to define a phenotype to be investigated.

To define the perturbations that determine or change a phenotype , which is, in this case the dynamics of the insulin signaling in liver cells, constraints are assigned to components to mimic the perturbations on the network. If a protein is constantly inhibited (in an experiment), we restrain the state of its corresponding variable in the model to be always in either 0 (lower than control) or 1 (control state).

In the simulation each run starts with its own set of randomly generated initial states and a simulation result represents the dynamic profile of a single cell in the population. By assuming that cells response independently to a signal, we can simulate a large number of independent runs to mimic a population effect and measure the average evolving profile for the population.

Time is modeled by regular intervals called time-steps. Since most components in our network are kinases or phosphotases, and most reactions are protein phosphorylation and

dephosphorylation, we assume that the duration of the activation/inhibitions and the decay processes in the signaling transduction are comparable and approximated by one time-step. Since the reaction rates may be different from cell to cell even for the same interaction, we apply asynchronous updating of the state: $S_i^n = f_i(S_{(}j_1)^{(}m_1), S_{(}j_2)^{(}m_2), \ldots, S_{(}j_k)^{(}m_l))$, $m_l \in n-1, n$ where S_i^n is the state of component i at time-step n, and f_i is the transition function associated with i and its regulators j_1 to j_k, and the time-point corresponding to the last change of the regulators can be either the last or current round of updates.

This is an example of how biological assumptions are applied to build a systems model of gene network, more details of how this particular signaling network is analyzed by modeling and experimentation can be found in our paper [68].Overall, this discrete dynamic model provides an *in silico* model framework that integrates potential interactions and assesses the contributions of the various interactions on the dynamic behavior of the signaling network. Simulations with the model generated testable hypothesis on the response of the network upon perturbation, which were experimentally evaluated to identify the pathways that function in our particular liver cell system. The modeling in combination with the experimental results enhanced our understanding of the insulin signaling dynamics and aided in generating a context-specific signaling network.

More importantly, by comparing model simulation with experiments, we found that even in this well-studied signaling system, there are many components and interactions that are not actually functioning, or not functioning in the way we thought they would be in our system, since our current understandings of regulatory network are based on experiments from different research groups, on different systems, under different conditions. For example, we found the AKT-PP2A feedback in the insulin signaling, in which PP2A can induce the

dephosphorylation of Akt, and thereby suppress the activity of Akt, is not functioning in the HepG2 liver cells, although it has been suggested as one important regulatory module in insulin signaling. These observations emphasize the importance and necessity in the identification of specific genes and their regulation to understand a phenotype.

In summary, systems modeling studies biological processes by systematically model the gene network and its response to different perturbations. The modeling studies provide a framework that can integrate experimental results to develop a functional context to address the complexity that arises in biological systems as dynamic interacting networks. The methods for modeling and simulation are generic but their biological applications have many challenges, since many biological assumptions need to be considered to describe an appropriate mathematical representation for the system of interest.

Chapter 2

Identification of condition specific genes and interactions — a multi-layer approach

An important topic in systems biology is the reverse engineering of regulatory mechanisms through reconstruction of context-dependent gene networks. A major challenge is to identify the genes and the regulations specific to a condition or phenotype of interest, given that regulatory processes are highly connected such that a specific response is typically accompanied numerous collateral effects.

In this chapter I will introduce and discuss the multi-layer approach that is able to reconstruct condition-specific genes and their regulation through an integrative analysis of large scale information of gene expression, protein interaction and transcriptional regulation. Application on the yeast dataset correctly identified a context-specific network and the major transcription factors that regulated at either the transcriptional or post-transcriptional level. Current approaches have difficulty specifying these regulators. Further application on human breast cancer identified Trop2 (TACSTD2) as a target gene, and discovered its regulation by transcription factors CREB as well as NFkB, the latter regulated at the post-transcriptional

level. The predictions were further confirmed through experimental studies.

2.1 The problem in current approaches to identify specific genes and interactions

The accumulation of high-throughput transcriptome data has driven the development and application of computational approaches to infer networks, to elucidate gene regulation and to identify targets. As introduced in chapter 1, there are many approaches developed and applied to identify genes and interactions. Initial network inference methods based on gene expression data were successful with prokaryotes [18]. However higher eukaryotes systems, with their higher number of genes, provided many more candidate genes (several hundred) and interactions (more than a thousand) (e.g. [69,70]. This generated numerous hypotheses, and with this sheer number most of the candidates cannot be investigated or validated through experiments, making it difficult to assess the utility of the proposed approach on these systems.

Further, a majority of the candidates are often related to general processes that are not specifically responsive to the condition being investigated. With this large number of possible candidates literature search to "manually" characterize many of these predicted candidates are typically performed to identify potential targets for further experimental investigation based on one's expertise. Alternatively, GSEA [71] and GO (`http://www.geneontology.org/`) annotations have been applied in many reverse engineering studies to interpret the results, which provide an understanding of the "general" processes involved rather than the direct molecular mechanisms.

To address the problem of large number of candidates that are hard to validate and characterize to answer biological questions, several studies integrated gene expression data with interaction networks constructed from protein-protein interaction (PPI) and protein-DNA (P-DNA) interaction information [59, 72]. In these studies statistical methods were applied on the gene expression data of the pair-wise interactions to identify active sub-networks or modules. In contrast, ARANCE based on gene expression data alone was able to successfully identify a transcriptional molecular interaction network by removing indirect interactions [19]. These approaches are based on structure learning methods that aim to infer "functional interactions" in accordance with certain presumed mathematical definitions, i.e. differential expression, correlation or mutual information between the gene expression. A correlation between two genes could suggest "regulating", "being regulated" or indirect relationship. However an absence of a correlation does not preclude a possible regulatory relationship, e.g. post-transcriptionally regulated interactions. Thus subsequent methods were developed to uncover these potential post-transcriptional interactions, given either a transcription factor and target gene pair along with a list of potential regulators or a transcription factor and its target genes [73, 74].

We note that the problem of generating too many hypotheses is because the predictions are not specific enough. For example, in biological systems a specific and direct response to a perturbation usually is accompanied by a cascade of collateral effects on many genes and dozens of regulatory modules in the network. For example, Stephanopoulos et al. [69] showed that in an experiment that knocked-out the GAL80 gene in yeast, and comparing the transcriptomes before and after the treatment showed that such modulation of a single pathway eventually caused a global effect throughout the whole bio-molecular inter-

action network. The specific response of the knock-out experiment is the activation of the galactose-processing pathways by eliminating GAL80's repression on the GAL4 transcription factor, nevertheless the repression of this one pathway resulted in hundreds of differentially expressed genes and dozens of activated modules, making it difficult to identify the essential "trigger", i.e. the specific pathway in response to the perturbation. Therefore, in a typical network inference or module analysis, many of the genes and modules identified would likely be such "side-effects" or collateral response rather than direct and specific effects. For example (Xia Yang et al., 2010) in identifying the activity of the p450 gene in human liver, network analysis (clustering and network reconstruction based on correlation) identified more than 5000 differentially expressed traits spanning many general functional modules including immune response, cell cycle, lipid metabolism, macromolecule biosynthesis, etc. This provides a "rough measurement" of the overall influence but is ineffective in identifying the specific pathway that regulates these effects and in guiding the experimental design for further in-depth functional studies. Therefore, the specific responses are usually concealed by many less specific effects, which are difficult to distinguished based on current network analysis methods.

2.2 Basic ideas to improve the specificity of prediction

In our network reconstruction framework, we propose to integrate microarray data from a diverse set of conditions to provide a common context (more and better controls) for the expression behaviors of genes, and apply advanced feature selection technique, to identify the target genes (in layer I) that are most specific to the condition being investigated. From the target genes, conditional gene regulations (in layer II), and conditional transcription

factor activity (in layer III) are then determined. Incorporating these diverse conditions for comparison in the feature selection of genes and interactions reduces the false positive rate and enhances the specificity.

2.2.1 Incorporating diverse conditions for the identification of genes

Identification of conditional specific genes or ranking of the changes in TF activity can be modeled as a process of "feature selection", as introduced in Chapter 1, which is to select the best features (genes, TFs) that can determine a class (phenotype/condition). Traditional methods to identify candidate genes typically compare two set of samples, e.g. treated and untreated condition.

In contrast to traditional methods, we propose to integrating multiple conditions of gene expression data and applying a context sensitive algorithm to identify the genes that specifically respond to the condition being investigated. Therefore, the computational approach that we are looking for, which is to identify features (genes and TFs) that can distinguish one phenotype from all the other phenotypes, should fulfill the following requirements from the biological aspect:

1. The approach aims to weight and rank genes according to their "importance".

2. The approach should account for the fact that features (genes) are not all independent. Gene expression is controlled by a complex regulatory network, thus there are intrinsic relationships between genes.

3. There are intriguing relationships between phenotypes. There are some phenotypes that may have transcriptomes similar to our phenotype of interest, these phenotypes

could be more important to compare with to understand the unique changes in our conditions.

Therefore, the approach should adopt a learning model that fulfills these requirements:

1. It should be a feature selection process to weight genes/features: i.e. it is a mapping from original feature space $\mathcal{X}$ to a new feature space $\mathcal{X}'$, by scaling each dimension with weight w. There should be some constraints on ws so as to ensure a unique solution.

$$f : x \to wx, \;\; \|w\|_2^2 = c, \; w \leq 0 \tag{2.1}$$

2. To account for the fact that features (genes) are not all independent, we should evaluate each feature in the context of other features and samples, which suggests a *local learning* model.

3. There are some phenotypes that may have transcriptomes similar to our phenotype of interest. Thus we try to find "nearest neighbors" for the samples by comparing transcriptomes. For a sample x_n, its nearest neighbor of the same phenotype is Hx_n, while its nearest neighbor from a different phenotype is Mx_n. We can define a "margin":

$$\rho_n = d(x_n, Mx_n) - d(x_n, Hx_n) \tag{2.2}$$

We try to find a mapping so that the distance between different phenotypes $d(x_n, Mx_n)$ are as large as possible, while samples from our phenotype $d(x_n, Hx_n)$should be as close

as possible. So the problem can be formularized as maximize the margin:

$$\max \sum_{n} \rho_n(w), \text{ with respect to } w, \text{ s.t. } \|w\|_2^2 = c,\ w \leq 0 \tag{2.3}$$

Based on our modeling of this learning problem, there is an algorithm family called "Relief" that can solve this optimization problem, which we applied in the identification of genes and transcription factors for a phenotype.

The ReliefF algorithm The basic procedure of Relief algorithm (Kira and Rendell, 1992) is shown as follows:

> Algorithm: Relief (theoretical computing time complexity $O(m * n * f)$)
> Input: n learning samples X, with f features, sampling parameter m
> Output: for each feature F_i a quality weight W_i
> Initiate: For $i = 1$ to N: $W_i = 0$
> For $l = 1$ to m:
> Randomly pick a sample x_k;
> find its nearest hit H and nearest miss M;
> For $i = 1$ to f: $W_i = W_i - dif(i, x_k, H)/m + dif(i, x_k, M)/m$
> Return W

Relief estimates the quality of features through a nearest neighbor comparison to account for the "local" context of features (gene expressions in our case). The algorithm selects neighbors from the same condition (hit) and different condition (miss) based on feature vectors (the transcriptome of a sample). The function $dif(i, S_1, S_2)$ calculates the difference between the values of feature i in two samples I_1 and I_2. m is the number of samples randomly sampled from a dataset of n total samples and f features. For each iteration, the weight of features are updated with respect to whether the feature differentiates two samples from the same condition (undesired property), and whether it differentiates samples from different condition (desired property).

When Relief is used to identify condition specific genes, the data matrix E (n genes $\times m$ samples) contains n feature vectors, and the feature vector for each gene is the expression of the gene in different samples. When Relief is used to identify condition specific TF activity change, the data matrix A (k TFs $\times m$ samples) contains k feature vectors, the feature vector for each TF is the summation of the expression of its target genes in different samples, i.e. $A = B \times E$, in which B (k TFs $\times n$ genes) is the TF-Gene binding matrix based on protein-DNA binding information.

2.2.2 Integrating multiple conditions to identify regulatory relationships

Traditional approaches to determine gene regulatory relationship are commonly based on mutual information (or correlation) between pairs of genes [19, 72], with comparison between a condition of interest and a reference condition. Instead of using a single "untreated" reference condition, we suggest incorporating multiple conditions to provide a better reference pool, and compute the difference between the conditional and unconditional mutual information (MI):

$$\text{MI}(\text{gene pair}|\text{condition}) - \text{MI}(\text{gene pair in all conditions}) \tag{2.4}$$

This idea is essentially an extension of the two-way relationship (gene x and gene y) to a three way dependencies (gene-pair x_1, x_2 and y) in calculating the mutual information from gene expression, such that the regulators of y, genes x_1 and x_2, could be readily distinguished. This approach, also called causal-filtering, was initially suggested by Bontempi et

al. [75], in which they compute the conditional mutual information between x_1 and x_2 given y, $MI(x_1, x_2|y)$, and the unconditional one $MI(x_1, x_2)$. When x_1 and x_2 are the regulators of y, one have $MI(x_1, x_2|y) - MI(x_1, x_2) > 0$. In our case, since the effectors/target genes y for the condition of interest is identified by Relief and actually determines the phenotype, thereby the three way dependencies (gene xs and y) are similar to conditional dependencies (gene xs and the condition of interest), thus we compute the differences between conditional mutual information and unconditional mutual information for genes on paths to the target genes in the reference network (PPI and P-DNA network), and each gene pair consists a potential "regulator" and a "target gene" in the network. Positive scores suggest a potential causal relationship while zero scores suggest an indirect or downstream effect and thus are removed from the network. Any node with zero score is filtered out unless it is a potential transcription factor of the targets or it serves as an intermediate step on the path of a causal (positive-scored) node to the targets.

2.3 The multi-layer approach and the proof-of-concept applications

We take an alternative approach to identifying targets that are specifically responsible for the condition of interest, and which may not necessary be hubs in a network [76]. Our approach separates the network reconstruction into multiple levels, each addressing a specific biological question wherein a particular scenario of biological regulation is modeled. The overall aim is to provide more specific regulatory hypotheses by incorporating interaction data, however in a very different fashion from prior approaches. In contrast to prior approaches [59, 72, 77]

wherein an interaction network is first constructed and the gene expression is then superimposed onto the network, we first identify the most specific genes involved and then build the interactions from there up, with the interactions obtained from PPI and P-DNA information. We separate the reconstruction of condition specific gene network into three layers (layer I, II and III). Layer I aims to identify the genes that have distinct expression pattern under the condition of interest, from which the conditional network is built. Once these "specific" genes are identified, the network is expanded in layer II based upon known and predicted interactions with these genes, obtained from the PPI and protein-DNA networks. A filtering approach based on mutual information is applied to the physical interaction network to reconstruct the regulatory pathway from the candidate genes. In layer III we infer the transcription factor (TF) activity to identify the major regulators in the gene network, accounting for post-transcriptional regulation. The multiple layers of learning with their distinct biological assumptions capture different biological features in the regulation to achieve reconstruction of condition-specific gene networks that accounts for both transcriptional and post-translational molecular interactions. We establish the accuracy of our methodology against synthetic datasets as well as a yeast dataset.

2.3.1 Layer I. Identification of candidate genes

We propose that integrating multiple conditions of gene expression data and applying a context sensitive algorithm Relief identifies the genes that specifically respond to the condition being investigated. Unlike previous applications of pair-wise comparison with Releif [78], we apply Relief to a diverse set of conditions. To model this, we generate a simulated dataset of 100 genes with 5 different conditions (conditions $A-E$) plus a reference condition (condition

ref), each condition has 10 samples. In each of these 5 conditions there are 10% genes that are uniquely changed (specific responses) and another 20% are shared changes (overlapping responses to represent collateral responses). The changes are simulated with a normal distribution of the mean and variance, different from the unchanged (control) condition. The goal of the gene selection/feature selection is to identify the 10 genes that change uniquely for a given condition (condition A). Different methods are applied to identify the 10% of the genes that change uniquely for a given condition. Traditional approach to identify differentially expressed genes compares condition A with the reference condition, while we suggest applying ReliefF on condition A against all the other conditions that are available. We compare two different approaches: traditional approach to identify differentially expressed genes based on t-test, and the ReliefF algorithm; under two different scenarios: with only condition A and reference control, and integrating all the conditions that are available and comparing against condition A. The ROC curves in Figure 2.1 on page 55 shows that Relief applied on multiple conditions performs the best.

The "MEGA" yeast microarray dataset from the Audrey Gasch lab (`http://gasch.genetics.wisc.edu/datasets.html`), which combines 500 yeast experiments from a variety of conditions, provides an integrated gene expression "reference pool" for analyzing yeast data. We evaluate our approach by reconstructing a gene network for the "AltCarb" condition, i.e. adding extra alternative carbon source, in which the essential regulatory pathway is known a priori- namely the GAL pathway. When adding extra carbon sources into the yeast media, if the carbon source is galactose, the GAL pathway, also known as the "Leloir pathway" will turn ON, if it is other types of carbon source, such as glucose, the Leloir pathway is turned OFF leading to "glucose repression" [79]. A successful gene selection procedure

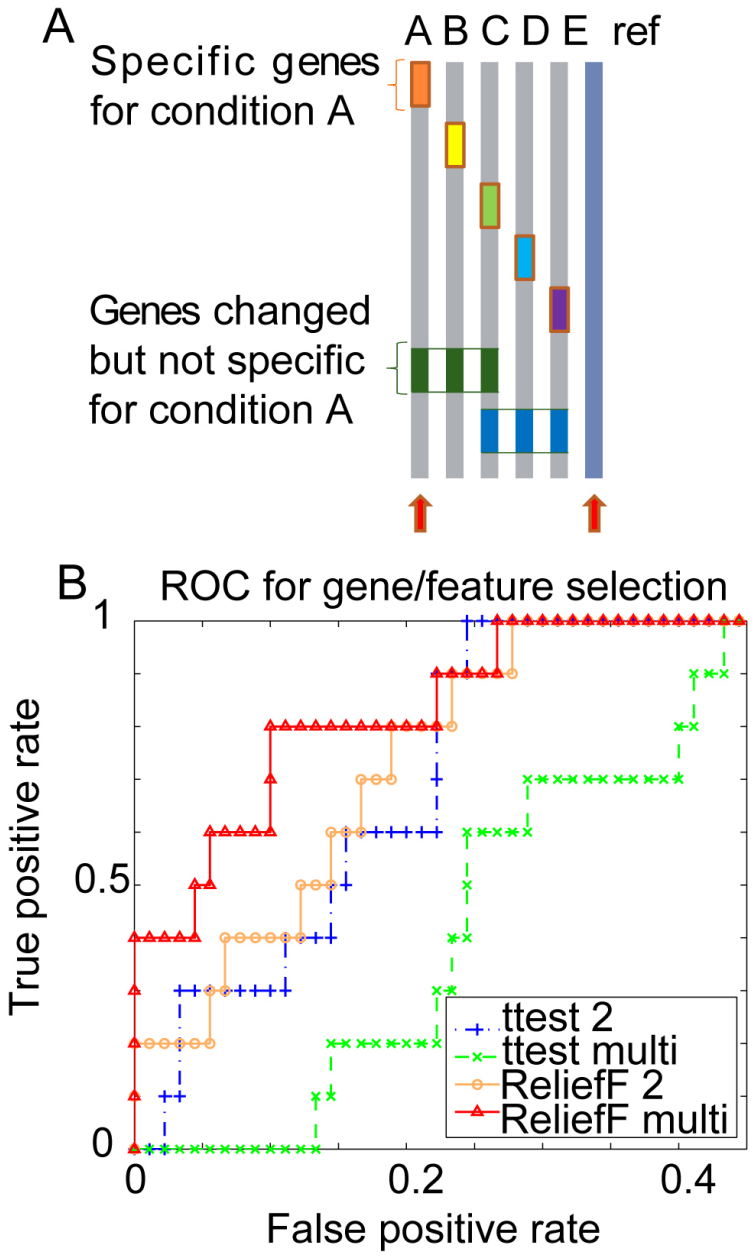

Figure 2.1: Proof-of-concept example of applying Relief on multiple conditions **A:** We suggest applying ReliefF on condition A against all the other conditions. **B:** The ROC curves for identification of specific genes for the condition of interest (condition A). The AUC (area under curve): ReliefF with multiple conditions > ReliefF with 2 conditions > t-test with 2 conditions > t-test with multiple conditions.

is expected to identify the entire GAL pathway-the four GAL genes (GAL1, GAL2, GAL7, GAL10), which are the essential transporters and enzymes for this condition.

In contrast to traditional analysis where conditions are compared with a same, untreated reference (Figure 2.2 on page 57), our algorithm analyzes an integrated reference pool as a representation of the diversity of gene expression behaviors to identify the most specific genes for a target condition (Figure 2.3 on page 58). The scoring of the genes for the "AltCarb" condition by Relief is shown in the figure. The four GAL genes are ranked as the top 4 genes by Relief, and their scores (0.3-0.4) are much higher than the rest of the other genes (majority of which scored less than 0.2). Therefore, our approach is able to re-discover all four specific genes in the yeast Leloir pathway, which constitutes the core metabolic processing by yeast in response to changes in the carbon sources.

To compare with a traditional approach, we also performed a per-gene permutation-based t-test between the galactose treated and the control (untreated) yeast samples, which is a standard method often applied to microarray analysis to uncover differential expression. This analysis identified 236 genes ($p < 0.01$) with GAL7, GAL2, GAL1, GAL10 ranked at 17th, 53th, 87th and 124th, respectively, based upon their p-value. Many of the genes identified are related to more general processes of stress or environmental changes, e.g. MSN2/4 for general stress responses, which are not directly associated with the condition under investigation. Therefore, our approach identified the most likely effectors, and the integration of multiple seemingly unrelated conditions, in fact, effectively reduced the number of non-specific gene candidates.

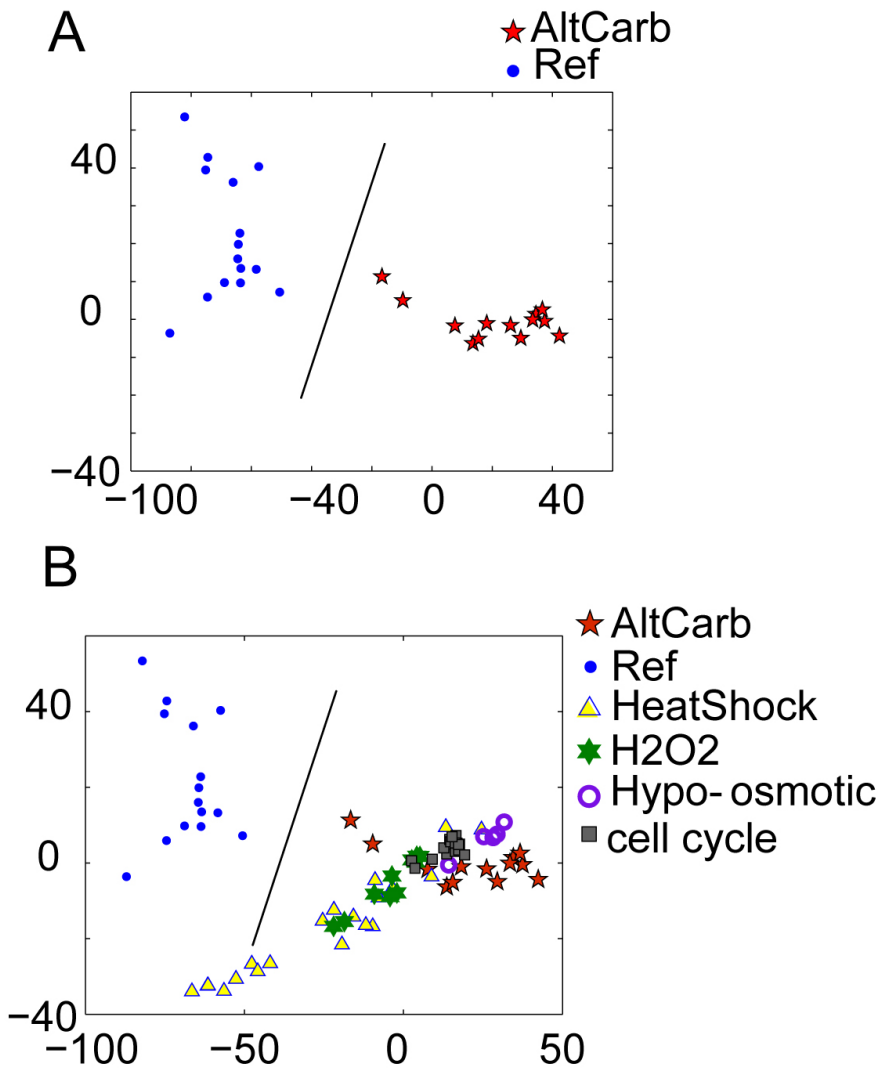

Figure 2.2: Comparing multiple conditions in yeast data. The samples "MEGA" yeast microarray dataset from are plotted in 2-D with their first two Principle Components. The condition "AltCarb" is used as an example. **A:** A traditional treated/untreated analysis compares the condition of interest with untreated yeast samples. The boundary is shown with a black line, and the genes that contribute to such boundary can be identified with feature selection approaches, e.g. t-test, ReliefF. **B:** We plot other conditions such as hyperoxide stimulation, heat stress, etc. The samples of these conditions are similar to samples in the "AltCarb" condition as compared with the same "untreated" reference samples, thus the gene lists that are identified for these different conditions based on such comparison could be similar.

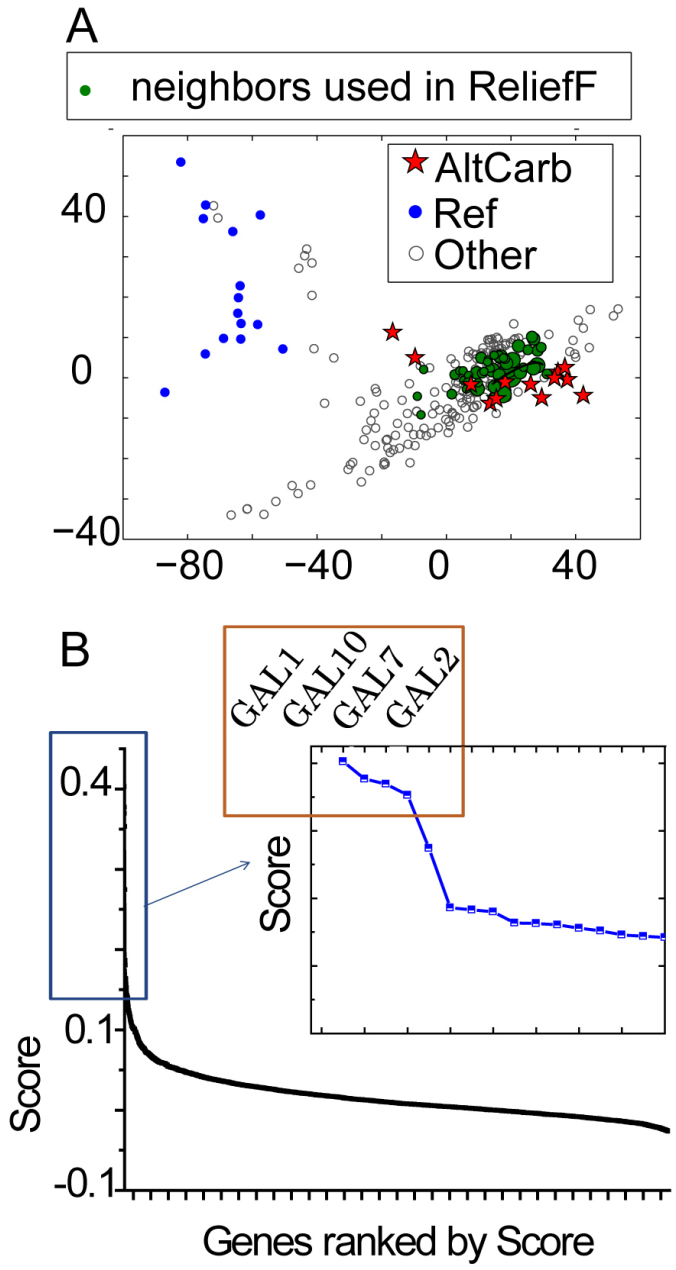

Figure 2.3: Application of Relief on the integrated dataset with multiple conditions **A:** The integrated dataset provides better coverage of the sample space, and ReliefF compares "AltCarb" samples with all other samples in different conditions to achieve better specificity. Nearest Neighbors of AltCarb condition used in the ReliefF procedure are shown in green. The closer a sample is to the "AltCarb", the more important it is in the ReliefF comparisons. **B:** The score of the yeast genes provided by the ReliefF analysis correlates with the importance or relevance of the gene to the specific condition. The genes behave distinctively in the conditions being investigated and are scored the highest to illustrate their importance.

2.3.2 Layer II. Identification of the potential gene regulatory relationships

To explore the regulatory mechanisms of the candidate gene "effectors" identified in layer I for a given condition, we integrate PPI and P-DNA or transcription regulatory network (TRN) (from motif search, ChIP-chip data, or literature information if available) data of the same system (denoted as the reference network), and apply mutual information on the expression data to identify potential regulators of these target genes. We incorporate expression data of multiple conditions to provide a better reference pool, and compute the difference between the conditional and unconditional mutual information in Layer II to identify the factors (i.e. genes or proteins) that directly (physically) interact with and causally regulate the target genes. The information on the physical interactions is derived from PPI and protein-DNA interaction data, and the potential causal factors are inferred based on the conditional mutual information computed from the gene expression data of these interactions.

We establish the accuracy of our methodology against a yeast dataset of conditional TF binding. There are 34 ChIP-chip datasets for yeast samples treated with H2O2 [52]. We used the TF-gene binding indicated in these conditional ChIP-chips experiments as the true TF-gene regulatory relationships, and applied our method in layer II to infer conditional regulatory information for these 34 TFs. Figure 2.4 on page 61 shows the ROC curves for inferring the TF-gene regulatory relationships. Incorporating information of potential transcriptional regulatory network (TRN) based on binding motifs on the gene promoters and literature information (solid lines in different colors) significantly reduce the false positive rate as compared with approaches that do not take advantage of such information, shown in

dotted lines (e.g. mutual information only based approaches). Our approach of incorporating multiple conditions enhances the specificity, which is shown by the red solid and dotted lines (corresponding to with and without TRN) as compared with the traditional setting denoted by the green solid and dotted lines.

We applied Layer II on the yeast "AltCarb" condition to reconstruct a condition specific network, which resulted in 27 nodes and 88 interactions. The GAL switch genes identified in Layer I are used as the targets in this layer to retrieve their potential regulators. Many known regulators of the GAL pathway, such as GAL4 and GAL80 are identified in the network because their expression level (e.g. GAL80) or their interacting partners' (e.g. GAL80 as the interacting partner of GAL4) expression level are correlated with the target genes under this condition, leading to a potential regulatory pathway to the target genes. For comparison, we applied ARACNE, another network reconstruction methods currently available based on mutual information, on the same yeast data for the same condition. The top ranked network module that is identified in ARACNE is the interaction "INH1-APA2" and consists of 2 nodes, in which the INH1gene is an inhibitor of ATPase and APA2 is involved in catabolism of bis(5'-nucleosidyl) tetraphosphates (based on annotation in SGD, `http://www.yeastgenome.org/`). There is no evidence in the current literature to suggest their involvement in response to the AltCarb condition, and further these two genes do not physically or directly interact with the GAL genes. To find the 4 GAL genes in the network required the threshold in ARACNE to be relaxed to allow 2050 nodes and 11,530 edges in the resultant network. In this large network the known, major regulator of the GAL switch, GAL4 and GAL80, were not connected to the four GAL genes. This suggests that many "collateral responses" may be stronger than the initial GAL response. Therefore, by

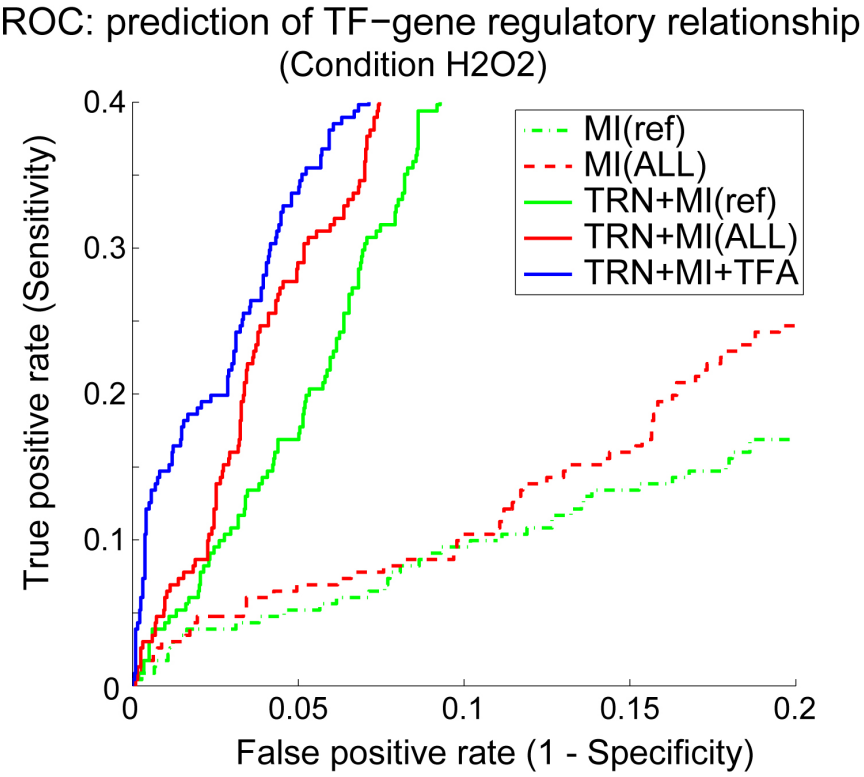

Figure 2.4: The ROC curves for inferring TF-gene regulatory relationships. The true TF-gene regulatory relationships were extracted from conditional ChIP-chip data on yeast (with binding p-value¡0.001) for the samples treated with H2O2. The prediction of TF-gene regulatory relationships are based on mutual information (MI) between TFs and their target genes. The traditional setting is $MI(condition\,of\,interest) - MI(a\,ref\,condition)$, shown in green dotted lines; while we propose to use a variety of conditions as reference: $MI(condition\,of\,interest) - MI(multiple\,conditions\,as\,refs)$, shown in red dotted lines. We apply the same approaches but incorporate the information of Protein-DNA interaction, shown in solid lines, green: traditional setting compared with reference condition, red: compared with a variety of conditions as reference. Further, we use the sum of the target gene expression as a feature of TF activity for a given condition, and apply Relief to identify the TFs and genes that have distinct activity and expression profile for the condition of interest (H2O2). Those TF-gene pairs with significant changes (top 30) on both TF activity and gene expression are elevated to the top of the list of potential TF-gene regulatory relationship based on MI measurement of the multiple condition setting. The result is shown in blue solid lines.

integrating the PPI and P-DNA interactions, and combining layers I and II, our approach effectively reduced the number of hypothesis, focused on the most specific candidates, and identified potential "causal" and "direct" regulators.

Nevertheless, similar to many previous network reconstruction approaches, layer II is based on statistical dependencies between gene expression, assuming a correlation between the expression of the regulator and its targets, which may not necessarily hold in all cases, especially when the transcriptional regulation involves a transcription factor that requires post-transcriptional modifications or co-factors to be activated [80]. Thus we add another layer in our framework, layer III, to address this challenge by accounting for the TF activity.

2.3.3 Layer III. Inference of TF activity and transcriptional regulation

In eukaryotes, post-transcriptional modifications or cofactors are required for many TFs to be activated to regulate their target gene expressions. Although such protein-level information cannot be directly measured in microarrays, the target genes' expression can reflect such regulatory events. Therefore, given potential target genes predicted from motif search, ChIP-chip data or literature information, we can infer the changes of TF activity.

We obtained target genes for each TF from the yeast P-DNA (or so-called TRN) network, and use the summation of its target gene expression level as the feature of a TF in the sample. With these features we can apply ReliefF to identify TFs that show distinct activity (i.e. features) in the condition of interest as compared with all the other conditions. Those TF-gene pairs for which both TF activity and gene expression changed significantly (e.g. within the top 30 of genes and within top 30 of the TFs scored in ReliefF) are raised to the top of

the list of potential regulatory interactions identified in layer II.

Layer III is applied to the yeast H2O2 condition, and the results are shown by the blue lines in Figure 2.4 on page 61. With the inference of the TF activity, the ROC shows a further increase in the specificity (achieve a false positive rate of less than 0.01) of the predicted relationships that are ranked at the top 10%.

We also applied layer III to identify the TFs that regulate the expression of the GAL genes under the AltCarb (adding extra carbon source) condition. We use the top 3 predicted TFs to reconstruct an essential regulatory network for the GAL system, based on known PPI and P-DNA binding information. The results shown in Figure 2.5 on page 64 demonstrate that our approach can re-discover the true network that includes the regulators GAL4, GAL80 and IMP2, the three specific TFs which regulates the GAL pathway for galactose utilization and glucose repression [79]. In contrast, previous approaches based on TF expression level (TYPE 1, approaches based on mutual information or correlation, e.g. Bayesian Network), differences in target and non-target gene expressions (TYPE 2), or correlations within target genes (TYPE 3) (as defined in Chapter 1) did not correctly identify the essential regulatory network (Figure 2.5 on page 64). In particular, GAL4 is post-transcriptionally regulated by its protein interaction with GAL80 [79], whose activity changes can be captured by our approach but not the other approaches compared. A GSEA based approach (e.g. MARINA) may be able to find the enrichment of the target gene groups for these TFs, however, it gives a lower specificity since many other TFs are identified to be more enriched for the AltCarb condition than the 3 GAL regulators.

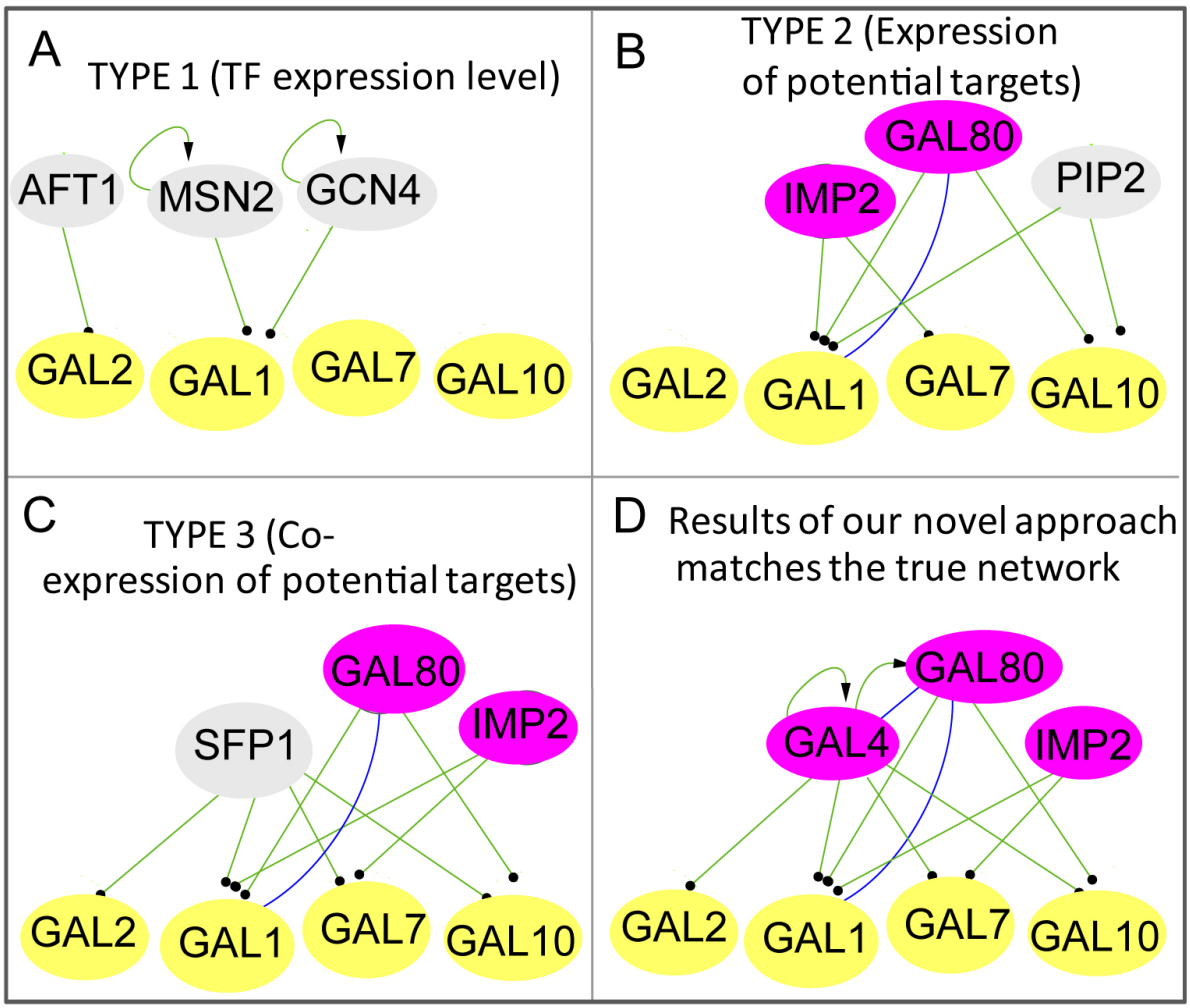

Figure 2.5: Network reconstruction of the GAL pathway. We estimate the activity of the 25 transcription factors that can bind to GAL genes in the transcriptional regulatory network (TRN) based on motif search and literature evidences, and use the top 3 TFs predicted to reconstruct the essential regulatory networks, with the interactions extracted from the TRN (green lines) and the PPI (blue lines). We compare different approaches in estimating the TF activity **A:** TYPE 1: TF activity is determined by its expression level, e.g. correlation or mutual information based approaches; **B:** TYPE 2: TF activity is determined by the expression level of their potential target genes in the TRN; **C:** TYPE 3: TF activity is implicated by the co-expression of their target genes; **D:** Our new approach: use the target gene expression information, i.e. sum of the target expressions and integrate a wide range of conditions to determine the TF activity. The true network includes regulators GAL4, GAL80 and IMP2, shown by the nodes colored in magenta in the figures, which are specific TFs regulating the GAL pathway for galactose utilization and glucose repression in the AltCarb condition. Nodes colored in grey are non-specific TFs for the AltCarb condition.

2.4 Applications on human breast cancer

2.4.1 Layer I: Genes identified for breast cancer

We apply our multi-layer inference approach on an integrated human gene expression dataset (ArrayExpress E-TABM-185) to identify potential biomarkers or targets and their regulators for ER (estrogen receptor)+ and ER- breast cancer subtypes. The dataset integrates 5897 microarray experiments on different human disease, which contains more than 1000 breast cancer samples. Applying differential expression analysis is problematic with such an integrated dataset. A per-gene permutation-based t-test results in thousands of differentially expressed genes for ER+ breast cancer. This high number of genes is hard to validate, while very few genes are identified for ER- breast cancer due to the heterogeneity among the samples in this subtype.

Layer I of our novel approach identified candidate biomarkers for both ER+ and ER- breast cancer, some of which are well-known targets for these subtypes. Expression profile of the predicted genes shows distinctive patterns in the breast cancer samples. The well-known target ESR1 (estrogen receptor alpha) is correctly identified as the top feature for ER+ breast cancer, while the TACSTD2 (tumor-associated calcium signal transducer 2) gene is identified for both ER+ and ER- breast cancer, which we recently discovered could be a potential target for both ER+ and ER- breast cancer subtypes [80].

2.4.2 Layer II: network for ER positive breast cancer

We then reconstruct a regulatory network for ESR1 to assess whether our approach can recapitulate the transcriptional regulators of the estrogen receptor. Human PPI and P-

DNA information is incorporated and layer II and layer III are applied. Finally we infer a regulatory network for TACSTD2 to identify potential regulators for this novel target. Figure 2.6 on page 67 shows the regulatory network inferred for ESR1, the nodes are sized by scores computed in layer II based on the differences between conditional and unconditional mutual information, genes with zero or negative scores (predicted not to regulate the target gene in the condition) without any positively scored interacting partners were filtered out. Based on the binding sites predicted by DECODE in the GeneCards data collection, 7 transcription factors (TP53, JUN, AHR, TFAP2C, GATA3, FOXO3, REST) are selected as potential regulators of ESR1. From BioGrid we extract a list of proteins that may physically interacts with ESR1 or any of these transcription factors, which results in a large interaction network. Our inference approach is then applied to determine which of these components may be the effective "causal factor" of ESR1's specific expression pattern in breast cancer, results in a conditional gene network with 143 proteins in this figure. Of all the 7 transcription factors, 3 (TP53, JUN, GATA3) have positive "scores" in the causal filtering approach, suggesting they can regulate ESR1 expression in cancer cells, and such regulation may also be affected by (but may not depend on) their interacting partners, especially for those of which have higher scores. The 4 other transcription factors are predicted to possibility affect ESR1 but not through direct transcriptional regulation, i.e. the regulation/activity of the TF depends on a co-factor, which are interacting partners to the TFs as shown in the network (Figure 2.6 on page 67).

ESR1 has been extensively studied in breast cancer and a literature search confirmed that many of the predicted regulators are correct. For example, the transcription factors TP53, JUN, GATA3 that are predicted by our approach to directly regulate ESR1 expression,

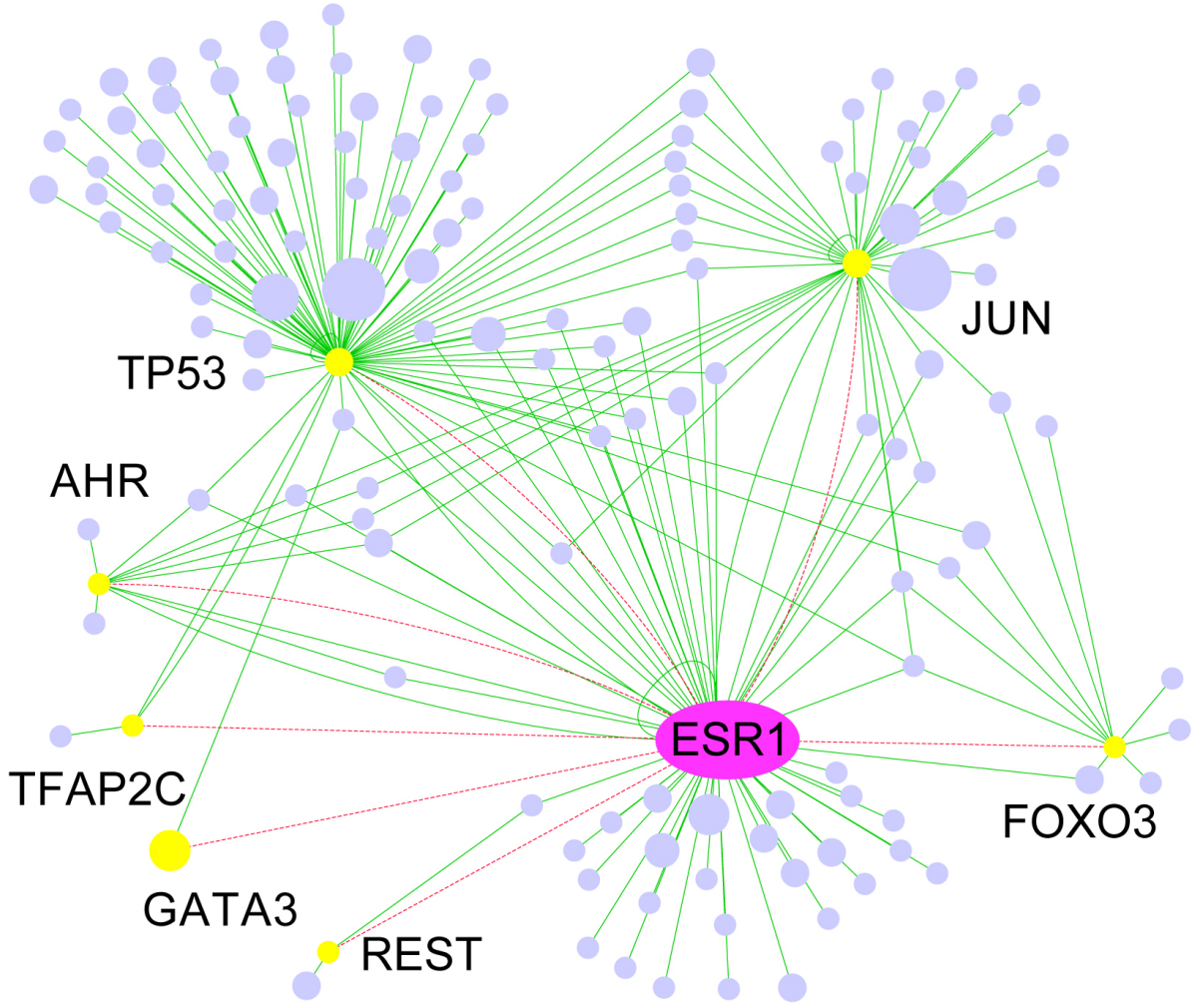

Figure 2.6: The regulatory network for ESR1. Potential transcription factors that can bind on ESR1 gene are colored in yellow. The causal impact (score) is represented by the size of the nodes in the network.

has been experimentally studied and shown to be major transcriptional regulators of the estrogen receptors. GATA3 binds to two cis-regulatory elements on the ESR1 gene and is required for RNA polymerase II to be recruited to the ESR1 promoter [81]. This is consistent with the conditional network constructed by our approach in which GATA3 has the highest score (causal impact) among the transcription factors. TP53 has been shown to bind to the promoter of ESR1 and regulate ESR1 expression in both ER+ (MCF-7) [82] and ER- (MDA-MB-468) breast cancer cells [83]. Finally, it has been previously confirmed that the transcriptional activation complex that was recruited to the ER promoter involves JUN [82], and JUN regulates ER transcription [84] [85].

The transcription factor TFAP2C has been shown to bind to the ESR1 promoter and regulate its expression [86] [87] [88], and silencing TFAP2C reduces significantly ESR1 expression and the estrogen response [89]. However, the activation of TFAP2C requires multiple co-activators, p300 or CBP, whose recruitment depends on the adaptor protein CITED2 [90]. Although the involvement of CITED2 in the activation of TFAP2 has not been shown in breast cancer cells, the knockout or mutation of CITED2 significantly diminished the TFAP2C transcriptional activity in liver cells [90, 91]. and in patients with heart defects [92]. These results support the possibility that CITED2, predicted by our conditional network to affect ESR1 expression through TFAP2C, may be an important regulator of ESR1 expression. Note that TFAP2C by itself has less of a direct "causal" impact based on the differences in its conditional and unconditional mutual information with ESR1, and thus suggestive of post-transcriptional activation.

Similarly, in our network FOXO3 has no direct "causal" impact on ESR1 (i.e., no differences in conditional and unconditional mutual information), but its interacting partners,

SIRT3, shows a difference, based on the scores computed in layer II. However, based on the literature FOXO3 was shown to bind the ESR1 promoter to regulate ESR1 expression [93]. Interestingly, SIRT3, one of the partners of FOXO3 that we predicted to directly impact ESR1, was previously suggested to be a co-activator that increases FOXO3a dependent gene expression [94].

In the network, the transcription factors, REST and AHR, are shown to have no impact on ESR1 expression although based on the binding sites on the ESR1 promoter these TFs could potentially regulate ESR1. To the best of our understanding from the literature, there is no experimental evidence to date that indicate either of the two TFs directly binds the promoter of ESR1 or regulates its expression. However, there is evidence showing that REST may be involved in activating estradiol (E2) stimulation [95] but its expression is higher in ER- than in ER+ patients [95]. Thus its role in E2 stimulation may require potential cofactors, interaction partners or modifications, which is consistent with a zero score in layer II.

There are many proteins that could directly interact with ESR1 and several were predicted to regulate its expression in our breast cancer conditional network for ESR1. From the literature we found experimental evidence to support many as transcriptional regulators of ER. For example, the 2 highest scored proteins are DDX17 and FLII. These factors are transcriptional co-activator of the ESR TF, and can bind to the ESR protein to enhance ESR activation of its own expression. DDX17 has been confirmed as an estrogen receptor alpha coactivator [96], while FLII is required for the recruitment of the SWI/SNF chromatin remodeling complex to enhance ER-mediated transcription of its target genes [97]. The largest family (MED21, MED12, MED7, MED6, MED16) among these proteins that interact with

ESR1 are the components of the mediator complex, which is a co-activator that serves as a scaffold for the assembly of a functional pre-initiation complex of RNA polymerase II, to enhance ER transcription and function [98]. Given that ESR1 can regulate its own mRNA expression [99], these co-factors of ER protein function may play a role in the regulation of ESR1 expression. The known targets for the pathogenesis of breast cancer—- BRCA1 is correctly identified in our network to interact directly and functionally with ESR1 to inhibit estradiol (E2)-stimulated ESR1 transcriptional activity [100].

2.4.3 Layer II and III: the Trop2 network

Given that our approach successfully identified many transcriptional regulators of ESR1. Inspired by the positive results obtained for ESR1, we applied this approach to identify the potential transcriptional regulators of Trop2. We previously identified Trop2 as an important biomarker for breast cancer [80], however, there is no information currently available on the regulation of TROP2. The "causal" network inferred for Trop2 is shown in Figure 2.8 on page 72. Based on this network we predicted CREB1 is a likely transcription factor that regulates TROP2 expression. Our experiment (Figure 2.7 on page 71) shows that the activation of CREB1 by FI (Forskolin-IBMX) induces an up-regulation of Trop2 expression level in the breast cancer cells, which supports our model prediction that CREB1 is a regulator of Trop2.

Although the human protein-DNA binding information is far from completed as compared with yeast, we obtained potential target genes of TFs for which binding motif is known, including TP53, JUN, REST for ESR network and NFkB1, CREB, Evi1 for Trop2 network. The protein-DNA interaction is obtained from TargetMine database [101] based on both

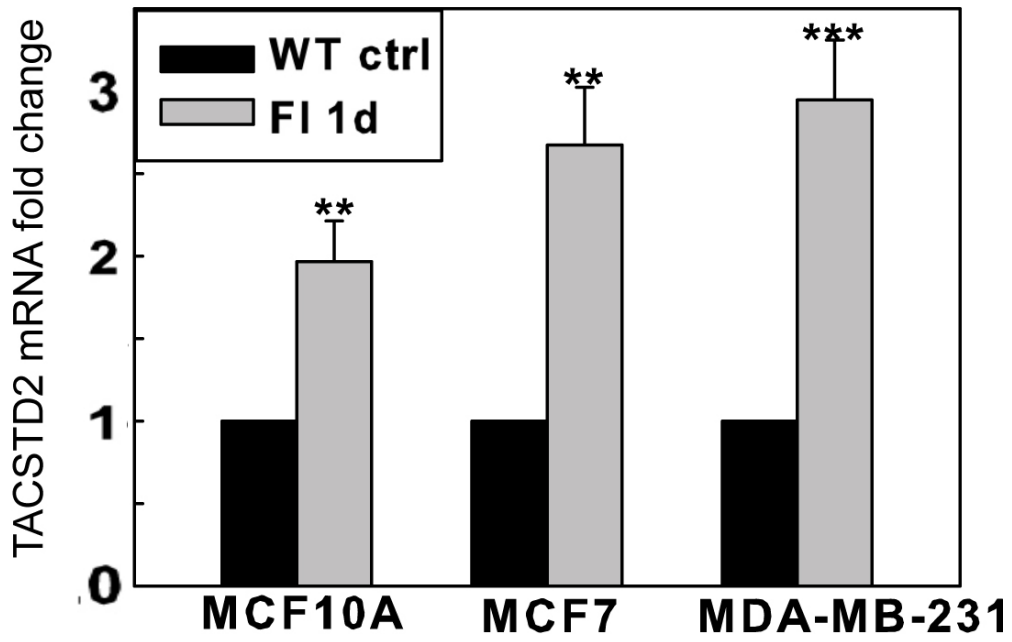

Figure 2.7: The mRNA-fold change of Trop2 in human mammary epithelial cell line and the different breast cancer cell lines upon FI treatment. Quantitative real-time PCR was performed to measure Trop2 mRNA expression levels in MCF10A, MCF7, and MDA-MB-231. The untreated cells (controls) and cells treated with $10\mu M$ forskolin and $100\mu M$ IBMX (FI) for 1 day ($n = 3$) are shown. **: $p < 0.01$, ***: $p < 0.001$.

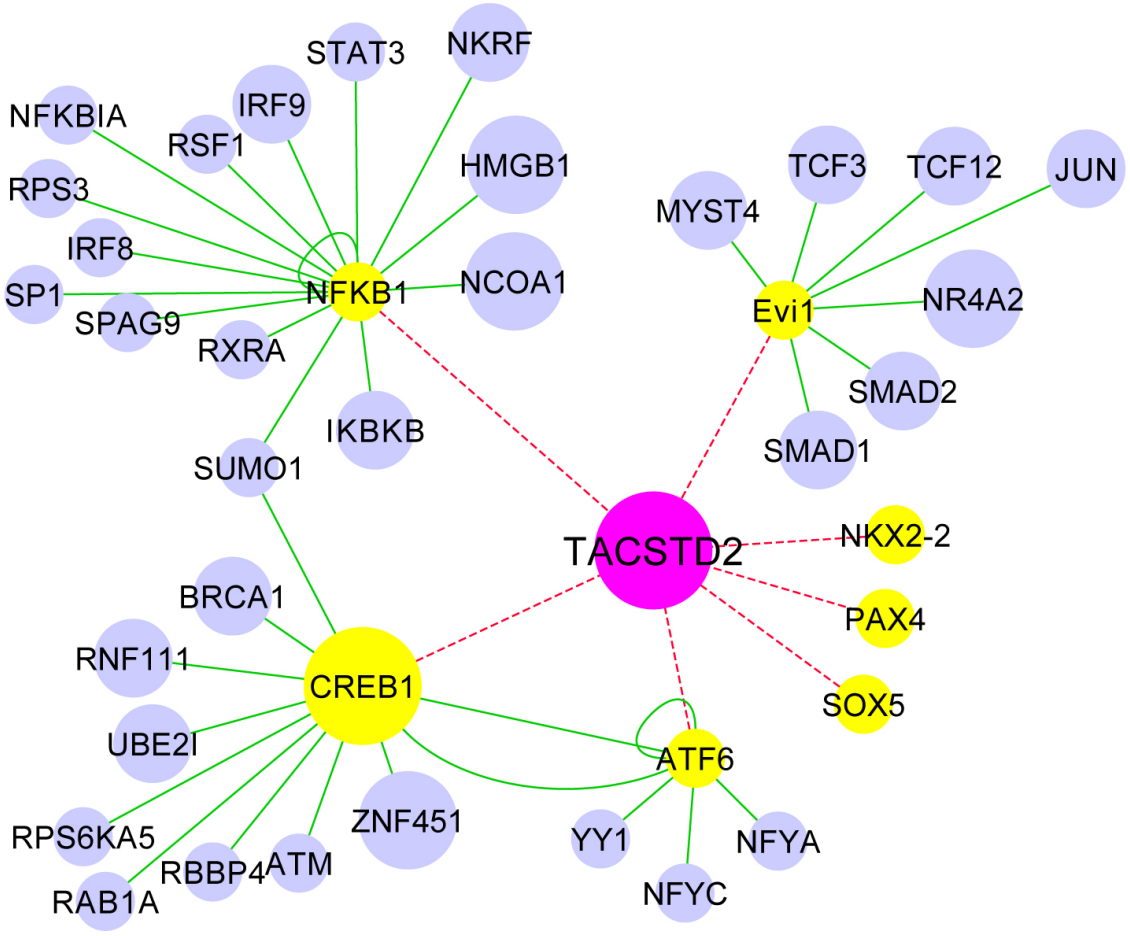

Figure 2.8: The regulatory network for TACSTD2. Seven transcription factors are predicted to bind to TACSTD2 based on motif search, including NFKB1, EVI1, CREB1, ATF6, NKX2-2, PAX4, and SOX5 (colored in yellow). Their interacting proteins are colored in blue. The causal impact (score) is represented with the size of the nodes in the network. Of the transcription factors that could regulate TACSTD2 only CREB1 shows a causal impact (a positive score), and has the highest score among all the proteins in the network that is connected to TACSTD2.

binding motif search on the gene promoter and literature curation. Application of Layer III on these TFs identified that in the ESR network, TP53 activity changes the most, and in the Trop2 network. NFkB1 activity changes more than CREB1 and Evi1. This predicts a potential regulatory role of TP53 on ESR, and NFkB1 on Trop2, and their activity could be controlled at the post-transcriptional level. It has been shown that TP53 transcriptionally regulates ESR1 expression and the regulation relies on many protein cofactors [82], which supports this prediction. For Trop2, our approach predicted another potential transcriptional regulator: NFkB1, which is likely regulated at the post-transcriptional level, given that the activity of NFkB1 is primarily controlled by its cytosol-to-nucleic translocation [102]. Experiment in human breast cancer cell lines (MCF10A and MDA-MB-231) shows that Trop2 gene expression is down-regulated within 2 hours upon either the inhibition of NFkB's protein activity, or reducing its translocation by inhibiting IKK (Figure 2.9 on page 74), which confirms a regulatory role of NFkB on the transcription of Trop2 gene. This would be difficult to identify with current approaches, for example, applying GSEA on the known TFs shows similar enrichment scores for almost all of the TFs, suggesting its lower specificity when applied on large datasets with diverse conditions. These results are similar to what is observed when GSEA is applied to the yeast dataset, where many of the enriched TFs show a zero p-value while GAL4 was not detected. These results support that our approach can identify TFs whose activity is regulated at the post-transcriptional level.

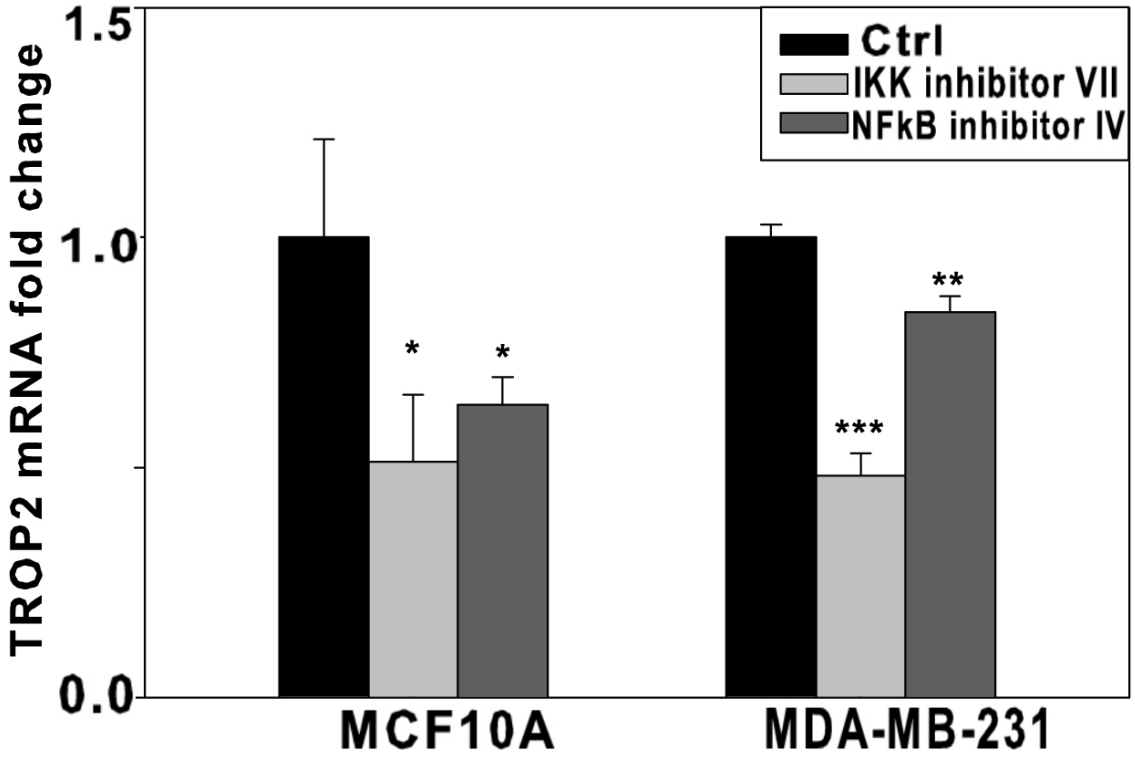

Figure 2.9: The TROP2 mRNA expression levels in different cell types. MCF10A and MDA-MB-231 were treated with IKK inhibitor VII and NF B activation inhibitor IV for 2 hrs, respectively. The TROP2 mRNA levels were measured by quantitative real-time PCR (n=3). * indicates $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. P value was compared to control.

2.5 Comparison of computational approaches to reconstruct gene network

There have been numerous computational approaches developed to reconstruct context dependent gene network based on gene expression profiles. The field has been studied and reviewed (e.g. [80]) extensively but predominantly from the perspective of computational assumptions and methodologies. We attempt to clarify the biological problems and hypothesis that can be solved or predicted by these different approaches. In Table 2.1 on page 77 we summarize and compare our approach with different computational approaches that are currently applied on gene network reconstruction. Approaches based on Bayesian network or correlation network (e.g. in module analysis) infers a "functional network" where the connections predict statistical influences (i.e. correlation) between gene expression, which does not necessary provide clear information on the regulatory relationships and mechanisms. ARACNe [19] and MINDy [73] aim to study transcriptional regulation, which is built on the correlation between the mRNA level of the TF and its target genes. MARINA [74] and NCA [103, 104] can account for post-transcriptional regulation and infers TF activity by incorporating known protein-DNA interactions. MARINA applies GSEA to select the active TFs, where the gene sets contain the targets genes of the TFs, while NCA requires many samples, to solve its complex parametric model. In contrast, the layer approach accounts for different levels of transcriptional regulation to identify condition specific targets and regulations. The applications of the layer approach on yeast and human breast cancer data demonstrate that by integrating multiple conditions, better specificity in target identification and network reconstruction can be achieved, which helps generate more specific

biological hypotheses on the target genes and their regulations.

	BIOLOGICAL PROBLEM TO BE SOLVED				
	Network model learned	**Identify specific genes**	**Identify TF regulations**	**account for TF activity**	**DATA**
BN	influence network	N/A	maybe	N/A	Expression
Module analysis	influence network	maybe ***(many hypothesis)***	N/A	N/A	Expression, PPI
ARACNe	transcriptional regulation	**No**	**Yes**	**No**	Expression
MINDy	transcriptional modifier/cofators	**No**	**No**	Yes, identify **modifiers**	Expression, ***given*** TF-gene, ***given*** modifiers
MARINA	transcriptional regulation	**No**	**No**	Yes, infer TF activity	Expression P-DNA
NCA	transcriptional regulation	**No**	**Yes**	Yes, infer TF activity	Expression ***(large amount)*** P-DNA ***(high quality)***
Three layer approach	transcriptional regulatory relationship	**Yes**	**Yes**	**Yes**	**Expression, PPI, P-DNA**

Table 2.1: Computational approaches for gene network reconstruction.

Chapter 3

Identification of novel targets by exploring gene switches

The learning approaches to identify genes and their regulation based on gene expression, e.g. network reconstruction introduced in Chapter 1 and the multi-layer approach discussed in Chapter 2, are able to find important genes and interactions for a phenotype, such as the Trop2 gene and its transcriptional regulators for human breast cancer. It is straightforward to confirm if an interaction exists in the condition, but one has to design more experiments to understand the functional role of a gene and the regulation in a phenotype. Although these computational approaches can predict candidate genes/biomarkers, they are not able to predict how and why these genes are important for the phenotype in a biological sense.

To answer how and why a gene is important for a phenotype, one has to understand how genes are regulating each other to determine a phenotype, which requires information on the regulatory mechanisms. Therefore, in this chapter, we explore a functional module of gene network—-the gene switches, with systems modeling, wherein we show that gene switches can generate specific pattern of gene expression which can be identified by mining microarray data. Our mining approach demonstrates that one can capitalize on genome-wide

expression profiling to capture dynamic properties of a complex network, thereby to predict gene switches that could be important for a phenotype because they can participate in cell fate decision.

3.1 Gene switches play essential role in cell fate decision, and could be good biomarkers and targets for cancer

Given the complexity of gene regulatory networks, knowledge of the properties of individual components in the network are not sufficient to elucidate the cell physiology. Thus systems biology has evolved to uncover "emergent properties" that arise from the intricate interactions of gene networks. One such emergent property, "switch-like behavior" or "bistability", describes a dynamic feature of a particular gene [105] to preferentially toggle between two steady-states. Multiple steady states are often observed in chemical and biochemical reactions (reviewed by [106]) and are characterized by a non-linear response. Bistability happens to be a special case involving two steady-states, giving rise to a "switch-like behavior". In biochemical reactions, such "bistable" behavior shows a sharp sigmoid function or a hysteresis structure (see examples in Figure 3.1 on page 80), whereby the state of the variable flips between high and low levels. Such an "all-or-none" state transition usually depends on a threshold, i.e., the concentration of the stimulator or regulator. Hysteresis depends further on the previous state of the system.

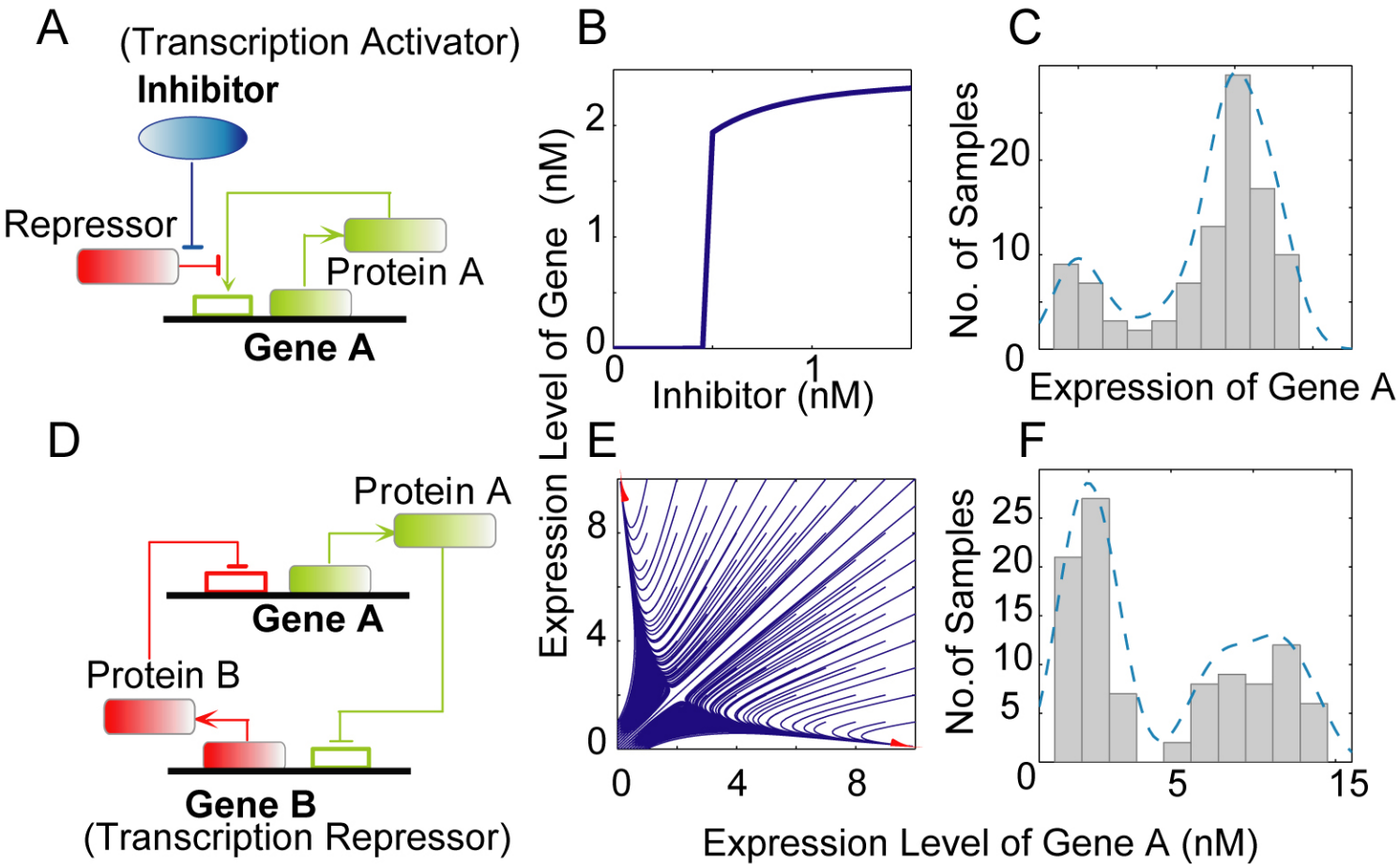

Figure 3.1: Dynamics of gene switches and bimodality in their expression profiles **A:** A synthetic genetic switch that contains a repressed positive feedback is stimulated by an inhibitor of the repressor. **B:** The stimulation-response curve of the genetic circuit. **C:** The histogram of the steady-state gene expression level of 100 random sample simulations of the genetic switch shown in A. **D:** A synthetic genetic toggle switch that contains double negative feedbacks. **E:** The state space of the gene expression. Each trajectory (blue lines) is the response curve with respect to a particular initial condition. The red arrows are the two attractor-states. **F:** The histogram of the steady state gene expression level.

The expression level of a gene switch does not change gradually but rather has two distinct steady-states: HIGH or LOW, ON or OFF, ALL or NONE. The ability of switches to convert a graded signal into a binary response ensures that a cell responds in a decisive manner or unambiguously commit to a specific program [107]. Furthermore switches have been noted for their noise-filtering capacity. Endogenous noise are typically lower for fully repressed or induced expression states than in a gene where the state changes continuously [108] [109].

Bistable behavior of gene switches have been reported to play pivotal roles in many important aspects of cell physiology, including cell fate decisions, cell cycle control, and cellular responses to environmental stimulation [110] [106]. *E.coli lac* operon is a famous gene switch that uses a hysteretic feedback to decide between glucose and lactose utilization [111]. Many bistable systems have been discovered in bacteria, including the genetic transformation in *Bacilius subtillis* and sporulation in many bacterial species [112]. In mammalian systems, gene switches and bistability have been postulated as the underlying mechanism for cellular differentiation, but rarely has this been confirmed experimentally, until recently with the work on neutrophil differentiation [113]. Another interesting observation is that cells have "memory", and hysteresis has been shown to govern short-term memory in lymphoid cells, preserving information of past encounters with antigen [24]. Thus, the discovery of gene switches in cellular responses has become a milestone in molecular biology and prompt strong interest in understanding the function and design of gene networks [111].

Switches play a central role in cell decision, and the ability to predict whether switches can occur without a priori detail information of the network would be significant. For instance, the ability to identify which genes are turned on or off in cancer versus normal cells would have a tremendous impact on identifying the most pertinent molecular signatures

or targets for drug therapy. Therefore a major challenge confronting the field, which we address in this study, is how to effectively identify gene switches or bistable states by mining high-throughput data. An approach that could predict switches based on high-throughput data not only provides candidates of biomarkers but also associates the candidates with a potential biological function/mechanism for the phenotype.

3.2 How to identify gene switches

Despite the importance of gene switches, identifying multiple steady-states, and in particular switches, has been difficult. Our understanding of gene switches has been mostly based on simulations of generic feedback circuits and well-characterized biological modules [114–117]. Theoretical studies of feedback circuits have elucidated general principles of network dynamics, but they usually lack solid evidence to associate these principles with real physiological processes in cells. Few studies have succeeded in demonstrating functional roles of actual switches in biological systems by coupling detailed kinetic modeling with rigorous experimentation [118] [24]. This is because well-characterized models with equations and kinetic parameters are difficult to obtain for real, complex biological systems, in part because current techniques are not able to quantitatively measure reaction constants at the single-cell resolution for all the network components. Alternatively, researchers in synthetic biology have designed artificial gene networks with specific functions and implemented the interactions by manipulating or bringing together exogenous genetic components [119] [120] [121]. Thus current methods of experimentally studying switches have been limited to well-characterized or synthetic small modules.

Previous computational approaches addressed this question by analyzing the network

topology. These studies assume that bistability requires particular feedback structures [107, 122], and discovered dynamic features by searching for these structures (e.g. positive feedbacks) in protein-protein interaction and protein-DNA interaction networks [123]. However, these feedback structures do not ensure switch-like behavior. From modeling and simulations of genetic circuits, positive feedback (or even feedback itself) has been shown to be neither necessary [124] [110] nor sufficient [125] to ensure switch-like behavior. Furthermore, it is less likely that one can uncover a dynamic property from static networks.

Alternatively, we theorize that the dynamic "behavior" of a switch could be identified by analyzing the gene expression profiles from a wide range of conditions. We propose a top-down mining approach to identify gene switches from microarray gene expression data. Taking advantage of the tremendous amount of expression data, our approach aims to identify bimodality, which we hypothesize is an essential characteristic of a gene switch.

3.2.1 Simulation of kinetic models of gene switch

A gene switch has two steady states, which will produce a bimodal distribution in its expression profile when sampled across different conditions. Figure 3.1 on page 80 show the gene network topology of two typical regulatory circuits that exhibit bistable behavior. The ordinary differential equations (ODEs) for the synthetic systems are as follows:

- Positive self-feedback:

$$\frac{\mathrm{d}A}{\mathrm{d}t} = p[\frac{A^2}{1+A^2}][\frac{1}{1+R^2}][1-\frac{A}{2.5}] - \deg(A) \tag{3.1}$$

$$R(i) = 10[1-\frac{K_i}{1+K_i}] \tag{3.2}$$

A represents the expression level of Gene A, $pA^2/[1+A^2]$ describes the self-binding and activation of the transcription, and $1/1+R^2$ is the effect of the repressor, in which R depends on the stimulation —the concentration of the inhibitor i. $deg(A)$ is a linear function for the degradation of A. The model is constructed by [126] for a mammalian cell system.

- Double-negative feedback:

$$\frac{\mathrm{d}A}{\mathrm{d}t} = \frac{a}{1+B^2} - \deg(A) \tag{3.3}$$

$$\frac{\mathrm{d}B}{\mathrm{d}t} = \frac{a}{1+A^2} - \deg(B) \tag{3.4}$$

A,B represents the expression level of Gene A and B, respectively. a is a parameter about the strength of the cross-repression of the two genes. deg is a linear function for degradation. The model is constructed by [119] in *E.Coli*

Figure 3.1 on page 80 part A shows the positive self-feedback transcriptional system under the control of a transcriptional repressor. Part D in the figure shows the double-negative feedback system, also known as a toggle switch, which produces mutually exclusive activation of two genes. Both circuits have been synthesized and implemented in cell systems [126] [119] to confirm their switching behavior. Simulations based on the kinetic models of these systems [126] [119] confirm the on/off and toggle-like switching behavior in their response curve and state space. By simulating random samples from a wide range of conditions with different initial states, this unique feature of two distinct steady-states of gene switches results in a gene expression histogram profile containing two modes Figure 3.1 on page 80. This bimodality is observed despite the noise (20% Gaussian noise) imposed on the parameters.

3.2.2 Data mining to identify gene switches

A challenge in experimentally identifying gene switches is their population effect. In single cell experiments, if obtainable, the response curves would represent individual cell measurements, and a gene that switches will exhibit a steep jump between the steady states. However many biological measurements (RT-PCR, Western Blotting), including microarray analysis, provide the population-average. In fact, even with single cell measurements, individual clones can contribute cell-cell variances, with differences in the protein expression levels across different cells. In Figure 3.2 on page 86 the cell-cell variance is modeled by a Gaussian distribution in the protein expression and different cells in a clone would then respond differently to stimulation, leading to a continuous change in the averaged response curve. This explains, in part, the difficulty in identifying switches through standard experiments.

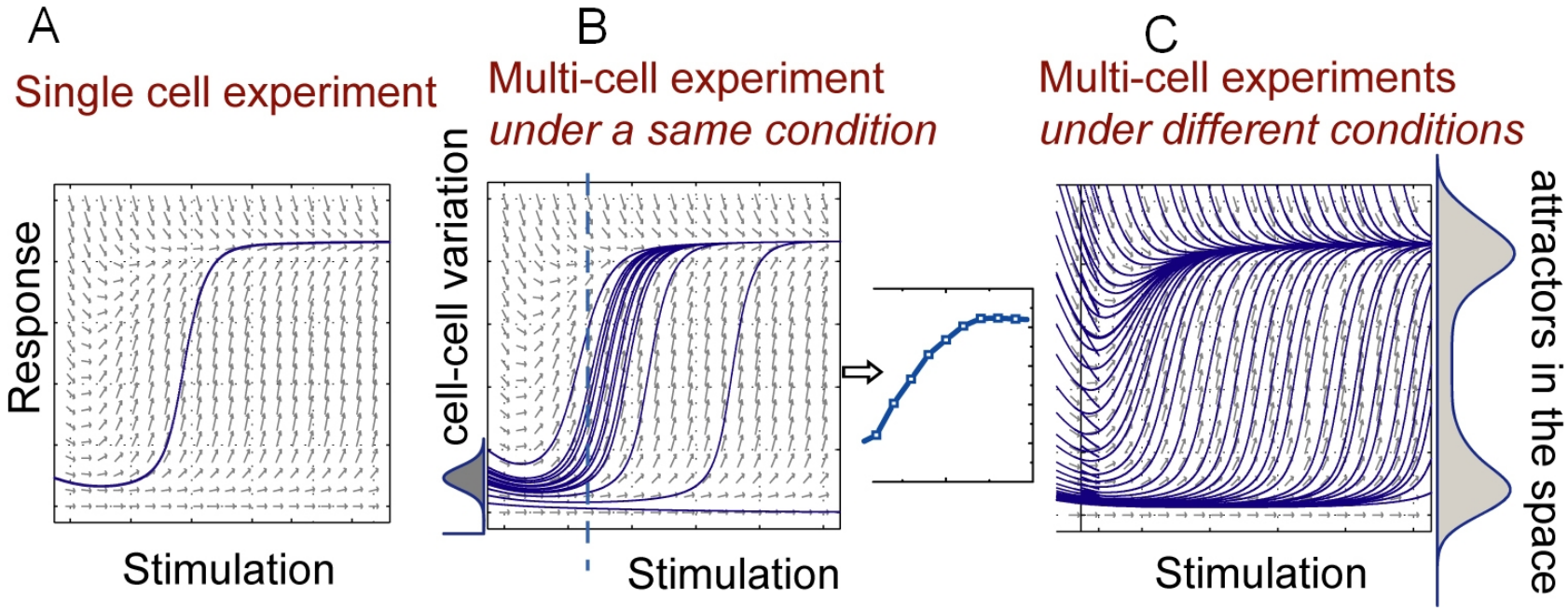

Figure 3.2: Schematic representation of a phase plane of a gene switch. **A:** Single cell dose response experiments should be able to measure the response curve and uncover the switch-like behavior. **B:** Experimental measurements of the average expression level of a cell population will mask the switching behavior. A Gaussian distribution is plotted to represent the cell-cell variances in the population. Different cells, according to their initial gene expression level, could have different response curve (blue trajectories). The averaging of the variation in the responses results in a seemingly graded response. **C:** Experiments across a range of different conditions allowing for the sampling of a large state space recover the switch-like behavior. Each sample could fall in the neighborhood of a possible steady state (points on the blue trajectory). The steady states (on/off states of the gene switch) are the dense regions of the possible response curves in the state space, i.e. the samples occurs at higher frequencies in these states, which results in a bimodal distribution in the observed profiles.

We proposed that an unbiased sampling across a range of different conditions could address this issue and help reveal the dynamic feature of gene switches. In Figure 3.2 on page 86 we show analytically, potential response-curves (the blue trajectories) in the whole state space of a gene switch. Each sample within the system would asymptotically approach one of the two possible steady states (dark blue region). Since the on/off states are the steady states which most cells will concentrated in upon stimulation, the samples will have higher probability of staying in these states, leading to a bimodal distribution in the observed expression profiles.

To identify bimodality researchers have used the DIP statistics [127] or Gaussian mixturemodel to [128](`http://www.astro.lsa.umich.edu/~ognedin/gmm/`) identify bimodal distributions. Since what we need is not only a quantity for bimodality, but also an explicit separation of the conditions into two categories corresponding to two expression levels, we choose the Gaussian mixture model. The Expectation Maximization algorithm is implemented to separate the data distribution into two Gaussian models. The criterion used to assess the fitting is the Akaike information criterion from information theory.

$$AIC = 2k - 2\log(L) \tag{3.5}$$

Where k represents the number of parameters, and L is the goodness of fit, defined by the likelihood of observing the data given the model (one or two Gaussian in this case). We use the differences between AICs of the two models, the "ΔAIC", to compares the fit with a unimodal vs. a bimodal distribution. ΔAIC provides a measure for an "unconditional" bimodality in which the profiles show bimodal but the condition for the "switch" is yet to be investigated. When a particular condition is specified (e.g. the breast cancer phenotype),

the separation D can be used to identify if there is a distinct state for the condition, or, a "conditional specific" bimodality:

$$D = \frac{\mu_1 - \mu_2}{\sqrt{(\sigma_1^2 + \sigma_1^2)/2}} \tag{3.6}$$

Where (μ_1, σ_1) are the mean and deviation of samples in the specified condition, while (μ_2, σ_2) are the mean and deviation for all other samples. We perform theoretical analysis and provide proof-of-concept applications on both synthetic and yeast microarray datasets.

3.2.3 A Proof-of-Concept application of the E2F-Rb network

The E2F-Rb network is a well-characterized system in mammalian cell fate determination, whereby the Retinoblastoma (Rb) protein regulates the transcriptional factor, E2F, to control the restriction point for the G1-S transition in cell cycle [129]. A simplified kinetic model was constructed for the E2F-Rb system [118], in which two genes Myc and CycD (Cyclin D: Cdk4,6) are activated by sufficient growth signal (serum) to induce E2F activation, which then directs the synthesis of downstream factors, such as CycE for DNA replication. The E2F self-activation and CycE-mediated E2F activation constitute two positive feedbacks in the system. It then was experimentally [118] confirmed that the level of E2F switches ON or OFF for cell-proliferation and cell-cycle arrest, respectively, suggesting E2F acts as a gene switch, while CycD and Myc do not show such switch-like behaviors.

We perform simulations based on the kinetic model [118] to generate a synthetic gene-expression dataset. The stimulation-response curve of a single-cell is shown in Figure 3.3 on page 90, and confirms a graded response for Myc/CycD and bistable dynamics for E2F. The downstream factor CycE, controlled by E2F, also shows a switch-like response. Intro-

ducing a distribution in the expression level to represent cell-cell variation within a clone, and averaging multiple simulations shows that population averaging for any one condition disguises the switch-like behavior and is indistinguishable from a graded responses, which is consistent with previous RT-PCR experiments [118].

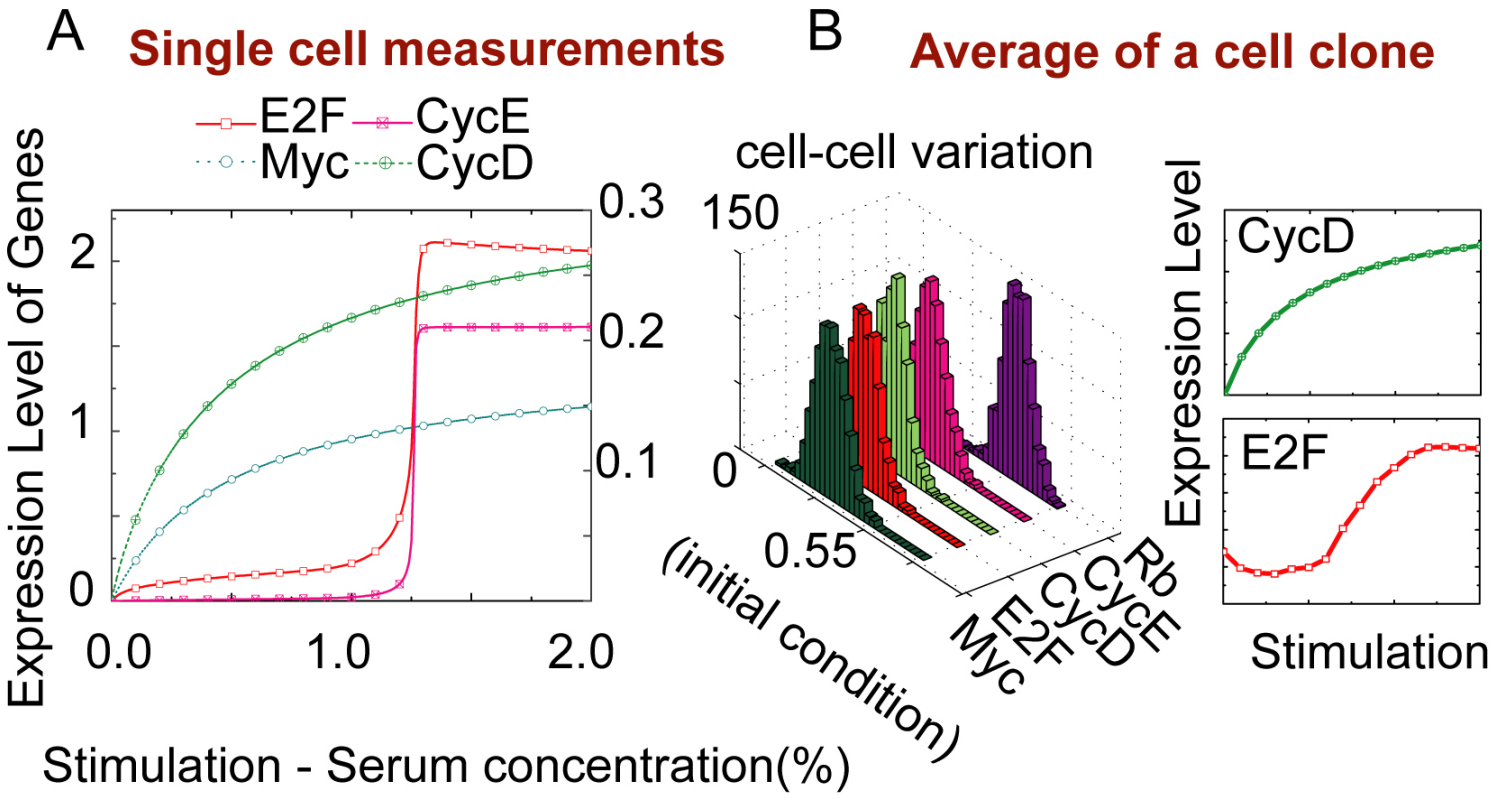

Figure 3.3: Proof-of-concept example: simulation of the E2F-RB network. **A:** Simulation of the kinetic model based on a fixed initial condition represents measurement at the single cell-resolution of the system. The response curves of serum stimulation on the different genes in the model are plotted. **B:** Assign a Gaussian distribution with small variances on the initial gene expression level of the untreated cells to represent the cell-cell variation in a clone. Simulation-results are computed by averaging the responses of 100 cells in a clone.

We then simulate 100 cell clones, each clone with a random initial condition, and measure the steady state expression level of the network components for each clone. In this way we synthetically generate 100 "microarrays" for 100 different conditions. It is clear genes that have two steady states, i.e. E2F and CycE (effector of the gene switch), exhibit two distinct modes in their expression profiles (Figure 3.4 on page 92). Each gene's ΔAIC value is calculated from the synthetic expression data and the switches exhibit higher ΔAIC values than the non-switches. Thus ΔAIC can be used to rank and help uncover genes that are bistable.

3.2.4 A Proof-of-Concept application to Yeast microarray data

We apply our mining approach to an integrated yeast microarray dataset containing 500 yeast experiments and calculate the ΔAIC value for each gene in the dataset. With such a large set of conditions, the ΔAIC value is fairly robus. A histogram of the ΔAIC value among the yeast genes is shown in Figure 3.5 on page 93. Most genes have low ΔAIC, and their expression appear unimodal. However, a few genes have high ΔAIC values and clearly show bimodality.

The genes with high ΔAIC values have distinct states under different conditions. By collating and comparing those conditions under the two distinct expression states, one can potentially identify the phenotypes in which the genes are functioning. Given that a phenotypic ontology is not available, it is difficult to compare conditions. Nevertheless, one approach is to categorize conditions by the type of perturbations, e.g. heat shock (with different temperature and length of time), hypo-osmotic shock (different time points), and extra carbon sources (different carbon source), etc., and check if one of the two states of a

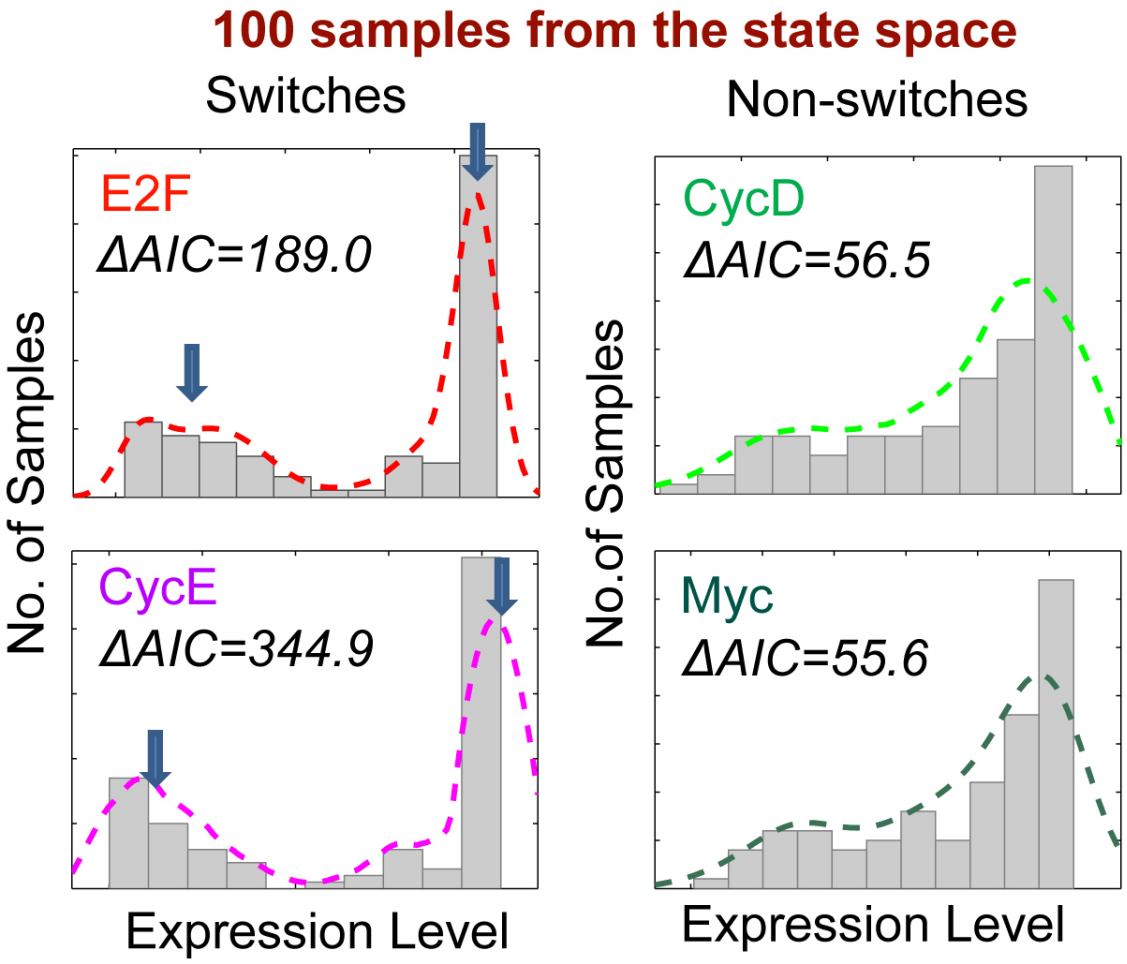

Figure 3.4: Mining approach to identify switches in the E2F-RB network. Sampling of 100 clones under different, randomly generated initial conditions. The simulation results are shown as histograms of the expression level of the different genes, together with their ΔAIC values.

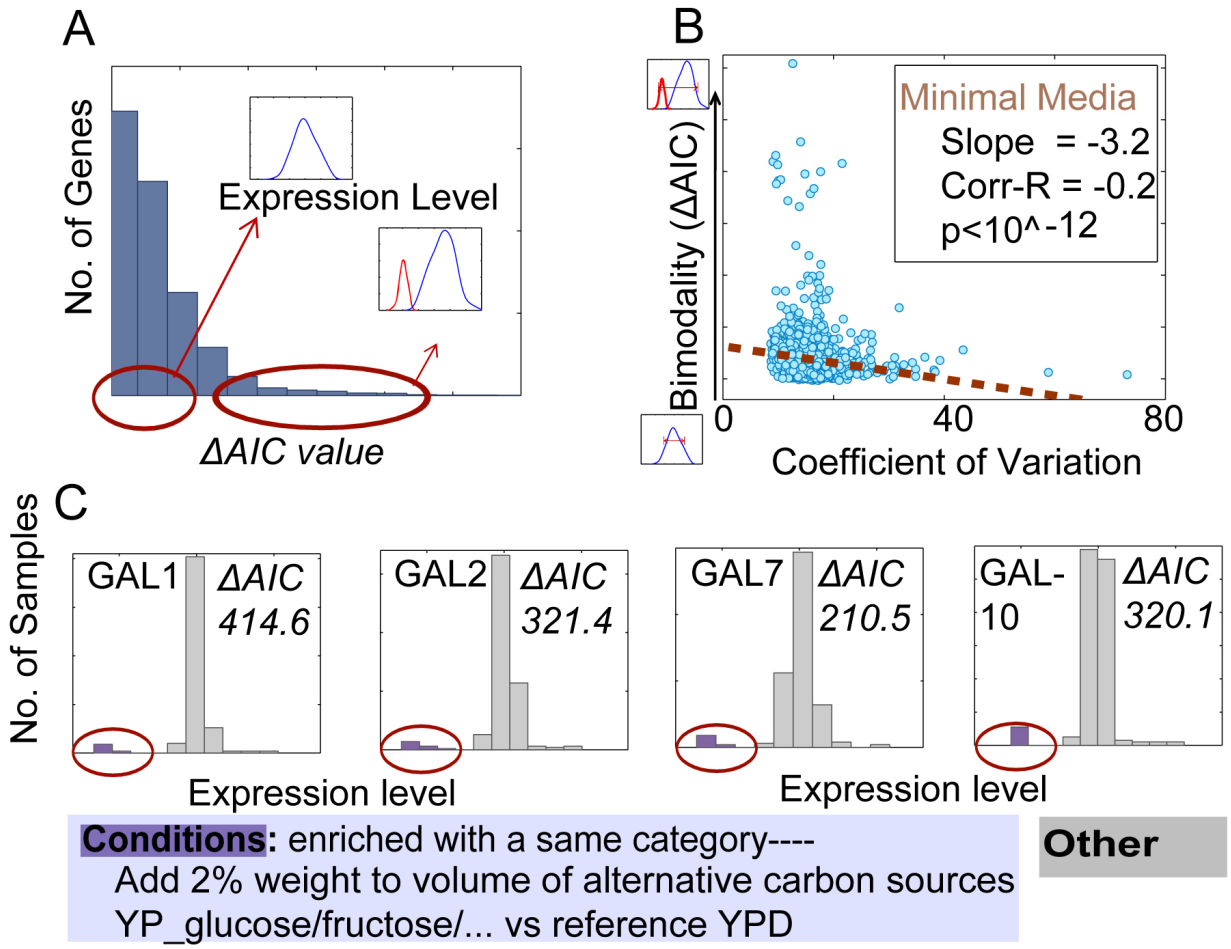

Figure 3.5: A proof-of-concept application on the yeast dataset. **A:** A histogram representing the distribution of AIC value among the yeast genes. Most genes have small ΔAIC and exhibit an uni-modal expression profile. A relatively small number of genes have high ΔAIC and show bimodality in their expression. **B:** A negative correlation between bimodality (described by ΔAIC) and expression noise (as described by the coefficient of expression variation) in the yeast genes. **C:** The four GAL genes show bimodality and one of their modes are enriched within the same condition/category.

putative switch is enriched within a category of conditions. Using this approach, we correctly uncovered genes that have switch-like behavior, namely GAL1, GAL2, GAL7 and GAL10 (Figure 3.5 on page 93). These genes all have ΔAIC values that rank among the top 5% and show bimodality, with one of their two modes containing conditions from the same category, i.e. "adding extra carbon sources". The bimodal profiles show that by adding 2% (weight to volume) extra carbon sources into the media, with the exception of galactose as the extra carbon source, the expression of these four genes shut down. It has been reported that these four genes function in the same pathway for galactose utilization, i.e., the well-known "GAL genetic switch" (review: [79]). The addition of alternative carbon sources results in "glucose-repression" of the GAL pathway. During this process, the high level of glucose or other carbon sources (other than galactose) induces the formation of the repressor complex (protein Mig1p and Cyc8-Tup1) and upon its binding to specific upstream repressing sequences (URSG) on the GAL promoters, it prevents the activation of these four GAL genes by the transcription factor GAL4, thereby turning off the galatose utilization pathway.

Current knowledge on the existence and functional machinery of other gene switches is limited. However we show next that by integrating information of the regulatory network and proteomic data, the genes with high ΔAIC obtained from our analysis could be possible switches or at least important genes with respect to the phenotypes. We calculate the ΔAIC values of transcription factors in the yeast transcriptional regulatory network (based on binding motif data), and observe that the leaf-nodes — genes that are only regulated by one factor and are not regulating any other transcription factors —- tend to have significantly lower ΔAIC value (average $\Delta AIC = 135 \pm 9$ compared with overall average $\Delta AIC = 223 \pm 58$ for transcription factors, $p < 0.01$). These genes which have few regulators and

do not transcriptionally control transcription factors are less likely to have feedbacks at the transcriptional level, and therefore switching dynamics. Thus the dynamic property we infer of the molecular components within a network is contingent on the network organization.

Next, we analyze single-cell proteomic data that includes noise in the protein expression measurements. We find a weak but significant negative correlation between the ΔAIC value of a gene and its coefficient of variation, which captures the noise of its protein expression (Figure 3.5 on page 93). This suggests that genes with higher ΔAIC value, showing bimodality, tend to express relatively lower levels of noise. This observation that genes with lower expression noise under normal conditions are more tightly controlled highlights their importance in the network, and is consistent with previous suggestions that gene switches have noise-filtering capacity [108] [109].

3.3 Identify characteristic signatures of human breast cancer

Since the state of gene switches in the genetic network governs the phenotype [130], we postulate that recognizing specific gene switches will enable one to identify biomarkers or molecular signatures that would be better drug targets for treating a disease. We demonstrate the utility of our mining approach in human breast cancer by analyzing a paired breast cancer/normal tissue expression dataset against the integrated human gene expression dataset.

We analyze the paired breast cancer/normal tissue expression dataset (GSE15852) (Pau Ni et al, 2010) against the integrated human gene expression dataset [131] to identify char-

acteristic signatures of human breast cancer. First, we calculate the separation value D [132] for the top 10% ranked genes by ΔAIC to examine whether the expressions of these genes are bimodal when comparing the breast cancer (1119) samples against all other phenotypes (4,777 samples for 300 conditions). Biologically this indicates whether a gene potentially shows bistability and could be involved in the "switching" or transition to a breast cancer phenotype. $D > 2$ has been suggested to indicate whether the separation into two Gaussian distributions or modes is distinctive [132]. Considering the large amount of noise in the microarray data, we accept separation values of greater than or close to 2 (¿1.8) to indicate bimodality, which results in 17 genes showing distinct bimodality in breast cancer.

Next, an independent microarray dataset (GSE15852) with 43 paired breast cancer samples of diverse histopathological characteristics is analyzed to test if the 17 genes are expressed differently and show distinct bimodality in breast tumor as compared to normal breast tissues. Comparing such "local" expression profile (paired normal and cancer conditions) with the "global" expression profile (across various conditions) identified that of these 17 genes, 12 genes (ESR1, SPDEF, IRX5, ERBB3, ERBB2,CRABP2,RAB25, FXYD3, TACSTD2, DSP, AGR2, CDH1) show bimodality in both datasets (Figure 3.6 on page 97 shows the flow chart of the procedure). One type of genes is bimodal within the breast cancer samples, herein denoted Type 1, with estrogen receptor-alpha (ESR1) having the highest separation. The other type of gene switch shows predominantly one modality within the breast cancer samples, herein denoted as Type 2, and is where we find the TACSTD2 (a.k.a. Trop2) gene having the highest separation value within this group.

Many of the genes that show Type 1 bimodal behavior also exhibit the biomdality within the breast cancer samples (Figure 3.6 on page 97). Known therapeutic targets for breast can-

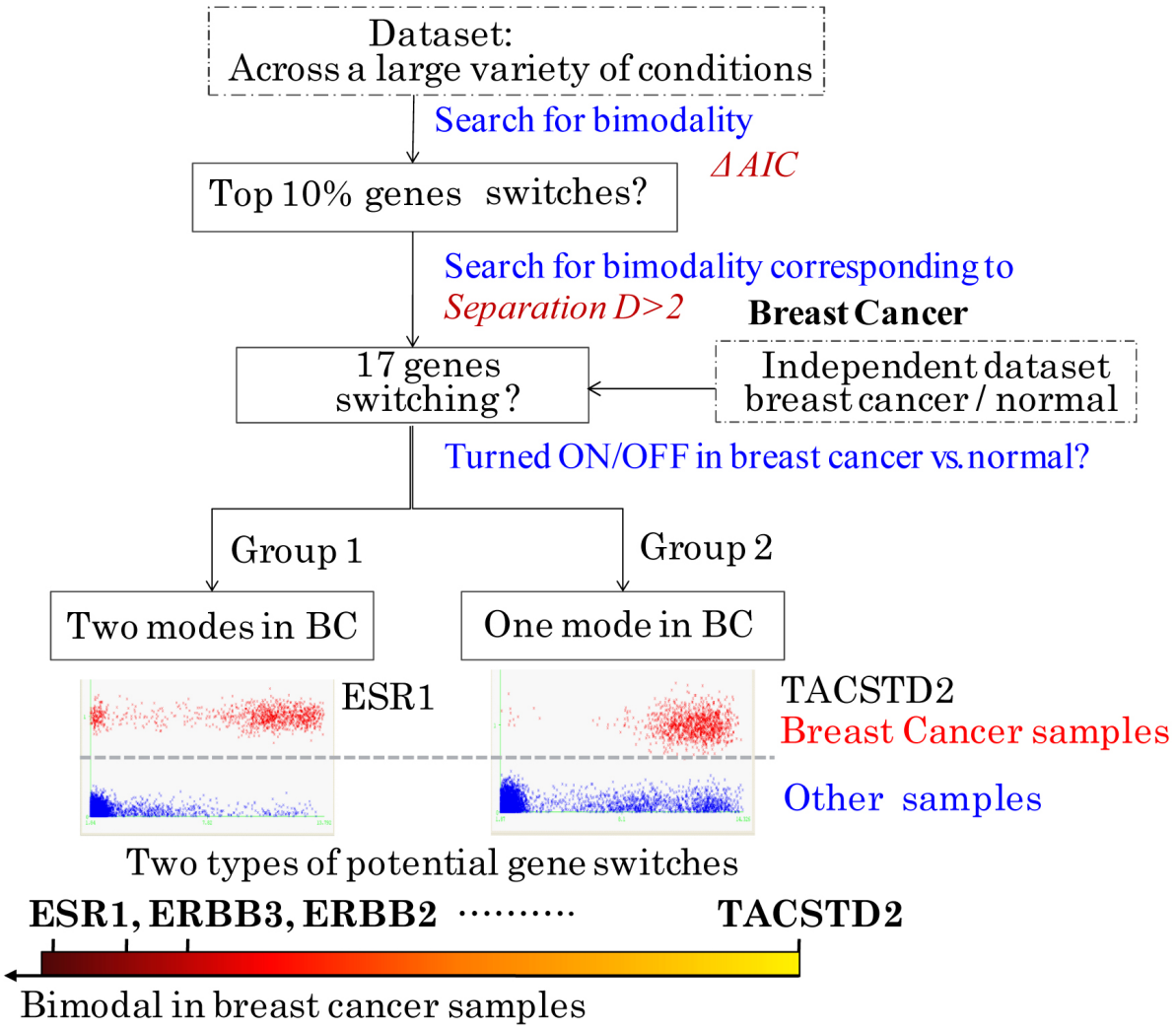

Figure 3.6: Identification of potential gene switches for breast cancer. We analyzed the integrated dataset to search for bimodality in the gene expression profiles. Genes are ranked based on their ΔAIC calculations. The top 10% are selected to compute the separation D with respect to breast cancer. An independent dataset is then used to further examine candidate genes.

cer, such as ESR1, ERBB2 (HER2) and ERBB3 (HER3), are identified as showing bimodality in their gene expression level in breast cancer. Their bimodality in the cancer samples represents well-known subtypes in breast cancer, i.e. ER+/ER- and HER2+/HER2- subtypes. ESR1 (estrogen receptor alpha) is a well-known transcription factor involved in the development and progression of breast cancer. Previous immunohistochemical analysis showed a bimodal distribution in estrogen receptors (ER) expression —- the majority of breast cancer patients express either ER-negative (low expression) or unambiguously ER-positive (high expression), of which ($\approx 80\%$) are ER+, while moderate ER immunostaining is rarely observed [133]. This supports our discovery of bimodality of the ESR1 gene expression within the breast cancer samples. It has been a decade since researchers attempted to explore the mechanism underlying such an all-or-none expression pattern of estrogen receptors. It was previously reported that the ESR promoter activity is increased by co-transfection of the wild-type ESR expression vector, suggesting a positive contribution of ESR to its own expression [99]. A recent study uncovered that miR-375 is involved in a forward feedback loop that regulates ESR1 expression, whereby ESR1 enhances miR-375 expression and miR-375 targets and reduces the expression level of RASD1 (ras dexamethasone-induced 1) gene, which is a transcriptional inhibitor of ESR1 [134]. These studies provide evidence of a potential positive-feedback (with a double-negative circuit) induced bistability of the ESR1 expression, where the topology is similar to a toggle-switch design. ERBB2 and ERBB3 interact with each other and are known to be transcriptionally regulated by ESR1 [135]. A recent study [136] identified a positive feedback of ERBB2 through the transcription factor c-Jun, which could provide a potential explanation for the bimodality observed for ERBB2.

The molecular characterization of the Type 1 genes (e.g. ESR, HER2) suggests the

development of therapies for ER+/PR+ and HER2+ would be effective for these breast cancer subtypes, however $\approx 15 - 20\%$ of the breast cancer tissues expressing low levels of these biomarkers (i.e. triple negative subtype) have poor prognosis and few treatment options. Moreover, patients that are responsive to commonly used drugs, such as tamoxifen (estrogen antagonist) and trastuzumab (anti-HER2 agent), eventually acquire resistance to the drugs. $\approx 30\%$ of tamoxifen-responsive tumors become resistant [137] [138], and the resistance invariably ensues at some point with trastuzumab. Given the increase in resistance to drugs that target the ESR receptor alternative therapeutic targets are needed.

The second type of potential gene switch, herein denoted as Type 2, shows unimodal behavior in the breast cancer tissue (Figure 3.6 on page 97) and is differentially expressed in almost all the paired breast tumor/normal tissues as compared with non-breast cancer samples. The top gene showing this type of switching behavior is Trop2. Type 2 gene switches uncovered by our analysis show a distinct state in the breast cancer samples, and could be a potential biomarker or drug target that does not rely on the ESR receptor. We characterized the Trop2 gene, and found it to be distinctively expressed at higher levels in almost all of the breast cancer samples, ER+/-, HER2+/- subtypes. We confirm that the expression of Trop2 gene is high in breast cancer cell lines MCF7 and MDA-MB-231 as compared with non-cancer cells (Figure 3.7 on page 100).

Our network reconstruction in Chapter 2 identified CREB and NFkB as potential transcription factors that regulate the expression of Trop2. We also observe a significant increase in the correlation between the expression level of CREB and Trop2 in the paired breast cancer dataset. The correlation coefficients in the normal breast tissue are $0.15, 0.06, 0.03$ for the three CREB probes in the Affymetrix array, and the correlation coefficients in the breast

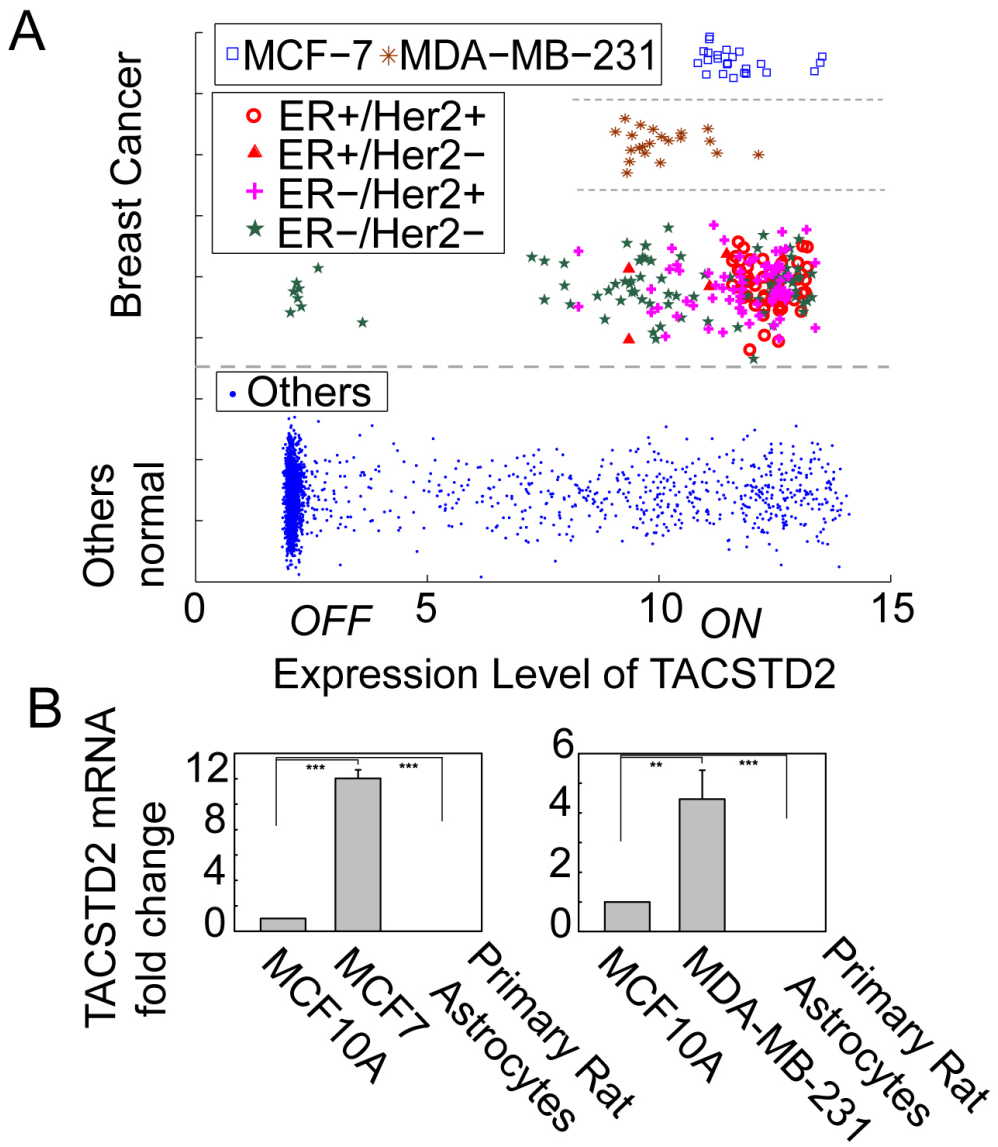

Figure 3.7: The expression profiles of Trop2 in breast cancer. **A:** The scatter plot shows the gene expression level of Trop2. x-axis indicating the expression level, the values in y-axis are randomly generated to reduce the overlap between samples. Subtypes of breast cancers are determined by their expression levels of ESR1, PR, and Her2. **B:** The Trop2 mRNA levels in human mammary epithelial cell line, MCF10A, in breast cancer cell lines, MCF7 and MDA-MB-231, and in primary rat astrocytes were measured by quantitative real-time PCR ($n = 3$). **: $p < 0.01$, ***: $p < 0.001$.

tumor tissues are $0.46, 0.21, 0.31$, respectively.

To assess the possible switching behavior of Trop2, we performed flow cytometry to probe the Trop2 protein level at single-cell resolution. For both MCF10A and MCF7 breast cell lines the Trop2 protein level shows a bimodal distribution in their cell population (Figure 3.8 on page 102), which is a property of a bistable system. We stimulated the cells with FI (Forskolin and IBMX) to induce cAMP, which is an activator of CREB [139], and measured the Trop2 levels. Both Trop2 mRNA and protein levels increased significantly upon stimulation, thereby supporting a possible transcriptional regulation by CREB. Upon activation of Trop2 by FI, a decrease in one of the modes with a concomitant increase in the other mode, instead of a gradual increase in the protein level, (Figure 3.8 on page 102) is indicative of a switching behavior. The activation essentially increases the number of cells with Trop2 levels at the ON state and decreases the cells with Trop2 at the OFF level. In contrast, the expression of the Trop2 protein in primary rat astrocytes shows a unimodal expression under the same test conditions. Furthermore stimulation of astrocytes by FI leads to a non-significant change in the protein level and with the cells predominantly remaining in the OFF steady-states.

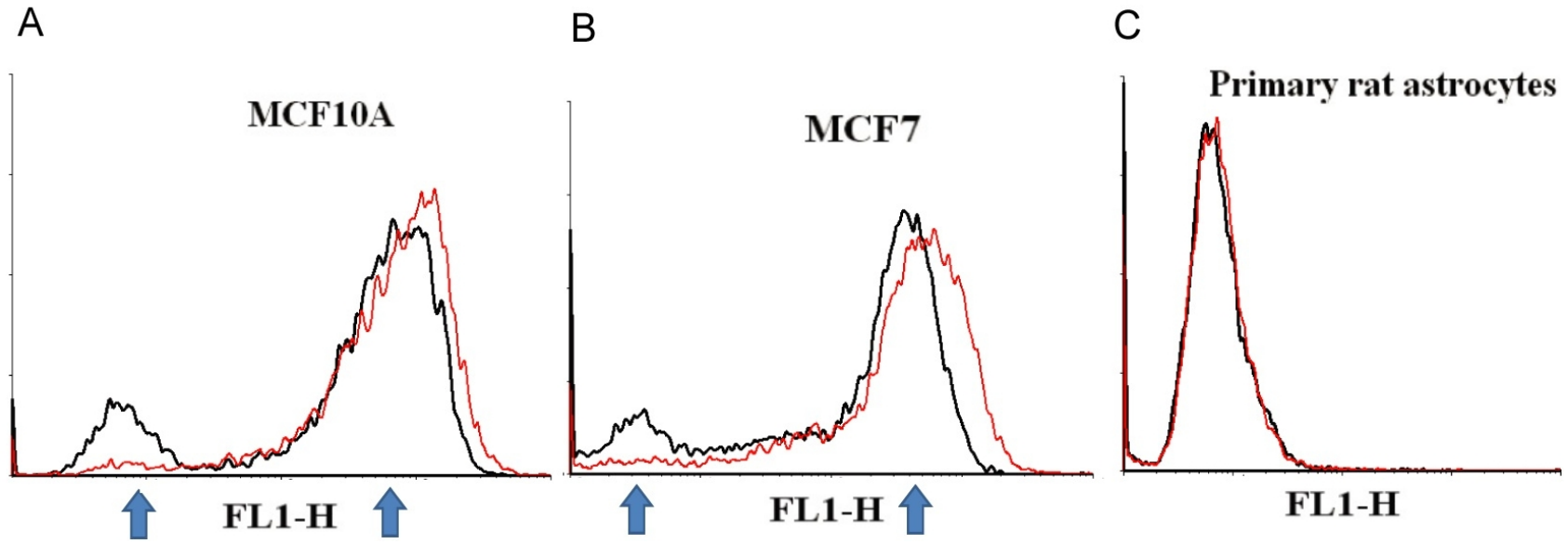

Figure 3.8: A switch-like behavior in Trop2 expression. Flow cytometry analysis of Trop2 expression in MCF10A, MCF7 and primary rat astrocytes (Black lines). The cells were treated with $10\mu M$ forskolin and $100\mu M$ IBMX (FI) for 1 day (Red lines) and the two modes of Trop2 in MCF10A, MCF7, and primary astrocyte cell population are pointed out by the blue arrows. Note the primary astrocytes have only one mode.

Our mining approach uncovered a unique expression pattern of Trop2 in breast cancer, and experiments confirmed Trop2 show bimodal behavior in breast cancer cell lines. Trop2 (Trop2) is a cell surface glycoprotein, first discovered to be highly expressed in trophoblast cells that become invasive and metastasized to form the outer layer of blastocyst in embryo development [140]. Recent studies, along with our analysis of breast cancer samples, found Trop2 to be highly expressed in a variety of epithelial cancers and show low to no expression in normal somatic cells. High expression of Trop2 in squamous-cell carcinoma [141], pancreatic [142], colorectal [143] and gastric [144] cancers have been associated with poor prognosis and higher incidence of metastasis and death. Trop2 was identified as an oncogene in colorectal cancer cells [145]. Although not essential for cell proliferation under normal condition, ectopic expression of Trop2 enhances anchorage-independent cell growth, promotes tumorigenesis and metastasis in colon cancer cells. Knock-down or inhibition of the protein reduces the invasiveness of aggressive colon cancer cells [145]. In our analysis we also found Trop2 to be highly expressed in many colon cancer samples and shows bimodality, however the percentage of colon cancer samples with Trop2 at the ON state ($\sim 60\%$) are less than in breast cancer ($\sim 99\%$), suggesting Trop2 could be a better target for breast cancer.

In previous microarray analysis of breast tumors, [146] Huang et al. studied "aggregate patterns of gene expression" with respect to lymph node status and recurrence, and identified "metagenes" that could predict the outcomes of the patients. Trop2 is found among the "metagenes" in their list; however the list consists of more than a hundred genes with potential predictive value. In contrast, we find the Trop2 gene to be the top gene in the list that shows the Type 2 behavior. Interestingly, the distinctive HIGH/LOW expression level of the Trop2 gene has been implicated as a marker for stem cell characteristics in

prostate basal cells [147] and hepatic oval cells [148]. The prostate basal cells and hepatic oval cells, considered progenitor cells, show HIGH expression of the Trop2 gene and maintain self-renewal capability [147] [148], and thereby implicating a potential role of Trop2 in cancer initiating stem cells. Although Trop2 has been reported to be associated with cancer, the regulatory mechanism of Trop2 remains unclear. Combining computational prediction (network reconstruction in Chapter 2) and experimental analysis, we found that CREB and NFkB could regulate Trop2 in breast cancer cells.

Since the completion of our study of network reconstruction and gene switches in breast cancer, two research papers have been published in ONCOGENE [149, 150] that provide experimental evidence in human cancer tissues to support Trop2 as an oncogene. These independent experimental studies confirm the effectiveness of our approach in predicting specific genes for a phenotype and potential targets for disease. Our discovery of the transcriptional regulators of Trop2 complements the network biology study of Trop2 in [150]. We identified novel regulators of Trop2 that have not been discovered as yet based on previous approaches of network reconstruction. Our discovery of NFkB as a regulator of Trop2, together with the evidences provided in [150] that Trop2 regulates NFkB expression, suggest a potential positive feedback structure that supports the "switching" behavior of Trop2 expression.

In this chapter we apply systems modeling to define a specific pattern that is emerged from the complex interactions in gene network—-the gene switches, and explore the pattern by mining gene expression data, to be able to provide more specific predictions on gene function. Our mining approach demonstrates that in the absence of a priori knowledge of the specific network architecture, one can capitalize on genome-wide expression profiling to

capture dynamic properties of a complex network. To the best of our knowledge, this is the first attempt in applying mining approaches to explore gene switches on a genome-scale and this is a first case a single cell level bimodality and bistability can be predicted from microarray data.

Researchers recognized that "genetic switches" behave in a discrete manner, but this feature is usually lost in biochemical analysis of large cell populations due to the difficulty in distinguishing between changes in the proportion of cells and their expression level in the two states [113]. For example, it is hard to determine from population measurements whether the expression level of a gene increases gradually by 70%, or whether 70% of the cells are "switched" ON. In this study, we provide an alternative approach to identifying possible gene switches by capitalizing upon the large amount of available microarray data. The large sample set enables the characterization of the state space by uncovering the presence of the two attractor-states where the majority of the samples should fall. Thus, if an ON/OFF switch behavior exists in a system the state space will show bimodality or bistability, which are relatively stable with respect to perturbations [151]. It has been suggested that bistability or multiple steady states [130] exists in large gene networks [152] [153], and these attractor-states represent different phenotypes [130]. Thus, by sampling across different conditions, which are less affected by population averaging, one can reveal this dynamic feature of regulatory networks. In this sense, our study provides a different perspective that takes advantage of the large integrated set of expression data, and suggests a mechanism-based framework to perform the meta-analysis. This approach of integrating microarray data from a diverse set of conditions provides a common "context" of gene behaviors, whereby one can obtain a better understanding of the specific function of a gene for a particular condition

under investigation.

By applying the computational analysis on human microarray data, we uncovered a unique expression pattern of Trop2 in breast cancer, and experiments confirmed Trop2 show bimodal behavior in breast cancer cell lines, further, our perturbation study suggest a potential bistable mechanism is involved. Therefore, not only does our computational approach predict biomarkers/targets, but also it can suggest the mechanism how and why the biomarker/target could be functionally important.

Chapter 4

Identification of genes (microRNAs) that determine a phenotype

Network reconstruction based on gene expression and interaction data in chapter 2 identifies genes and regulation that are specifically functioning in a phenotype and the mining approach in chapter 3 identifies potential gene switches. These learning approaches treat the genes as "features" and the phenotype as "labels" (a particular annotation on samples) to find features that are able to differentiate between different labels/annotations, by comparing the transcriptomes that represent different phenotypes/labels. Therefore, these approaches essentially identify genes and interactions that distinguish a phenotype. However, one of the key goals in systems biology is to understand how complex molecular and cellular outcomes arise from the dynamic interactions at the detailed mechanistic level, wherein the "phenotypes" are not independent labels/annotations for samples but involve many different aspects to define a biological state, which is reflected by the complex molecular and cellular outcomes. For example, a "cancer" phenotype represents very special genetic states, gene regulatory mechanisms, cellular behaviors, as well as specific metabolic states. In this sense, the learning approaches are limited by treating phenotypes as "labels"/annotations.

To predict genes that determine a phenotype, it is necessary to model the phenotype as a complex biological state and perform simulation studies to perturb the state, that is, as introduced in Chapter 1, to address the following two questions:

- How does one model a phenotype?
- How does one define the perturbations that determine or change a phenotype?

In this chapter, we study the *cancer* phenotype by modeling the human metabolic network, thereby "cancer" is no longer merely a label/annotation but modeled as an entire metabolic state. We developed a novel approach to simulate context dependent metabolic states upon perturbation of gene expression, which is then applied to predict microRNAs that can inhibit cancer growth.

We model the phenotype in terms of metabolic states because:

1. Metabolism is crucial to cell growth and proliferation. Deficiencies or alterations in metabolic functions are known to be involved in many human diseases. For example, the pathogenesis of diabetes results from malfunction in the regulation of metabolic pathways, leading to alterations in insulin signaling, oxidative metabolism, and lipid/fatty acid metabolism [154]. Dysregulation of the metabolic system is also implicated in carcinogenesis [155]. Most cancer cells have higher glycolytic rates, the so-called "Warburg effect" [156, 157] [158]. A recent study of breast cancer further uncovered alterations in glucose metabolism mediated by phosphoglycerate dehydrogenase (PHGDH) enzyme [159], whose expression was found to be associated with poor prognosis [160]. Since metabolism plays an essential role in cell growth and proliferation, genes regulating metabolism have been used as drug targets in the treatment of

cancer [161] [162] and other diseases involving metabolic disorders [163] [164], including diabetes, atherosclerosis and fatty liver disease. Thus, understanding the human metabolic system is important and provides a complementary approach to study and identify potential treatments for complex human diseases.

2. It is a complex system that is well-studied in characterizing many of its components/interactions and much of the detailed mechanistic information is available for systems modeling. A (steady) state of the system can be defined by metabolic fluxes distributed in the network and reliable techniques are available (experimental and computational metabolic flux analysis) to measure the state. Current reconstructions of the global human metabolic network provide a computational platform to integrate knowledge gained over the past 50 years of research on human metabolism [165], thus enable a systems modeling approach to study *in silico* the global effect of perturbations on the network to generate hypotheses and help understand the mechanisms underlying the genotype-phenotype relationship.

4.1 Systems modeling of a metabolic state

To model a phenotype as a metabolic state, we need to characterize the components and interactions in the system by the reconstruction of global human metabolic network, and we need a simulation approach that can be used to define a metabolic state.

4.1.1 Reconstruction of global human metabolic network

The global human metabolic network has been manually curated based on an extensive collection and evaluation of the genomic and bibliomic data. The first two installation of the network were released in 2007: the Edinburgh Human Metabolic Network [166] and the human Recon 1 [165], each contains a list of human reactions, metabolites and gene-protein-reaction relationships. The Gene-Protein-reaction (GPR) represents functional relationships between genes/proteins (e.g. enzymes) and the corresponding reactions they catalyze or control. For example, in human Recon 1, the genes are first mapped to their transcripts, accounting for alternative splicing. Then, based on Boolean rules of OR and AND, the transcripts are mapped to proteins. The proteins are then mapped to reactions by Boolean rules based on current knowledge of their effects on the reactions.

The two networks (Edinburgh Human Metabolic Network and the human Recon 1), developed independently by different research groups, consist of many different genes and reactions. The Edinburgh Human Metabolic Network contains more genes and metabolites, but was not compartmented in its initial release. Compartmentalization requires assignments of metabolic reactions into different cellular organelles (cytoplasm, nucleus, endoplasmic reticulum, mitochondria, lysosome, peroxisome, and Golgi apparatus) and accounts for the transportation and exchange of metabolites between organelles. Human Recon 1 is a compartmented network which could be used in reconstructing predictive models for systems biology studies, therefore, most of the recent applications have been based on Recon 1. An overview of the publications thus far that used Recon 1 is reviewed by Bordbar and Palsson [167]. Notably, in 2010, the compartmentalization of the Edinburgh Human Metabolic Network was completed and its current release is a compartmented, and more complete

human metabolic network [168].

The reconstruction of the global human metabolic network uses a bottom-up approach. Researchers begin by compiling reactions of cellular metabolism to build a network through the collection of gene annotations, enzymes and pathway information from genome (e.g. NCBI, Ensembl) and pathway (e.g. KEGG, ExPASy) databases. Researchers then refine the network by manually collating literature evidences, including journal articles, reviews and textbooks on metabolic functions, biomass composition, growth conditions and gene-reaction associations. The constructed draft network is converted to biochemical models to evaluate the basic functionality, and simulations are performed to check for consistency with the current knowledge. The whole process runs iteratively to incorporate as much information and minimize gaps and inconsistencies. The protocol for the reconstruction process is available in [169].

The major difference in a metabolic network as compared with other biological network, e.g. Protein-Protein Interaction, Protein-DNA network, is that the metabolic network represents a biochemical system that is charge-balanced, mass-balanced and compartmentalized, which not only provides information about whether there is an interaction, but also how it happens and what it is produced as a biochemical reaction, thus can be directly converted into mathematical equations based on the biochemical reactions for model predictions.

4.1.2 Modeling and simulation based on human metabolic network

A reconstructed human metabolic network can be represented by a system of stoichiometric reactions. This system of reactions can be modeled as ordinary differential equations,

however the reaction rate constants and metabolite concentrations are typically difficult to obtain, thereby limiting their applicability to small well-studied networks. However, since the stoichiometry of metabolic reactions are not organism or context-dependent but is fixed by mass balance, one could apply Constraint Based Modeling (e.g. Flux Balance Analysis, FBA [170]) to simulate the state of the system without detailed kinetic data, assuming that the flux distributions based on the stoichiometric mass balance are at steady state or pseudo-steady state.

Mathematical representation of reaction network and Constraint Based Modeling:

- Reactions: S (Stoichiometric Matrix), with m compounds (rows) and n reactions (columns). The stoichiometric coefficients are negative for the substrates of each reaction, and positive for the products.
- Flows: v (n by 1 vector) on all reactions
- Concentrations: X (m by 1 vector) of all compounds

Thus we have

$$dX/dt = Sv \tag{4.1}$$

Assuming pseudo-steady state, the time derivative is zero, therefore: $Sv = 0$. So the flux distribution v that satisfies this equation is in the null space of S.

In the human metabolic network, $n > m$ results in an under-determined system that does not have a unique solution. Adding constraints permit a "feasible" solution to the system of equations, for example, a "flux capacity" constraint determines the upper and lower bounds of the flux through a reaction. Imposing mass balance and capacity constraints will define the space of feasible steady-state flux distributions of the network. Geometrically the space

looks like a "flux cone" in the null space of S. A visualization of the "flux cone" is shown in [171] to demonstrate the way a solution space could be narrowed by the steady-state and capacity constraints. Further, in FBA we define an objective function Z, which is a linear function of fluxes. An objective function could be

$$Z = c^T V \tag{4.2}$$

in which c is a column vector to assign weights to each reaction, c^T is the transpose of the vector c, and V is the flux vector through all reactions. Optimization of the objective function Z identifies a unique (or multiple) set of flux configurations within the flux cone. The constrained linear optimization problem can be solved by linear programming [172] [171].

The form of the objective functions, constraints and the optimization problems can vary depending on the biological applications, which are different variants of the Constraint Based Modeling. For example, the Flux Sensitivity Analysis (FSA) estimates the objective flux change in response to perturbations in some reactions of interest [173]. The Flux Variability Analysis (FVA) explores the solution space to exam the maximum/minimum fluxes for each reaction. Further, current approaches to reconstruct context dependent metabolic networks are essentially different variants of the Constraint Based Modeling. A detailed review of the algorithms in the Constraint Based Modeling is provided in [174].

4.2 Prediction of the metabolic state-change upon perturbations

Metabolic genes could be differentially expressed under different conditions in different cell types, resulting in different metabolic states of the cell. Changes in gene expression drive changes in metabolic fluxes which manifest cellular phenotypes. This is the central idea to associate gene expression with metabolic network to determine a metabolic state for a phenotype (i.e. reconstruct context-dependent metabolic network) and study how perturbations in gene expression can change the phenotype.

A summary of the pipeline for the systems modeling and its applications based on human metabolic network is shown in Figure 4.1 on page 115. Since the global human metabolic networks (the human Recon 1 and Edinburgh Human Metabolic Network) are generic metabolic networks that collate information from all types of human cells, the reconstruction of a context-dependent (i.e. condition/cell-type/tissue/organ specific) network is required prior to *in silico* analysis of the particular system under investigation. Once the context-dependent reconstruction is obtained, one can simulate the metabolic phenotypes under different perturbations to identify essential gene targets or pathways, or predict cellular responses to different treatments.

4.2.1 Reconstructing context-dependent metabolic network

Similar to the reconstruction of the global human metabolic networks, one could follow the protocol (Thiele and Palsson, 2010a) to collate literature evidences and gene annotations, and manually identify the context-dependent reactions to reconstruct a condition-specific

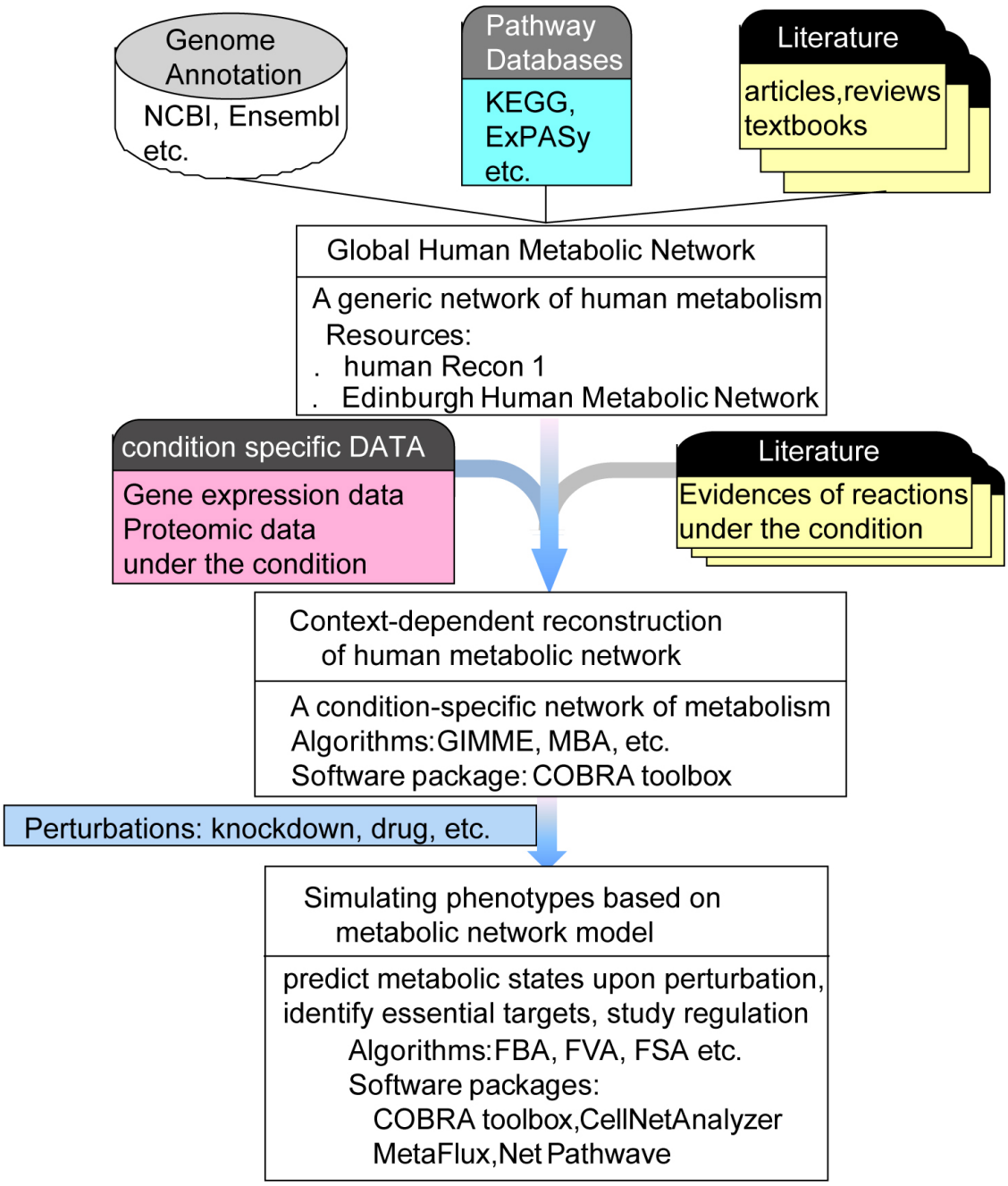

Figure 4.1: A pipeline for systems biology applications of human metabolic network.

metabolic network. This has been applied in [175] [176] to achieve comprehensive reconstructions of hepatic and neuronal cells. However, the process requires manual evaluation of thousands of papers, and curates thousands of genes and metabolites, which is lengthy, and requires tremendous effort and labor. Therefore, many studies have focused on automating the reconstruction of a cell-type or tissue-specific metabolic network, by incorporating high throughput gene expression data rather than manual curation of the cell/tissue-specific network.

Assuming changes in gene expression drives changes in the metabolic states; the basic idea in automated reconstruction is to identify active reactions by incorporating condition-specific gene expression profiles. Most reconstruction approaches start by analyzing the gene expression data to determine if a gene is "present" (highly expressed) or "absent" (low expression level) for the condition being investigated, then selects the active reactions according to their corresponding gene/enzymes' expression level. For example, the Gene Inactivity Moderated by Metabolism and Expression (GIMME) algorithm, developed by Becker et al. [177], uses gene expression data to determine active and suppressed genes with an expression threshold, and determines active reactions based on the state (active or suppressed) of the corresponding enzymes. "Inactive reactions" are removed unless they are required for a desired functionality (according to a predefined objective function). Another approach to study human tissue-specific metabolic states was developed by Shlomi et al. [178]. The approach does not require an objective function but matches active reactions with expression data by solving a network flux to maximize the number of enzymes that are highly expressed and catalyze flux-carrying reactions. For a better quality reconstruction, Jerby et al. [179] developed a Model Building Algorithm (MBA), which determines the active

"core reactions" with multiple sources of information including literature, transcriptome and proteomic data. The MBA then reconstructs a consistent network (no gaps or zero-flux reactions) with all the pre-defined core reactions (evidences obtained from both literature and data), adding as many of the likely active reactions (evidences obtained only from high throughput data), and as few of the other reactions. Nevertheless, manual curation is necessary with this approach to collate and analyze the literature and high-throughput data.

4.2.2 Simulating phenotypes based on metabolic network model

Constraint based modeling and simulation based on a condition-specific human metabolic network can be used to predict the flux distribution in the network for that specific metabolic phenotype. Such *in silico* analysis have been used to generate hypotheses on the cell growth, ATP production, or the states of specific metabolic functions upon perturbation [180]. For example, modeling and simulations of a metabolic model of human kidney reconstructed with the GIMME algorithm were used to evaluate the metabolic phenotypes associated with the side-effects of a drug treatment [181]. The side effect of a particular drug is determined by its off-targets, which are the enzymes/genes that are not the therapeutic targets but nevertheless are predicted to bind and be inhibited by the drug. FBA was performed on the perturbed network where the reactions catalyzed by the off-target enzymes were inhibited by the drug, to evaluate the systematic consequences of the drug and determine if the treatment leads to deficiency in metabolic functions [181]. This systematic approach has also been applied to metabolic disorders of the liver [179] and cancer cells [182] [183], and was able to correctly identify many of the genes essential for the metabolic disorders [179], interpret metabolic state-changes (e.g. Warburg effects [183]), and predict drug targets for

the metabolic system [182]. In the cancer study [182], gene expression in cancer cell lines is analyzed to identify highly expressed metabolic enzyme-encoding genes in cancer, which are used in MBA algorithm to reconstruct a "cancer metabolic network model". FBA is then applied to predict metabolic states (cell proliferation) across different gene knock-downs. Specifically, in each prediction one could turn off the reaction associated with the gene that is knock-down and apply a FBA on the constrained model to see if the cell proliferation (represented by an objective function) is reduced. The genes predicted to be important for cancer metabolism are confirmed to be highly essential in a shRNA gene knockdown dataset which lists experimental identified cancer growth-supporting genes. FBA is also applied on non-cancer cells (reconstructed metabolic network with expression data of normal cells) to determine genes important for normal cells' metabolism. Genes that only affect cancer cells are predicted to be drug targets. Many known targets of FDA-approved metabolic anticancer drugs are re-discovered in this approach. Folger et al. [182] further simulated double gene knockdowns to explore combinations of synthetic lethal drug targets.

4.3 A novel approach to reconstructing context-dependent networks

Applications of the metabolic model (i.e. to predict a metabolic state upon perturbation on gene expression) require reconstruction of context-dependent networks, thus computational approaches to automate the generation of condition-specific models are of great interest. Current approaches assume the reactions are either absent or present according to the ONs and OFFs state of the genes, which could be problematic since a small change in the ex-

pression analysis could result in very different lists of present and absent genes/reactions, which thereby affects the reconstruction. For example, a gene expressed a little bit lower than the median expression level in the transcriptome of a phenotype may be classified as absent/OFF, which would eliminate the reaction it associates with. However a gene is not necessarily entirely OFF when expressed in low levels, and a reaction may not necessarily be completely abandoned just because an enzyme catalyzing the reaction is expressed at a low level. The assumption to associate genes and reactions is questionable, thus current approaches have relied on the stoichiometric constraints to mitigate the potential inconsistencies. A recent study by Colijn et al. [184] associated gene expression levels with the constraints on the reactions, i.e. the lower a gene is expressed, the lower the flux that could be conducted through its corresponding reactions:

$$\max_{V} Z = c^T V, \text{ subject to } SV = 0, \; x_j \leq V_j \leq y_j \tag{4.3}$$

In which x_j and y_j are determined by gene expression level. Although it could be more sensitive to changes (noises) in gene expression by directly associating gene expression level with flux bounds, the approach is less dependent on the discretization of gene expression and the present/absent call of the reactions. It has been applied to bacteria to predict metabolic modulators and responses to different drugs [184].

Inspired by this study [184], we develop a novel approach to reconstruct human context-dependent metabolic network to associate gene expression *changes* with constraints on the reactions. The model is based on two assumptions that differ from previous approaches:

1. The information on the context-dependent gene expression, e.g. microarray data, tissue specific gene database, studies of gene expression in the literature, provide information

on the relative **gene expression state *change*** between phenotypes, rather than the "expression level/state" of a gene. A single "expression value" cannot indicate a gene's expression level (or ON/OFF state) without a comparison involving different phenotypes, because for different genes, the amount of expression that is required to turn them ON should be very different. For example, signaling molecules could have large impact when they are expressed in a small amount, while housekeeping genes may require a constitutively large amount of expression to support cell survival, thus the expression level of these two types of genes (i.e. signaling molecules: LOW, housekeeping: HIGH) within the samples for a given phenotype do not correlate with their ON/OFF states.

2. The state-change of the genes/enzymes, as indicated by the context information, determines the state-change of reactions, by modulating the *maximum capacity* of a reaction tunnel. This is different from previous approaches that associate gene expression with reaction states by either removing a reaction when the corresponding gene expression is low, or correlating the flux through the reaction with gene expression level. Our assumption is based on the enzyme kinetics (Michaelis-Menten equation) which shows that the maximum reaction rate (i.e., the maximum capacity of flux through a reaction tunnel) is positively correlated with the concentration/activity of the gene/enzyme that catalyzes the reaction:

$$V_{max} = K_{cat}[E] \tag{4.4}$$

In which the K_{cat} is a reaction constant and $[E]$ is the concentration of the enzyme. Although the constant K_{cat} is less available and could be different in different enzymes, our first assumption compares the maximum capacity of fluxes in two conditions

V_{new}/V_{ref}, thus we have:

$$\frac{V_{new}}{V_{ref}} = \frac{[E_{new}]}{[E_{ref}]} \tag{4.5}$$

Where K_{cat} cancels (assuming the enzyme function does not change, i.e. no significant mutations occurred, which holds in perturbation applications of a given cell type or a given patient in a relatively short time-scale). Thus, the upper bounds of the reactions in the context dependent network can be computed by V_{ref} based on a reference network, and the $[E_{new}]/[E_{ref}]$ represents the state-change (fold change) in gene expression.

Our approach is designed to reconstruct metabolic network that is specific for a phenotype/condition, by incorporating context information of the gene expression on the particular phenotype of interest. The process is shown in Figure 4.2 on page 123. Based upon a reference metabolic network that have been defined (e.g. the generic human metabolic network) and the context information, we estimate the activity change of each reaction to adjust the upper bounds (i.e. the maximum capacity) on the reaction tunnel, which results in a new metabolic network model that is specific to the condition. There are two steps in the algorithm:

1. **Compute the states of the genes under the condition of interest:** the context information is used to determine the change of gene expression between the reference condition and the condition of interest. For example, based on microarray data, the state of a gene changes when it is differentially expressed, and the amount of changes $\delta(S_g)$ can be computed based on the "log fold-change" obtained from microarray analysis. Qualitative information from the literature or tissue specific dataset directly determines the state of a gene under a condition, but the amount of state-change is

computed by subtracting the reference state from the state under the condition, in order to determine the state of the reactions in the next step. In these cases the state of the genes is represented by discrete variables $S_g = \{-1, 0, 1\}$ (i.e. low, median, high), and the state-changes (plus 1: unregulated; minus 1: downregulated) depend on the context information. When we start from the generic human metabolic network, the reference condition is defined by having all the genes initially in state 1 and the initial bounds are computed with Flux Variance Analysis.

2. **Compute the states of reactions:** based on the GPR (Gene-Protein-Reaction) information in the network, genes that are components of an enzyme complex have the "AND" relationship, thus the change of the activity of these complexes are defined by $\delta(S_c) = \min\{\delta(S_{g1}), \delta(S_{g2}), \ldots, \delta(S_{gj})\}$ in which $\delta(S_{gj})$ are the state-change of gene component j in the complex c. Complexes/genes that are different isozymes for the same reaction have the "OR" relationship, thus the amount of state-change on a reaction is then defined by the state-changes of the isozymes: $\delta(S_r) = \max\{\delta(S_{c1}), \delta(S_{g2}), \ldots\}$. Finally, the upper bound of a reaction is changed: $V_{new} = V_{old} * 2^{\delta(S_r)}$ since the state-changes are defined as the "log fold change".

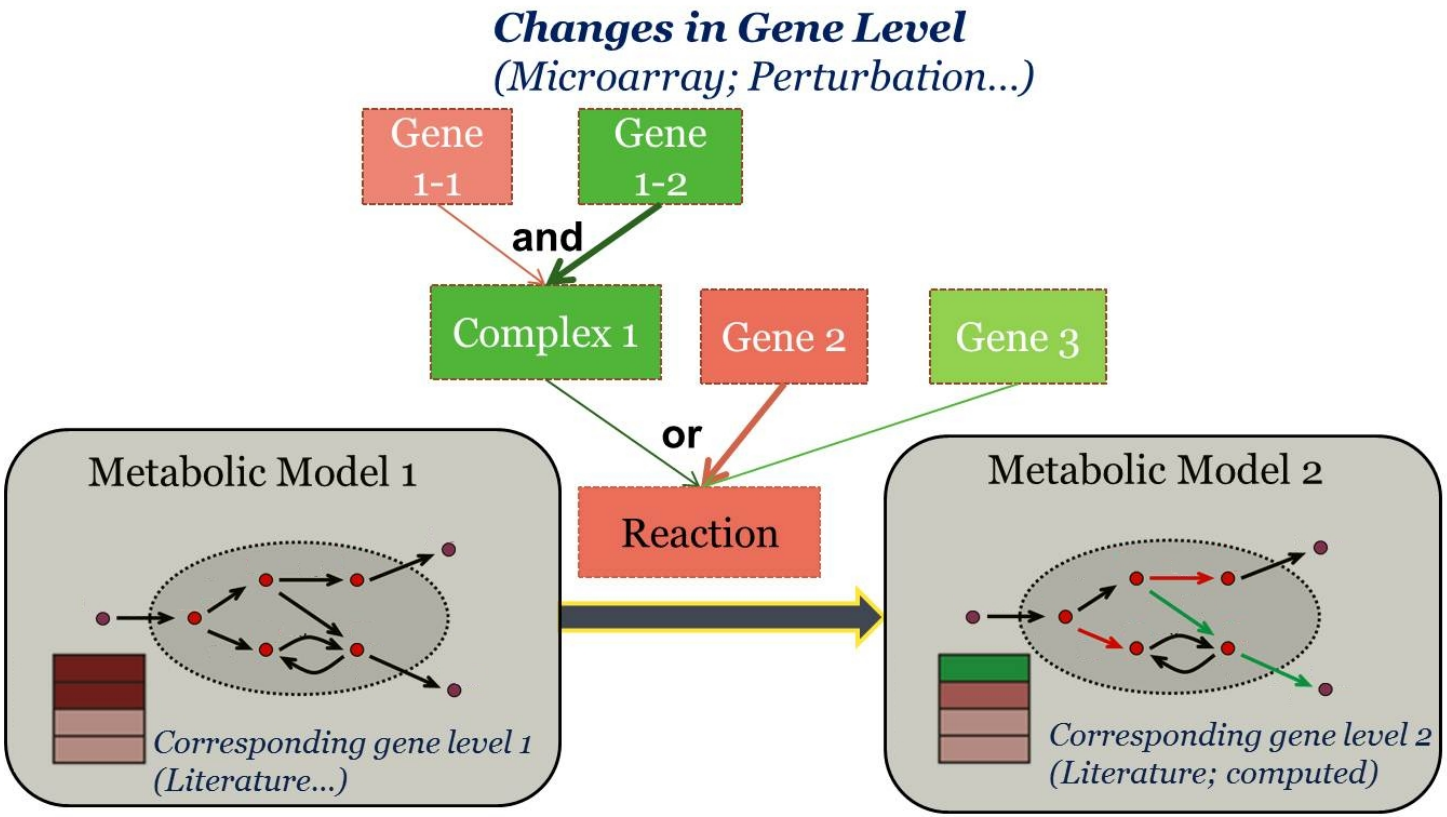

Figure 4.2: Reconstructing context specific metabolic network. A gene expression phenotype is associated with a metabolic state. The context information provides the state-change of the gene expression between two phenotypes. The state-change of each reaction is then determined by the state-change of the genes that regulate the reaction.

We apply this novel approach to model metabolic network of human liver cancer and predict microRNAs that can inhibit cancer growth.

4.4 Prediction of therapeutic microRNA based on condition specific metabolic network

MicroRNA expression has been found to be deregulated in human cancer [185]. For example, Calin et al. [186] observed down-regulation of miRNAs mir-15 and mir-16 in most lymphocytic leukemia patients, which suggests that miRNAcould be involved in cancer. He et al. [187] were the first to identify a potential non-coding oncogene, miR-17-92, which promotes c-Myc-induced tumorigenesis in mice. In another study, Johnson et al. [188] discovered a tumor suppressor miRNA let-7 which inhibits expression of the oncogene RAS in lung cancer cell lines.

A global decrease in miRNA levels has been observed in human cancers [189,190]. Knockout of the miRNA processing enzymes Drosha and Dicer enhances cancer cell growth in vitro and their invasiveness in mice [191], which confirms that the widespread reduction in miRNA expression could promote tumorigenesis. Therefore, miRNAs may have an intrinsic function in tumor suppression, and could be alternative therapeutic targets. Synthetic miRNA can be introduced into mammalian systems [192], and a pioneer study of therapeutic miRNA delivery of miR-26s in HCC (hepatocellular cancer) mice model successfully inhibited tumor cell proliferation and induced cancer-specific apoptosis [193].

Although miRNAs may be good alternative targets for cancer treatment, it has been difficult to identify which miRNAs to target for a particular type of cancer, since the un-

derlying mechanisms of why and how miRNAs are involved in cancer are largely unknown. Experimental evidences of how miRNAs regulate their targets in cancer cells have been limited. The expression of single targets, such as RAS [188], or E2F1 [194], has been shown to be regulated by miRNAs in explaining their association with cancer. However, each miRNA may regulate various target genes, and the same miRNA may have oncogenic or anti-tumorgenic activity depending on the context or cell type in which the targets are expressed. Current computational studies focus primarily on analyzing miRNA expression profiles to identify signatures that can separate particular cancer from normal samples [195], or combining miRNA and gene expression data to identify a "context-specific" target in the cancer samples [196] Nevertheless it remains unclear whether the altered miRNA expressions are the cause or consequence of the carcinogenesis processes, and which miRNAs could be good targets for treatment.

We propose a different approach to tackling this problem, by integrating miRNA target prediction, metabolic modeling and the context-specific gene expression data. We apply our novel approach of metabolic network reconstruction to reconstruct a context-specific metabolic system for human liver, and flux analysis based on the system is used to predict the miRNAs whose overexpression or delivery could inhibit cancer cell growth by inhibiting its target metabolic genes.

4.4.1 Prediction of the metabolic state in liver cancer cells upon perturbation of gene expression induced by miRNAs

Information of human liver specific gene expression is obtained from a curated dataset in the literature [178]. Our approach is then applied to reconstruct a liver cancer specific

metabolic system from the human generic metabolic network, and use the system to predict the growth and proliferation rate of cancer cells. Cancer cells are assumed to modulate their metabolic functions to be able to support their rapid growth and proliferation, thus a biomass production function was defined as their *objective function* based on experimentally measured DNA/RNA/Amino acids/Lipid composition in cancer cells [183]. We use FBA to optimize this biomass function in the liver specific metabolic network, and the biomass production rate reflects the maximum rate of cell growth and proliferation that can be achieved.

Metabolic gene targets of each microRNA are obtained from TargetScan, which is based on sequence complementarity and conservation of the targeting sites within vertebrates [197], and only those 153 conserved miRNA family with conserved binding sites (across mammalians) are included in our analysis. The gene targets are assumed to be inhibited (50% knock down) upon overexpression or therapeutic delivery of their miRNA regulator, since miRNA would bind to these target mRNAs by base pairing and forms RNA-induced silencing complex which then causes inhibition of protein translation and/or degradation of the mRNAs [198]. A conditional specific metabolic system is constructed upon overexpression of each miRNA with their targets turned off and simulated to obtain their maximum achievable biomass production rate F, which is compared with the rate F_0 in "wild-type" liver cancer without the miRNA up-regulation. A score $F_0 - F$ is computed for each miRNA to indicate their ability to reduce cancer growth. The miRNAs are ranked by this score. The procedure is shown in Figure 4.3 on page 127.

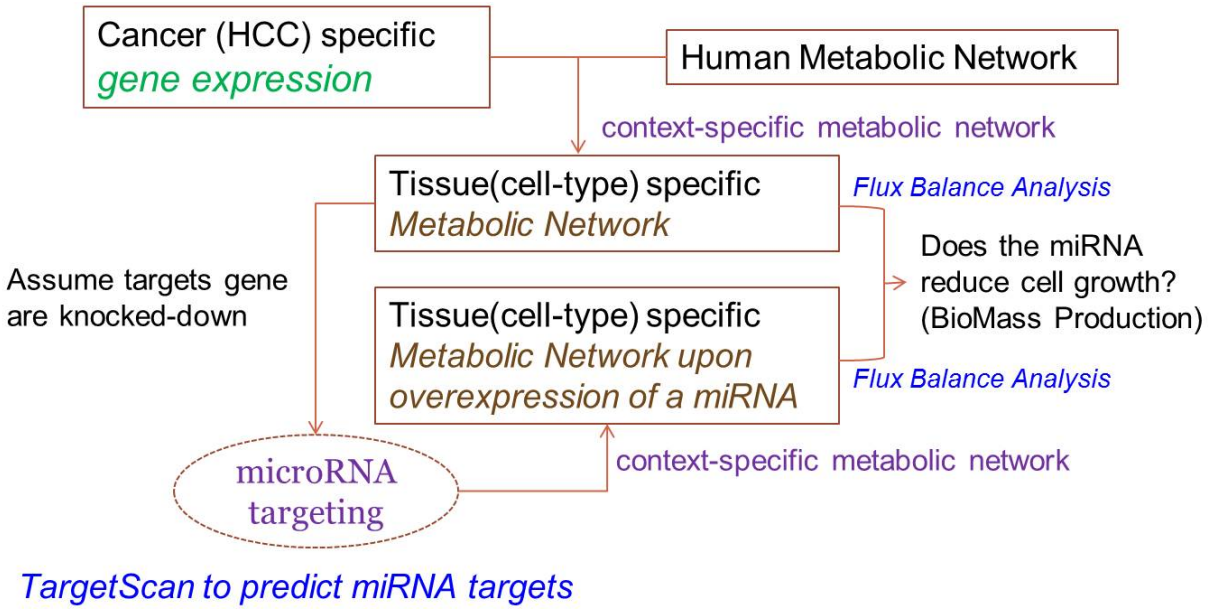

Figure 4.3: A pipeline for predicting therapeutic miRNAs for human liver cancer.

To assess the accuracy of the prediction, we search in the literature for the 153 miRNA families and collate a test set including 41 miRNAs that have been experimentally studied in liver cancer, in which there are *in vivo* or *in vitro* evidences showing 23 of these inhibits liver cancer growth/metastasis if over-expressed while 18 do not. Based on our scoring and ranking of miRNAs, we plot the ROC curve for the predictions of these 41 miRNAs in the test set (the black line in Figure 4.4 on page 129), and the result shows above 82% accuracy (Area Under Curve). To compare, we build a similar prediction system but using the GIMME approach for network reconstruction that was previously applied to human metabolism (the green line in Figure 4.4 on page 129). The accuracy of GIMME is 64%, only slightly better than random ($23/41 \approx 56\%$), and is much lower than the predictions based on our novel approach, for example, with the same cut-off with a false positive rate of 0.33, our approach achieves 0.91 precision (true positive rate) which is much higher than GIMME (0.57).

Current studies in the role of microRNAs in liver cancer (reviewed in [199]) have been focused on their targets in the signaling process, involving the apoptosis pathway (e.g. Bcl-w, Ras), cell cycle progression and migration/invasion signaling (e.g. CDK, cyclins, PI3K signaling PTEN, c-Met, FOS). From an alternative perspective, our predictions are based on metabolic system thus the results suggest that miRNAs could regulate cancer growth by modulating metabolic functions directly. Thus we further use the reconstructed metabolic system for liver cancer to explore the potential mechanisms by which miRNAs inhibit cancer growth.

Metabolic functions modulated by miRNAs to inhibit cancer growth To study the metabolic functions that miRNAs modulate to inhibit cancer growth, we average the flux change induced by the top 50 miRNAs (predicted to inhibit cancer growth) and by the bot-

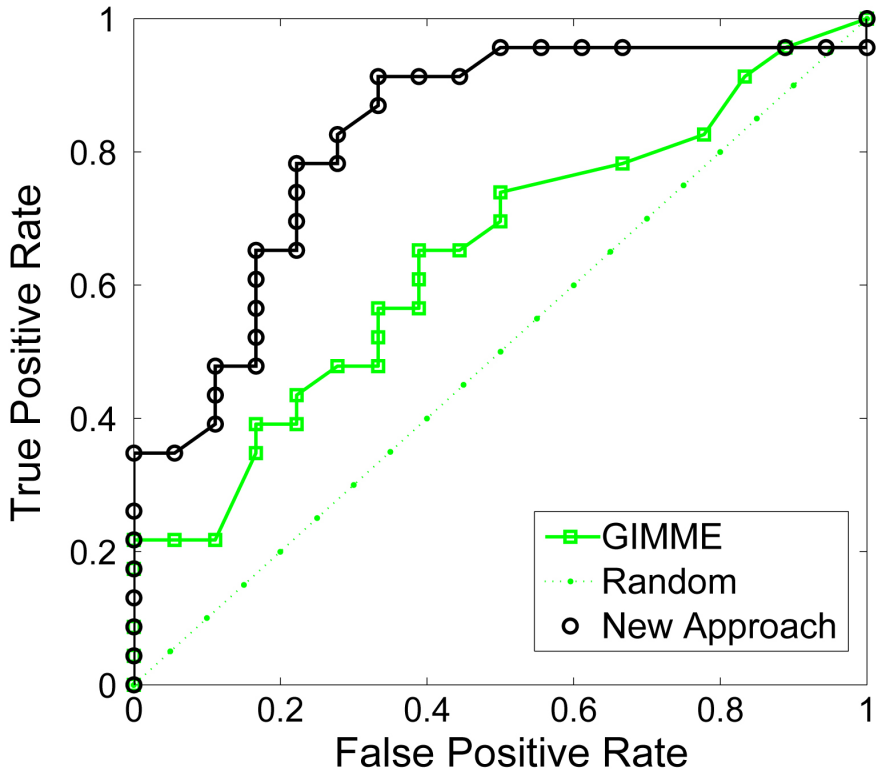

Figure 4.4: The ROC curve of the prediction based on a test-set. Black line shows the prediction result from our approach of network reconstruction, and the gree line shows the result when we apply GIMME.

tom 50 (predicted not to inhibit cancer growth) ranked based on their scores, and compared them to identify the subsystems that change the most. We found that the largest flux changes occured for the production of nucleotides and amino acids, the glycolysis/gluconeogenesis system, Citric Acid Cycle and pyruvate metabolism, as well as the transport processes across the cell membrane, mitochondrial and peroxisomal membranes. These are essential metabolic functions that support the biosynthetic processes and energy requirements of cells. Tumors specific metabolic machinery regulates these processes to facilitate cell growth and proliferation [157], and we show that these miRNAs could alter processes that inhibit cancer growth.

One of the most profound cancer biochemical phenotype is the *Warburg effect* [156]. Cancer cells metabolize glucose at high rates, and shift the flux downstream of glucose from mitochondrial tricarboxylic acid cycle to rapid anerobic glycolysis, thereby producing vast amounts of lactate that is secreted from the cancer cells [157]. The Warburg effect is known as a metabolic adaptive response in cancer cells to satisfy the high demand of the molecular building blocks of the cell, i.e. nucleotides, fatty acids, lipids and amino acids as well as ATP for facilitating proliferation [157, 183]. An indicator of the Warburg effect is the excessive lactate production, thus we tested the flux change in response to miRNA perturbations on the reaction catalyzed by lactate dehydrogenase which concerts pyruvate into lactate. The result (Figure 4.5 on page 132) shows that the flux is decreased significantly (on average by 50%) by the overexpression of the miRNAs that scored higher (top 50), while the bottom 50 miRNA have less impact on this flux, which suggests that miRNAs could inhibit liver cancer growth by mitigating the Warburg effect to alter the cancer metabolic phenotype.

miRNA target metabolic enzymes that are essential for cancer growth Each miRNA could have multiple targets, but there is no significant correlation between the number of targets in metabolism and their ability to inhibit cancer growth, nor is there significant difference in the number of targets between miRNAs predicted to inhibit cancer growth and those that are not tumor suppressors. We hypothesize that there are some important enzymes, upon targeting, whereby the cell growth/proliferation is reduced. We apply our reconstruction and simulation approach to estimate the biomass production change upon knocking down each metabolic gene individually, and identified 48 genes (among more than 1900 metabolic genes) that are predicted to be essential for the growth and proliferation of liver cancer cells, in which 24 of them can be targeted by miRNAs. We show these genes in Table 4.1 on page 133. We search for genome-wide pooled shRNA screen data [200] that is available in breast cancer, pancreatic and ovarian cancer cell lines and found that many genes in our list, including PYGB, GBE1, SCD, Enolase, pyruvate kinase and solute carrier family proteins have been identified as essential genes in more than 2 cancer cell lines wherein their knockdown by shRNA significantly reduced the survival and proliferation rate of the cells. Since there are no liver cell lines in these studies, we further looked into a collection of liver cancer gene signatures in the Liverome database [201] to determine if these genes are differentially expressed in liver cancer tissues and/or in highly invasive liver cancer cell lines. We found most of these genes have been identified as gene signatures for liver cancer and/or for the invasiveness/metastasis of liver cancer, based on previous gene expression analysis (Table 4.1 on page 133).

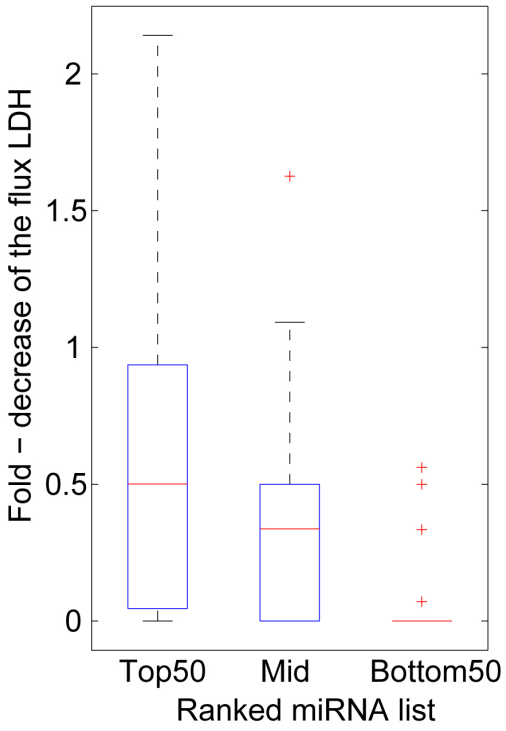

Figure 4.5: The flux change in response to miRNA perturbations on the reaction catalyzed by lactate dehydrogenase.

Gene Name	Description	Essential # cell lines	Signature cancer/normal	Signature invasiveness	text mining
PYGB	phosphorylase, glycogen	2	*	*	
PGM1	phosphoglucomutase 1		*		
GBE1	glucan (1,4-alpha-), branching enzyme 1	7			
GYG1	glycogenin 1			*	
GYS1	glycogen synthase 1 (muscle)			*	
PLA2G4A	phospholipase A2 (cytosolic, calcium-dependent)				
HADHA-B	hydroxyacyl-Coenzyme A dehydrogenase		*	*	
SCD	stearoyl-CoA desaturase (delta-9-desaturase)	2	*		
DLD	dihydrolipoamide dehydrogenase		*	*	
ENO1, ENO3	enolase 1 and 3	4	*	*	**x**
PKLR, PKM2	pyruvate kinase	7	*	*	
RPE	ribulose-5-phosphate-3-epimerase				
RPIA	ribose 5-phosphate isomerase A		*		
SLC25A13, A18, A22, A3-6	solute carrier family 25 (mitochondrial carrier: glutamate, phosphate, nucleotide)	14	*	*	
GALT	galactose-1-phosphate uridylyltransferase				

Color code for Metabolic sub-systems:		
Glycolysis	Pentose Phosphate Pathway	Glycogenolysis/Glycogenesis
Galactose pathway	Lipid/fatty acid Metabolism	Multi-functional

* based on expression level in liver cancer samples or cell lines

Table 4.1: The essential metabolic enzymes for human liver cancer that can be targeted by miRNAs

Although the evidences we found suggest these genes could be important for liver cancer, they have not been fully explored — according to text mining studies in the abstracts of pubmed literatures [201], there are only one enzyme (enolase) that have been experimentally studied and associated with liver cancer. However, some of them have been studied in other types of cancers to demonstrate their functional role in tumor growth. The glycolytic pathway is directly responsive to the Warburg effect in cancer, thus the key glycolytic enzymes, Enolases and pyruvate kinases in our list, have been shown to promote cancer invasion and proliferation [202–204]. Knockdown of pyruvate kinase PKM2 reverses the Warburg effect and suppresses tumorigenesis in mice model [202]. The high expression of Enolase correlates with poor prognosis in breast cancer and a decrease in its expression in tamoxifen-resistant breast cancer cells significantly augments the effectiveness of tamoxifen treatment [203]. The Pentose Phosphate Pathway uses glucose to generate ribose rings which are essential for the synthesis of DNA and RNA. Both of the two key enzymes which can catalyze the production of ribose-5-phosphate are on our list of essential genes: RPE and RPIA. Our simulation predicts the down regulation of either of them reduces DNA/RNA synthesis, which is consistent with a recent study in pancreatic cancer [205] that showed knockdown of either of the two enzymes (or both) reduced glucose flux into nucleotide production and suppressed tumor growth. The fatty acid metabolism and lipid synthesis are also very important in cell proliferation as they are primary components of cellular membrane. Enzyme SCD activity is involved in the synthesis of unsaturated fatty acids, and is shown as a major factor that could promote the oncogenic process and suggested to be a therapeutic target for prostate cancer [206–209]. The other enzymes HADH and PLA2G4A have been less studied in cancer metabolism. In addition to these three pathways, there are five genes involved in glycogenol-

ysis/glycogenesis predicted to be essential genes for liver cancer, as well as multifunctional genes DLD and solute carrier family proteins. These processes have not been explored in liver cancer thus far in the literature, and could be interesting targets to study experimentally. Figure 4.6 on page 136 summarizes the metabolic pathways and processes that could be affected by the essential enzymes we identified which are targeted by the miRNAs identified in our analysis.

We check the miRNAs that target these enzymes that could be essential for liver cancer. There is a strong association between targeting these enzymes and the miRNA's ability to inhibit liver cancer growth. 90% of the miRNAs ranked at the top 50 based on their score target at least one of these essential enzymes, while none of the miRNAs at the bottom 50 target any of these enzymes. Among the 90% of the miRNAs ranked at the top 50, most of them (80%) inhibit cancer growth more than by knocking-down any of the single essential enzymes that they potentially target. These observations suggest that miRNAs modulate metabolic function by simultaneously targeting metabolic enzymes that are essential for cell growth and proliferation. Table 4.2 on page 138 shows the 5 miRNAs that have the highest score and their essential metabolic gene targets predicted in our analysis, as well as previous studies showing their signaling targets and involvement in liver cancer. There are a few studies that explore the regulation of cancer metabolic functions by miRNAs (see review [210]) but currently they focus only on glycolysis, in which essential enzymes PGM1, ENO1 are shown to be regulated by miRNA-29a, miR-17-92, nevertheless it is unclear whether the miRNAs directly target these enzymes. The possible interactions between miRNA, signaling process, transcription factors and the expression of metabolic genes complicated the study in exploring function roles of miRNA in cancer. We provide an alternative modeling approach

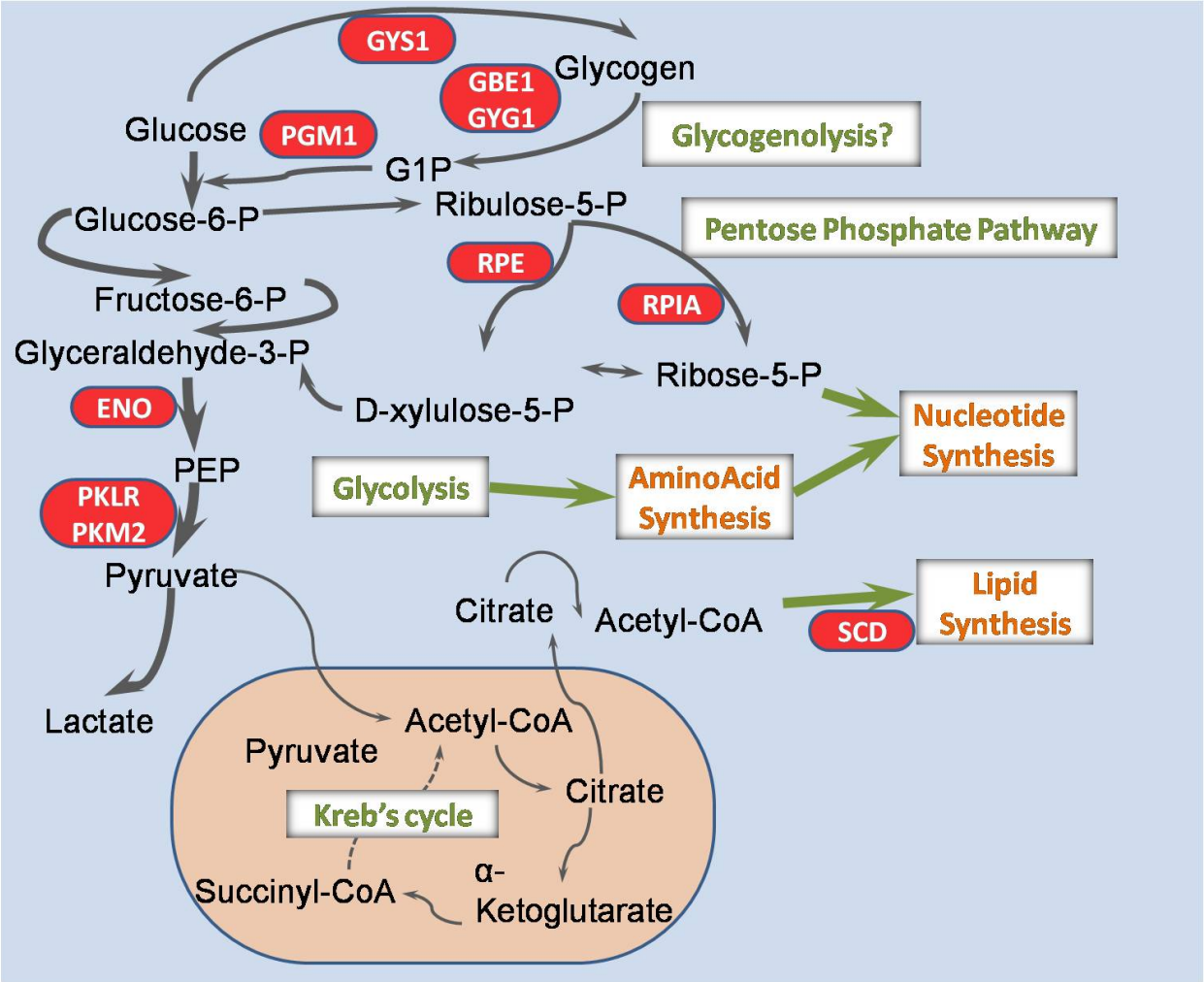

Figure 4.6: The metabolic pathways and processes that could be affected by the essential enzymes.

that focus on direct metabolic targets of miRNAs, but the limitation is that there could be indirect effects potentially responsive to signaling and transcription regulations mediated by miRNAs. Future studies in miRNAs will be aiming to incorporate more complicated gene networks in regulating the expression of metabolic genes, as well as account for the potential interactions between miRNA, signaling, transcription and metabolic systems to achieve a systematic understanding of cancer metabolism.

miRNA family	Score	Metabolic genes predicted to target	Previous studied	Expression in HCC	inhibit growth, invasion, metastasis	Target genes (direct?)	Known regulation	Ref (PMID)
miR-1/206	4.63	PLA2G4A, SCD, SLC25A22	miR-1	Down	yes, tested in many cell lines	FoxP1, MET, HDAC4	Methylation	18593903
miR-23ab	4.34	RPIA, SLC25A4, DLD	miR-23b		yes, tested only in SKHep1C3 cells	uPA, c-Met	?	19490101
miR-124/506	4.24	RPIA, GYS1, SCD, PGM1, SLC25A13, HADHA	miR-124	Down	yes, both in vivo and in vitro, phase I clinical trial starts 2012	Stat3, IL6R, CDK6?	HNF4a, Methylation	22153071 19843643
miR-122	3.68	PKM2, GYS1	miR-122	Down	yes, both in vivo and in vitro	cyclinG1, Bcl-w, p53?, ADAM17	HNFs	19617899 19296470 19584283 18291553
miR-150	1.36	GYS1	miR-150		yes, tested only in CD133+ liver cancer stem cells	c-Myb	?	22025269

Table 4.2: The 5 miRNAs ranked at the top in our analysis and their potential target genes

In summary, we developed a novel approach to simulate context dependent metabolic states upon perturbation of gene expression, which is able to incorporate the metabolic network, gene expression states to predict the steady state flux distributions in the cancer metabolic system. A condition specific metabolic system was constructed for human liver cancer (HCC) and overexpression of each miRNA was simulated to predict their effect on reducing cancer cell growth. Compared with experimental evidences that we collected from the literature, our approach achieved 80 percent accuracy in predicting miRNAs that can suppress metastasis and progression of liver cancer. Our approach can be used as a framework to explore the mechanism by which miRNAs modulate metabolic functions to affect cancer growth. We analyze the metabolic functions altered by miRNAs and identify the essential metabolic genes that are targeted by the miRNAs. We suggest that miRNAs modulate metabolic function by directly targeting metabolic enzymes that are essential for cell growth and proliferation.

Chapter 5

Conclusion

The recent advent of high-throughput technology has enabled the global analysis of genes, proteins, and their interactions, driving the development and application of computational approaches to study gene regulation on the genome scale, by reconstructing *in silico* the regulatory interactions of gene networkS. The challenge is to determine:

- What are the essential entities (genes, proteins) in a network that confer a phenotype?
- How do these entities work together to regulate biological processes?

We focus on the analysis of gene expression. Our research aim to develop novel approaches to identify genes and their regulation that are involved in conferring a phenotype. Given the complexity of gene regulatory networks, knowledge of the properties of individual components in the network are not sufficient to elucidate the cell physiology. As reported in Chapters 2-4, the research has focused on 2 aspects of this problem:

Regulatory processes are highly connected such that a specific response is typically accompanied numerous collateral effects. As also pointed out by Kholodenko et al. in [211]

> "A 'local' perturbation that is initially confined to a particular network node can propagate and cause widespread "global" changes in the network and thereby mask immediate connections and routes. This issue is particularly pertinent to large omics data sets, because even in response to a single local perturbation, the omics snapshots of the cellular state arise from a plethora of interactions spreading through cell networks."

To identify genes and interactions that are specific to a phenotype/condition, not only should we filter out the "false positives", which is the key goal of many previous studies in statistical analysis and data mining of biological data, but also, more importantly, to identify the *specific* responses rather than the many general/collateral effects. Therefore, in our network reconstruction framework, we design a multi-layer approach that is able to reconstruct condition-specific genes and their regulation through an integrative analysis of large scale information of gene expression, protein interaction and transcriptional regulation. We propose to integrate microarray data from a diverse set of conditions to provide a common context (more and better controls) for the expression behaviors of genes, and apply advanced feature selection technique, to identify the target genes that are most specific to the condition being investigated. From the target genes, conditional gene regulations, and conditional transcription factor activity are then determined. We show that incorporating these diverse conditions for comparison in the feature selection of genes and interactions enhances the specificity of the predictions.

Systematic understanding of a biological network. As a complex system, there are "emergent properties" that arise from the intricate interactions of gene networks, by which cells process external and internal signals to determine a phenotype, i.e. a state of the system. It is the interactions of the genes and proteins that determine a phenotype, thus the challenge is to understand how complex molecular and cellular functions and responses arise from these

dynamic interactions. We have been studying an important dynamic feature in the gene regulatory network—the "switch-like behavior", and propose a top-down mining approach to exploring gene switches on a genome-scale level. Our mining approach demonstrates that one can capitalize on genome-wide expression profiling to capture dynamic properties of a complex network. In another study, we aim to understand systematically the regulation of human metabolic functions. A novel approach is proposed to tackle this challenge by integrating metabolic modeling and context-specific gene expression data to simulate context dependent metabolic states upon perturbation of gene expression (e.g. induced by miRNA).

Overall these explorations in the field of computational systems biology aim to *identify genes and their regulation that determines a phenotype.* Applications on human breast cancer identified Trop2 as a target gene, a potential gene switch, and regulation of this gene by transcription factors, CREB as well as NFkB. Studies on the network reconstruction of human metabolic system predicted therapeutic miRNAs for human liver cancer, and suggested miRNAs could be implicated in the metabolic regulation of cancer. Future studies in systems biology could combine reverse engineering approaches based on the omics data and systems modeling of gene network to achieve a coherent mechanistic picture of biological regulation. For example, we have thus considered only the direct targets of miRNAs in metabolic system as discussed in Chapter 4, but there could be indirect targets whose expression is altered by TFs that are regulated by miRNAs. Therefore one could further incorporate the transcriptional regulatory network to describe regulations at this level. miRNA expression profiles and gene expression profiles upon perturbation of miRNAs, if available, could be integrated to identify the lowly expressed miRNAs for a given cancer type as well as infer the miRNA-target relationships. Through the reverse-engineering of a miRNA-target net-

work representing an integrated picture of transcriptional regulation, gene expression and miRNA expression, one would be able to determine which genes are likely the true (direct or indirect) targets of a miRNA under the condition of interest, and how the target genes could respond to the over-expression or delivery of the miRNA. This way, one could have a more realistic model of target genes' expression that respond to the over-expression of a particular miRNA, instead of assuming all predicted gene targets to be inhibited upon over-expression or therapeutic delivery of their miRNA regulators. The reverse engineering approach can be integrated into the systems modeling of metabolic network to account for the layer of transcriptional regulation, thus to achieve a better model in predicting the condition specific metabolic states upon perturbation of miRNAs. Furthermore, there are other layers of regulation of gene expression and enzyme activity, including post-transcriptional and post-translational processes. Many genes involved in post-transcriptional regulation of the metabolic genes or enzymes could also affect the metabolic states, which are excluded from current applications, and there are genetic changes that could be important in cancer phenotypes. Although it is difficult to have a comprehensive model due to limitations of our current knowledge of biological processes, nonetheless we could incorporate more information to provide a more comprehensive model of the regulatory network given the increasing amount of transcriptome data, protein interaction data, genomic data and other high-throughput data available.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] P Nurse. Reductionism and explanation in cell biology. *Novartis Found. Symp.*, 213:93–101; discussion 102–105–93–101; discussion 102–105, 1998.

[2] G D Vladutiu. Heterozygosity: an expanding role in proteomics. *Mol. Genet. Metab.*, 74(1-2):51–63, October 2001.

[3] Hui Ge, Albertha J M Walhout, and Marc Vidal. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.*, 19(10):551–560, October 2003.

[4] Yuri Lazebnik. Can a biologist fix a radio?–or, what i learned while studying apoptosis. *Cancer Cell*, 2(3):179–182, September 2002.

[5] Isabelle Guyon and Andr Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March 2003.

[6] Yvan Saeys, Iaki Inza, and Pedro Larraaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, October 2007.

[7] Cosmin Lazar, Jonatan Taminau, Stijn Meganck, David Steenhoff, Alain Coletta, Colin Molter, Virginie de Schaetzen, Robin Duque, Hugues Bersini, and Ann Nowe. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, pages –, February 2012.

[8] Sandrine Dudoit, Yee Hwa Yang, Matthew J. Callow, and Terence P. Speed. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *STATISTICA SINICA*, 12:111–139, 2002.

[9] A D Long, H J Mangalam, B Y Chan, L Tolleri, G W Hatfield, and P Baldi. Improved statistical inference from dna microarray data using analysis of variance and a bayesian

statistical framework. analysis of global gene expression in escherichia coli k12. *J. Biol. Chem.*, 276(23):19937–19944, June 2001.

[10] P Baldi and A D Long. A bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519, June 2001.

[11] Lin Deng, Jian Pei, Jinwen Ma, and Dik Lun Lee. A rank sum test method for informative gene discovery. In *KDD '04*, pages 410–419, New York, NY, USA, 2004. ACM.

[12] Rainer Breitling, Patrick Armengaud, Anna Amtmann, and Pawel Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, 573(1-3):83–92, August 2004.

[13] Sandrine Dudoit, Juliet Shaffer, and Jennifer Boldrick. Multiple hypothesis testing in microarray experiments. *U.C. Berkeley Division of Biostatistics Working Paper Series*, pages –, August 2002.

[14] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[15] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol*, 3(2):185–205, April 2005.

[16] Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[17] Patrick E Meyer, Kevin Kontos, Frederic Lafitte, and Gianluca Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J Bioinform Syst Biol*, pages 79879–79879, 2007.

[18] Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5(1):e8–e8, January 2007.

[19] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruc-

tion of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1:S7–S7, 2006.

[20] Jing Yu, V. Anne Smith, Paul P. Wang, Er J. Hartemink, and Erich D. Jarvis. Using bayesian network inference algorithms to recover molecular genetic regulatory networks. pages –, 2002.

[21] Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene van Someren, and Reinhard Guthke. Gene regulatory network inference: data integration in dynamic models-a review. *BioSystems*, 96(1):86–103, April 2009.

[22] Wei-Po Lee and Wen-Shyong Tzou. Computational methods for discovering gene networks from expression data. *Brief. Bioinformatics*, 10(4):408–423, July 2009.

[23] Timothy S Gardner and Jeremiah J Faith. Reverse-engineering transcription control networks. *Phys Life Rev*, 2(1):65–88, April 2010.

[24] Debopriya Das, Matteo Pellegrini, and Joe W. Gray. A primer on regression methods for decoding cis-regulatory logic. *PLoS Comput Biol*, 5(1):e1000269–e1000269, January 2009.

[25] Adam A Margolin and Andrea Califano. Theory and limitations of genetic network inference from microarray data. *Ann. N. Y. Acad. Sci*, 1115:51–72, December 2007.

[26] Benjamin de Bivort, Sui Huang, and Yaneer Bar-Yam. Dynamics of cellular level function and regulation derived from murine expression array data. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51):17687–17692, December 2004.

[27] Jan Ihmels, Gilgi Friedlander, Sven Bergmann, Ofer Sarig, Yaniv Ziv, and Naama Barkai. Revealing modular organization in the yeast transcriptional network. *Nat. Genet*, 31(4):370–377, August 2002.

[28] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences*, 107(14):6286–6291, March 2010.

[29] Kuang Lin and Dirk Husmeier. Modelling transcriptional regulation with a mixture of factor analyzers and variational bayesian expectation maximization. *EURASIP J Bioinform Syst Biol*, pages 601068–601068, 2009.

[30] Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet*, 34(2):166–176, June 2003.

[31] E Segal, T Raveh-Sadka, M Schroeder, U Unnerstall, and U Gaul. Predicting expression patterns from regulatory sequence in drosophila segmentation. *NATURE*, 451(7178):535–U1–535–U1, January 2008.

[32] Chiara Sabatti and Gareth M. James. Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, 22(6):739–746, March 2006.

[33] Jason Ernst, Oded Vainas, Christopher T Harbison, Itamar Simon, and Ziv Bar-Joseph. Reconstructing dynamic regulatory maps. *Mol. Syst. Biol*, 3:74–74, 2007.

[34] Michael A Beer and Saeed Tavazoie. Predicting gene expression from sequence. *Cell*, 117(2):185–198, April 2004.

[35] A P Gasch, P T Spellman, C M Kao, O Carmel-Harel, M B Eisen, G Storz, D Botstein, and P O Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, 11(12):4241–4257, December 2000.

[36] P T Spellman, G Sherlock, M Q Zhang, V R Iyer, K Anders, M B Eisen, P O Brown, D Botstein, and B Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, 9(12):3273–3297, December 1998.

[37] Pedro T. Monteiro, Nuno D. Mendes, Miguel C. Teixeira, Sofia d'Orey, Sandra Tenreiro, Nuno P. Mira, Helio Pais, Alexandre P. Francisco, Alexandra M. Carvalho, Artur B. Lourenco, Isabel Sa-Correia, Arlindo L. Oliveira, and Ana T. Freitas. Yeastract-discoverer: new tools to improve the analysis of transcriptional regulatory associations in saccharomyces cerevisiae. *Nucl. Acids Res.*, 36(suppl1):D132–136–D132–136, January 2008.

[38] Miguel C Teixeira, Pedro Monteiro, Pooja Jain, Sandra Tenreiro, Alexandra R Fernandes, Nuno P Mira, Marta Alenquer, Ana T Freitas, Arlindo L Oliveira, and Isabel S-Correia. The yeastract database: a tool for the analysis of transcription regulatory associations in saccharomyces cerevisiae. *Nucleic Acids Res*, 34(Database issue):D446–451–D446–451, January 2006.

[39] Yanxin Shi, Itamar Simon, Tom Mitchell, and Ziv Bar-Joseph. A combined expression-interaction model for inferring the temporal activity of transcription factors. pages 82–97, Singapore, 2008. Springer-Verlag.

[40] Markus J. Herrgard, Markus W. Covert, and Bernhard . Palsson. Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Research*, 13(11):2423–2434, November 2003.

[41] Bryan J Venters and B Franklin Pugh. How eukaryotic genes are transcribed. *Crit. Rev. Biochem. Mol. Biol*, 44(2-3):117–141, June 2009.

[42] Eric H. Davidson. *Genomic Regulatory Systems: In Development and Evolution*. Academic Press, January 2001.

[43] Sergey V Nuzhdin, Anna Rychkova, and Matthew W Hahn. The strength of transcription-factor binding modulates co-variation in transcriptional networks. *Trends Genet*, 26(2):51–53, February 2010.

[44] Zeba Wunderlich and Leonid A Mirny. Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet*, 25(10):434–440, October 2009.

[45] Xiao-yong Li, Stewart MacArthur, Richard Bourgon, David Nix, Daniel A Pollard, Venky N Iyer, Aaron Hechmer, Lisa Simirenko, Mark Stapleton, Cris L Luengo Hendriks, Hou Cheng Chu, Nobuo Ogawa, William Inwood, Victor Sementchenko, Amy Beaton, Richard Weiszmann, Susan E Celniker, David W Knowles, Tom Gingeras, Terence P Speed, Michael B Eisen, and Mark D Biggin. Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm. *PLoS Biol*, 6(2):e27–e27, February 2008.

[46] Duygu Ucar, Andreas Beyer, Srinivasan Parthasarathy, and Christopher T Workman. Predicting functionality of protein-dna interactions by integrating diverse evidence. *Bioinformatics*, 25(12):i137–144–i137–144, June 2009.

[47] Robert J White and Andrew D Sharrocks. Coordinated control of the gene expression machinery. *Trends Genet*, 26(5):214–220, May 2010.

[48] Maria J Amorim, Cristina Cotobal, Caia Duncan, and Juan Mata. Global coordination of transcriptional control and mrna decay during cellular differentiation. *Mol Syst Biol*, 6:380–380, 2010.

[49] AJ Hartemink, DK Gifford, TS Jaakkola, and RA Young. Combining location and expression data for principled discovery of genetic regulatory network models. pages 449, 437–449, 437, 2002.

[50] Adriano V Werhli and Dirk Husmeier. Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat Appl Genet Mol Biol*, 6:Article15–Article15, 2007.

[51] Shao-Shan Carol Huang and Ernest Fraenkel. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci Signal*, 2(81):ra40–ra40, 2009.

[52] Christopher T Harbison, D Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D Macisaac, Timothy W Danford, Nancy M Hannett, Jean-Bosco Tagne, David B Reynolds, Jane Yoo, Ezra G Jennings, Julia Zeitlinger, Dmitry K Pokholok, Manolis Kellis, P Alex Rolfe, Ken T Takusagawa, Eric S Lander, David K Gifford, Ernest Fraenkel, and Richard A Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, September 2004.

[53] Yongping Pan, Chung-Jung Tsai, Buyong Ma, and Ruth Nussinov. Mechanisms of transcription factor selectivity. *Trends Genet*, 26(2):75–83, February 2010.

[54] Douglas F Browning and Stephen J Busby. The regulation of bacterial transcription initiation. *Nat. Rev. Microbiol*, 2(1):57–65, January 2004.

[55] Lacramioara Bintu, Nicolas E Buchler, Hernan G Garcia, Ulrich Gerland, Terence Hwa, Jan Kondev, and Rob Phillips. Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev*, 15(2):116–124, April 2005.

[56] Lacramioara Bintu, Nicolas E Buchler, Hernan G Garcia, Ulrich Gerland, Terence Hwa, Jan Kondev, Thomas Kuhlman, and Rob Phillips. Transcriptional regulation by the numbers: applications. *Curr. Opin. Genet. Dev*, 15(2):125–135, April 2005.

[57] Anthony Gitter, Zehava Siegfried, Michael Klutstein, Oriol Fornes, Baldo Oliva, Itamar Simon, and Ziv Bar-Joseph. Backup in gene regulatory networks explains differences between binding and knockout results. *Mol. Syst. Biol*, 5:276–276, 2009.

[58] Zhanzhi Hu, Patrick J Killion, and Vishwanath R Iyer. Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet*, 39(5):683–687, May 2007.

[59] Igor Ulitsky and Ron Shamir. Identification of functional modules using network topology and high-throughput data. *BMC Systems Biology*, 1(1):8–8, January 2007.

[60] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, April 2005.

[61] Olaf Wolkenhauer and Mihajlo Mesarovic. Feedback dynamics and cell function: Why systems biology is called systems biology. *Mol Biosyst*, 1(1):14–16, May 2005.

[62] Steven S. Andrews and Adam P. Arkin. Simulating cell biology. *Current Biology*, 16(14):R523–R527–R523–R527, July 2006.

[63] O Wolkenhauer, S N Sreenath, P Wellstead, M Ullah, and K-H Cho. A systems- and signal-oriented approach to intracellular dynamics. *Biochem. Soc. Trans.*, 33(Pt 3):507–515, June 2005.

[64] Mihajlo D. Mesarovic, M. and Yasuhiko Takahara. *General Systems Theory: Mathematical Foundations*. Academic Press, February 1975.

[65] Xuerui Yang, Aritro Nath, Michael J Opperman, and Christina Chan. The double-stranded rna-dependent protein kinase differentially regulates insulin receptor substrates 1 and 2 in hepg2 cells. *Mol. Biol. Cell*, 21(19):3449–3458, October 2010.

[66] Fangting Li, Tao Long, Ying Lu, Qi Ouyang, and Chao Tang. The yeast cell-cycle network is robustly designed. *Proc. Natl. Acad. Sci. U.S.A.*, 101(14):4781–4786, April 2004.

[67] Maria I Davidich and Stefan Bornholdt. Boolean network model predicts cell cycle sequence of fission yeast. *PLoS ONE*, 3(2):e1672–e1672, 2008.

[68] Ming Wu, Xuerui Yang, and Christina Chan. A dynamic analysis of irs-pkr signaling in liver cells: a discrete modeling approach. *PLoS ONE*, 4(12):e8040–e8040, 2009.

[69] Gregory Stephanopoulos, Hal Alper, and Joel Moxley. Exploiting biological complexity for strain improvement through systems biology. *Nat Biotech*, 22(10):1261–1267, October 2004.

[70] Florence Jaffrezic and Gwenola Tosser-Klopp. Gene network reconstruction from microarray data. *BMC Proceedings*, 3(Suppl 4):S12–S12, 2009.

[71] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A*, 102(43):15545–15550, October 2005.

[72] Trey Ideker, Owen Ozier, Benno Schwikowski, and Andrew F Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1:S233–240–S233–240, 2002.

[73] Rui-Sheng Wang, Guangxu Jin, Xiang-Sun Zhang, and Luonan Chen. Modeling post-transcriptional regulation activity of small non-coding rnas in escherichia coli. *BMC Bioinformatics*, 10(Suppl 4):S6–S6, 2009.

[74] Celine Lefebvre, Presha Rajbhandari, Mariano J Alvarez, Pradeep Bandaru, Wei Keat Lim, Mai Sato, Kai Wang, Pavel Sumazin, Manjunath Kustagi, Brygida C Bisikirska, Katia Basso, Pedro Beltrao, Nevan Krogan, Jean Gautier, Riccardo Dalla-Favera, and Andrea Califano. A human b-cell interactome identifies myb and foxm1 as master regulators of proliferation in germinal centers. *Mol Syst Biol*, 6:377–377, June 2010.

[75] Gianluca Bontempi and Patrick E Meyer. Causal filter selection in microarray data. *ICML 2010*, pages –, 2010.

[76] Katia Basso, Adam A Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. Reverse engineering of regulatory networks in human b cells. *Nat. Genet.*, 37(4):382–390, April 2005.

[77] Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3(1):–, October 2007.

[78] Yuhang Wang and F. Makedon. Application of relief-f feature filtering algorithm to selecting informative genes for cancer classification using microarray data. pages 497–498. IEEE, August 2004.

[79] Marta Rubio-Texeira. A comparative analysis of the gal genetic switch between not-so-distant cousins: Saccharomyces cerevisiae versus kluyveromyces lactis. *FEMS Yeast Research*, 5(12):1115–1128, 2005.

[80] Ming Wu, Li Liu, and Christina Chan. Identification of novel targets for breast cancer by exploring gene switches on a genome scale. *BMC Genomics*, 12:547–547, 2011.

[81] Jrme Eeckhoute, Erika Krasnickas Keeton, Mathieu Lupien, Susan A. Krum, Jason S. Carroll, and Myles Brown. Positive cross-regulatory loop ties gata-3 to estrogen receptor expression in breast cancer. *Cancer Research*, 67(13):6477–6483, July 2007.

[82] Stephanie Harkey Shirley, Joyce E Rundhaug, Jie Tian, Noirin Cullinan-Ammann, Isabel Lambertz, Claudio J Conti, and Robin Fuchs-Young. Transcriptional regulation of estrogen receptor-alpha by p53 in human breast cancer cells. *Cancer Res.*, 69(8):3405–3414, April 2009.

[83] Mozhgan Rasti, Rita Arabsolghar, Zahed Khatooni, and Zoherh Mostafavi-Pour. p53 binds to estrogen receptor 1 promoter in human breast cancer cells. *Pathology Oncology Research: POR*, pages –, June 2011.

[84] V Doucas, G Spyrou, and M Yaniv. Unregulated expression of c-jun or c-fos proteins but not jun d inhibits oestrogen receptor activity in human breast cancer derived cells. *EMBO J.*, 10(8):2237–2245, August 1991.

[85] L M Smith, S C Wise, D T Hendricks, A L Sabichi, T Bos, P Reddy, P H Brown, and M J Birrer. cjun overexpression in mcf-7 breast cancer cells produces a tumorigenic, invasive and hormone resistant phenotype. *Oncogene*, 18(44):6063–6070, October 1999.

[86] E R Schuur, L A McPherson, G P Yang, and R J Weigel. Genomic structure of the promoters of the human estrogen receptor-alpha gene demonstrate changes in chromatin structure induced by ap2gamma. *J. Biol. Chem.*, 276(18):15519–15526, May 2001.

[87] L A McPherson and R J Weigel. Ap2alpha and ap2gamma: a comparison of binding site specificity and trans-activation of the estrogen receptor promoter and single site promoter constructs. *Nucleic Acids Res.*, 27(20):4040–4049, October 1999.

[88] George W Woodfield, Yizhen Chen, Thomas B Bair, Frederick E Domann, and Ronald J Weigel. Identification of primary gene targets of tfap2c in hormone responsive breast carcinoma cells. *Genes Chromosomes Cancer*, 49(10):948–962, October 2010.

[89] George W Woodfield, Annamarie D Horan, Yizhen Chen, and Ronald J Weigel. Tfap2c controls hormone response in breast cancer cells through multiple pathways of estrogen signaling. *Cancer Res.*, 67(18):8439–8443, September 2007.

[90] Jose Braganca, Jyrki J Eloranta, Simon D Bamforth, J Claire Ibbitt, Helen C Hurst, and Shoumo Bhattacharya. Physical and functional interactions among ap-2 transcription factors, p300/creb-binding protein, and cited2. *J. Biol. Chem.*, 278(18):16021–16029, May 2003.

[91] Jose Braganca, Tracey Swingler, Fatima I. R. Marques, Tania Jones, Jyrki J. Eloranta, Helen C. Hurst, Toshihiro Shioda, and Shoumo Bhattacharya. Human creb-binding protein/p300-interacting transactivator with ed-rich tail (cited) 4, a new member of the cited family, functions as a co-activator for transcription factor ap-2. *Journal of Biological Chemistry*, 277(10):8559–8565, March 2002.

[92] Silke Sperling, Christina H Grimm, Ilona Dunkel, Siegrun Mebus, Hans-Peter Sperling, Arno Ebner, Raffaello Galli, Hans Lehrach, Christoph Fusch, Felix Berger, and Stefanie Hammer. Identification and functional analysis of cited2 mutations in patients with congenital heart defects. *Hum. Mutat.*, 26(6):575–582, December 2005.

[93] Shangqin Guo and Gail E Sonenshein. Forkhead box transcription factor foxo3a regulates estrogen receptor alpha expression and is repressed by the her-2/neu/phosphatidylinositol 3-kinase/akt signaling pathway. *Mol. Cell. Biol.*, 24(19):8681–8690, October 2004.

[94] Kristi Muldoon Jacobs, J Daniel Pennington, Kheem S Bisht, Nukhet Aykin-Burns, Hyun-Seok Kim, Mark Mishra, Lunching Sun, Phuongmai Nguyen, Bong-Hyun Ahn, Jaime Leclerc, Chu-Xia Deng, Douglas R Spitz, and David Gius. Sirt3 interacts with the daf-16 homolog foxo3a in the mitochondria, as well as increases foxo3a dependent gene expression. *Int. J. Biol. Sci.*, 4(5):291–299, 2008.

[95] Michael W. Bronson, Sara Hillenmeyer, Richard W. Park, and Alexander S. Brodsky. Estrogen coordinates translation and transcription, revealing a role for nrsf in human breast cancer cells. *Mol Endocrinol*, 24(6):1120–1135, June 2010.

[96] M Watanabe, J Yanagisawa, H Kitagawa, K Takeyama, S Ogawa, Y Arao, M Suzawa, Y Kobayashi, T Yano, H Yoshikawa, Y Masuhiro, and S Kato. A subfamily of rna-binding dead-box proteins acts as an estrogen receptor alpha coactivator through the n-terminal activation domain (af-1) with an rna coactivator, sra. *EMBO J.*, 20(6):1341–1352, March 2001.

[97] Kwang Won Jeong, Young-Ho Lee, and Michael R Stallcup. Recruitment of the swi/snf chromatin remodeling complex to steroid hormone-regulated promoters by nuclear receptor coactivator flightless-i. *J. Biol. Chem.*, 284(43):29298–29309, October 2009.

[98] Yun Kyoung Kang, Mohamed Guermah, Chao-Xing Yuan, and Robert G Roeder. The trap/mediator coactivator complex interacts directly with estrogen receptors alpha and beta through the trap220 subunit and directly enhances estrogen receptor function in vitro. *Proc. Natl. Acad. Sci. U.S.A.*, 99(5):2642–2647, March 2002.

[99] C G Castles, S Oesterreich, R Hansen, and S A Fuqua. Auto-regulation of the estrogen receptor promoter. *J. Steroid Biochem. Mol. Biol.*, 62(2-3):155–163, June 1997.

[100] S Fan, J Wang, R Yuan, Y Ma, Q Meng, M R Erdos, R G Pestell, F Yuan, K J Auborn, I D Goldberg, and E M Rosen. Brca1 inhibition of estrogen receptor signaling in transfected cells. *Science*, 284(5418):1354–1356, May 1999.

[101] Li Chen, George Wu, and Hongkai Ji. hmchip: a database and web server for exploring publicly available human and mouse chip-seq and chip-chip data. *Bioinformatics*, pages –, March 2011.

[102] T. D. Gilmore. Introduction to nf-—[kappa]—b: players, pathways, perspectives. *Oncogene*, 25(51):6680–6684, October 2006.

[103] Simon J Galbraith, Linh M Tran, and James C Liao. Transcriptome network component analysis with limited microarray data. *Bioinformatics*, 22(15):1886–1894, August 2006.

[104] James C. Liao, Riccardo Boscolo, Young-Lyeol Yang, Linh My Tran, Chiara Sabatti, and Vwani P. Roychowdhury. Network component analysis: Reconstruction of regulatory signals in biological systems. *PNAS*, 100(26):15522–15527, December 2003.

[105] Uri Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall/CRC, July 2006.

[106] John J Tyson, Reka Albert, Albert Goldbeter, Peter Ruoff, and Jill Sible. Biological switches and clocks. *J R Soc Interface*, 5 Suppl 1:S1–8–S1–8, August 2008.

[107] James E. Ferrell and Wen Xiong. Bistability in cell signaling: How to make continuous processes discontinuous, and reversible processes irreversible. *Chaos*, 11(1):227–236, March 2001.

[108] William J Blake, Mads KAErn, Charles R Cantor, and J J Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637, April 2003.

[109] Hunter B Fraser, Aaron E Hirsh, Guri Giaever, Jochen Kumm, and Michael B Eisen. Noise minimization in eukaryotic gene expression. *PLoS Biol*, 2(6):e137–e137, April 2004.

[110] Joseph R Pomerening. Uncovering mechanisms of bistability in biological systems. *Curr. Opin. Biotechnol*, 19(4):381–388, August 2008.

[111] Benno Muller-Hill. *The Lac Operon: A Short History of a Genetic Paradigm*. Walter de Gruyter, September 1996.

[112] Wiep Klaas Smits, Oscar P. Kuipers, and Jan-Willem Veening. Phenotypic variation in bacteria: the role of feedback regulation. *Nat Rev Micro*, 4(4):259–271, April 2006.

[113] Hannah Chang, Philmo Oh, Donald Ingber, and Sui Huang. Multistable and multistep dynamics in neutrophil differentiation. *BMC Cell Biology*, 7(1):11–11, 2006.

[114] Tetsuya Kobayashi, Luonan Chen, and Kazuyuki Aihara. Modeling genetic switches with positive feedback loops. *J. Theor. Biol*, 221(3):379–399, April 2003.

[115] Patrick B Warren and Pieter Rein ten Wolde. Chemical models of genetic toggle switches. *J Phys Chem B*, 109(14):6812–6823, April 2005.

[116] Ozlem Demir and Isil Aksan Kurnaz. An integrated model of glucose and galactose metabolism regulated by the gal genetic switch. *Comput Biol Chem*, 30(3):179–192, June 2006.

[117] Anna Ochab-Marcinek. Predicting the asymmetric response of a genetic switch to noise. *J. Theor. Biol*, 254(1):37–44, September 2008.

[118] Guang Yao, Tae Jun Lee, Seiichi Mori, Joseph R. Nevins, and Lingchong You. A bistable rb-e2f switch underlies the restriction point. *Nat Cell Biol*, 10(4):476–482, April 2008.

[119] Timothy S. Gardner, Charles R. Cantor, and James J. Collins. Construction of a genetic toggle switch in escherichia coli. *Nature*, 403(6767):339–342, January 2000.

[120] Emery Conrad, Avraham E Mayo, Alexander J Ninfa, and Daniel B Forger. Rate constants rather than biochemical mechanism determine behaviour of genetic clocks. *J R Soc Interface*, 5 Suppl 1:S9–15–S9–15, August 2008.

[121] J Christopher Anderson, Christopher A Voigt, and Adam P Arkin. Environmental signal integration by a modular and gate. *Mol. Syst. Biol*, 3:133–133, 2007.

[122] J E Ferrell and E M Machleder. The biochemical basis of an all-or-none cell fate switch in xenopus oocytes. *Science*, 280(5365):895–898, May 1998.

[123] Tetsuya Shiraishi, Shinako Matsuyama, and Hiroaki Kitano. Large-scale analysis of network bistability for human cancers. *PLoS Comput Biol*, 6(7):e1000851–e1000851, July 2010.

[124] Nick I. Markevich, Jan B. Hoek, and Boris N. Kholodenko. Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. *The Journal of Cell Biology*, 164(3):353–359, February 2004.

[125] Ertugrul M Ozbudak, Mukund Thattai, Han N Lim, Boris I Shraiman, and Alexander Van Oudenaarden. Multistability in the lactose utilization network of escherichia coli. *Nature*, 427(6976):737–740, February 2004.

[126] Beat P. Kramer and Martin Fussenegger. Hysteresis in a synthetic mammalian gene network. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9517–9522, July 2005.

[127] P. M. Hartigan. Algorithm as 217: Computation of the dip statistic to test for unimodality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 34(3):320–325, January 1985.

[128] Alexander L Muratov and Oleg Y Gnedin. Modeling the metallicity distribution of globular clusters. *1002.1325*, pages –, February 2010.

[129] Joseph R. Nevins. The rb/e2f pathway and cancer. *Human Molecular Genetics*, 10(7):699–703, April 2001.

[130] Sui Huang, Gabriel Eichler, Yaneer Bar-Yam, and Donald E Ingber. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys. Rev. Lett*, 94(12):128701–128701, April 2005.

[131] Helen Parkinson, Misha Kapushesky, Nikolay Kolesnikov, Gabriella Rustici, Mohammad Shojatalab, Niran Abeygunawardena, Hugo Berube, Miroslaw Dylag, Ibrahim Emam, Anna Farne, Ele Holloway, Margus Lukk, James Malone, Roby Mani, Ekaterina Pilicheva, Tim F Rayner, Faisal Rezwan, Anjan Sharma, Eleanor Williams, Xiangqun Zheng Bradley, Tomasz Adamusiak, Marco Brandizi, Tony Burdett, Richard Coulson, Maria Krestyaninova, Pavel Kurnosov, Eamonn Maguire, Sudeshna Guha Neogi, Philippe Rocca-Serra, Susanna-Assunta Sansone, Nataliya Sklyar, Mengyao Zhao, Ugis Sarkans, and Alvis Brazma. Arrayexpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*, 37(Database issue):D868–872–D868–872, January 2009.

[132] Keith A. Ashman, Christina M. Bird, and Stephen E. Zepf. Detecting bimodality in astronomical datasets. *The Astronomical Journal*, 108:2348–2348, December 1994.

[133] Laura C Collins, Maria L Botero, and Stuart J Schnitt. Bimodal frequency distribution of estrogen receptor immunohistochemical staining results in breast cancer: an analysis of 825 cases. *Am. J. Clin. Pathol*, 123(1):16–20, January 2005.

[134] Pedro de Souza Rocha Simonini, Achim Breiling, Nibedita Gupta, Mahdi Malekpour, Mahmoud Youns, Ramesh Omranipour, Fatemeh Malekpour, Stefano Volinia, Carlo M Croce, Hossein Najmabadi, Sven Diederichs, Ozgr Sahin, Doris Mayer, Frank Lyko, Jrg D Hoheisel, and Yasser Riazalhosseini. Epigenetically deregulated microrna-375 is involved in a positive feedback loop with estrogen receptor alpha in breast cancer cells. *Cancer Res*, 70(22):9175–9184, November 2010.

[135] D Harari and Y Yarden. Molecular mechanisms underlying erbb2/her2 action in breast cancer. *Oncogene*, 19(53):6102–6114, December 2000.

[136] Ali Naderi, Ji Liu, and Glenn D Francis. A feedback loop between bex2 and erbb2 mediated by c-jun signaling in breast cancer. *Int. J. Cancer*, pages n/a–n/a–n/a–n/a, 2011.

[137] Elizabeth A. Musgrove and Robert L. Sutherland. Biological determinants of endocrine resistance in breast cancer. *Nat Rev Cancer*, 9(9):631–643, 2009.

[138] Rebecca B Riggins, Randy S Schrecengost, Michael S Guerrero, and Amy H Bouton. Pathways to tamoxifen resistance. *Cancer Lett*, 256(1):1–24, October 2007.

[139] Linxia Zhang, Linsey C Seitz, Amy M Abramczyk, Li Liu, and Christina Chan. camp initiates early phase neuron-like morphology changes and late phase neural differentiation in mesenchymal stem cells. *Cell. Mol. Life Sci*, 68(5):863–876, March 2011.

[140] M Lipinski, D R Parks, R V Rouse, and L A Herzenberg. Human trophoblast cell-surface antigens defined by monoclonal antibodies. *Proc. Natl. Acad. Sci. U.S.A*, 78(8):5147–5150, August 1981.

[141] Dominic Fong, Gilbert Spizzo, Johanna M Gostner, Guenther Gastl, Patrizia Moser, Clemens Krammel, Stefan Gerhard, Michael Rasse, and Klaus Laimer. Trop2: a novel prognostic marker in squamous cell carcinoma of the oral cavity. *Mod. Pathol*, 21(2):186–191, February 2008.

[142] D Fong, P Moser, C Krammel, J M Gostner, R Margreiter, M Mitterer, G Gastl, and G Spizzo. High expression of trop2 correlates with poor prognosis in pancreatic cancer. *Br J Cancer*, 99(8):1290–1295, 2008.

[143] Takahiro Ohmachi, Fumiaki Tanaka, Koshi Mimori, Hiroshi Inoue, Katsuhiko Yanaga, and Masaki Mori. Clinical significance of trop2 expression in colorectal cancer. *Clin. Cancer Res*, 12(10):3057–3063, May 2006.

[144] G Muhlmann, G Spizzo, J Gostner, M Zitt, H Maier, P Moser, G Gastl, M Zitt, H M Mller, R Margreiter, D Ofner, and D Fong. Trop2 expression as prognostic marker for gastric carcinoma. *J. Clin. Pathol*, 62(2):152–158, February 2009.

[145] Jianbo Wang, Ryan Day, Yiyu Dong, Steven J. Weintraub, and Loren Michel. Identification of trop-2 as an oncogene and an attractive therapeutic target in colon cancers. *Molecular Cancer Therapeutics*, 7(2):280–285, February 2008.

[146] Erich Huang, Skye H Cheng, Holly Dressman, Jennifer Pittman, Mei Hua Tsou, Cheng Fang Horng, Andrea Bild, Edwin S Iversen, Ming Liao, Chii Ming Chen, Mike West, Joseph R Nevins, and Andrew T Huang. Gene expression predictors of breast cancer outcomes. *The Lancet*, 361(9369):1590–1596, May 2003.

[147] Andrew S. Goldstein, Devon A. Lawson, Donghui Cheng, Wenyi Sun, Isla P. Garraway, and Owen N. Witte. Trop2 identifies a subpopulation of murine and human prostate basal cells with stem cell characteristics. *Proceedings of the National Academy of Sciences*, 105(52):20882–20887, December 2008.

[148] Mayuko Okabe, Yuko Tsukahara, Minoru Tanaka, Kaori Suzuki, Shigeru Saito, Yoshiko Kamiya, Tohru Tsujimura, Koji Nakamura, and Atsushi Miyajima. Potential hepatic stem cells reside in epcam+ cells of normal and injured mouse liver. *Development*, 136(11):1951–1960, June 2009.

[149] M Trerotola, P Cantanelli, E Guerra, R Tripaldi, A L Aloisi, V Bonasera, R Lattanzio, R de Lange, U H Weidle, M Piantelli, and S Alberti. Upregulation of trop-2 quantitatively stimulates human cancer growth. *Oncogene*, pages –, February 2012.

[150] E Guerra, M Trerotola, A L Aloisi, R Tripaldi, G Vacca, R La Sorda, R Lattanzio, M Piantelli, and S Alberti. The trop-2 signalling network in cancer growth. *Oncogene*, pages –, May 2012.

[151] Daniel Kaplan and Leon Glass. *Understanding Nonlinear Dynamics*. Springer, April 1995.

[152] K. Kappler, R. Edwards, and L. Glass. Dynamics in high-dimensional model gene networks. *Signal Processing*, 83(4):789–798, April 2003.

[153] Atsushi Mochizuki. An analytical study of the number of steady states in gene regulatory networks. *J. Theor. Biol*, 236(3):291–310, October 2005.

[154] Aleksej Zelezniak, Tune H. Pers, Simo Soares, Mary Elizabeth Patti, and Kiran Raosaheb Patil. Metabolic network topology reveals transcriptional regulatory signatures of type 2 diabetes. *PLoS Comput Biol*, 6(4):e1000729–e1000729, April 2010.

[155] M G Vander Heiden, S Y Lunt, T L Dayton, B P Fiske, W J Israelsen, K R Mattaini, N I Vokes, G Stephanopoulos, L C Cantley, C M Metallo, and J W Locasale. Metabolic pathway alterations that support cell proliferation. *Cold Spring Harbor Symposia on Quantitative Biology*, pages –, January 2012.

[156] O WARBURG. On the origin of cancer cells. *Science*, 123(3191):309–314, February 1956.

[157] Matthew G. Vander Heiden, Lewis C. Cantley, and Craig B. Thompson. Understanding the warburg effect: The metabolic requirements of cell proliferation. *Science*, 324(5930):1029–1033, May 2009.

[158] Ralph J DeBerardinis, Julian J Lum, Georgia Hatzivassiliou, and Craig B Thompson. The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. *Cell Metab.*, 7(1):11–20, January 2008.

[159] Jason W Locasale, Alexandra R Grassian, Tamar Melman, Costas A Lyssiotis, Katherine R Mattaini, Adam J Bass, Gregory Heffron, Christian M Metallo, Taru Muranen, Hadar Sharfi, Atsuo T Sasaki, Dimitrios Anastasiou, Edouard Mullarky, Natalie I Vokes, Mika Sasaki, Rameen Beroukhim, Gregory Stephanopoulos, Azra H Ligon, Matthew Meyerson, Andrea L Richardson, Lynda Chin, Gerhard Wagner, John M Asara, Joan S Brugge, Lewis C Cantley, and Matthew G Vander Heiden. Phosphoglycerate dehydrogenase diverts glycolytic flux and contributes to oncogenesis. *Nat Genet*, 43(9):869–874, 2011.

[160] Sirkku Pollari, Sanna-Maria Kknen, Henrik Edgren, Maija Wolf, Pekka Kohonen, Henri Sara, Theresa Guise, Matthias Nees, and Olli Kallioniemi. Enhanced serine production by bone metastatic breast cancer cells stimulates osteoclastogenesis. *Breast Cancer Res. Treat.*, 125(2):421–430, January 2011.

[161] Natalie J Serkova, Jennifer L Spratlin, and S Gail Eckhardt. Nmr-based metabolomics: translational application and treatment of cancer. *Curr. Opin. Mol. Ther.*, 9(6):572–585, December 2007.

[162] C M Galmarini, F Popowycz, and B Joseph. Cytotoxic nucleoside analogues: different strategies to improve their clinical efficacy. *Curr. Med. Chem.*, 15(11):1072–1082, 2008.

[163] Bei B. Zhang, Gaochao Zhou, and Cai Li. Ampk: An emerging drug target for diabetes and the metabolic syndrome. *Cell Metabolism*, 9(5):407–416, May 2009.

[164] Joel P. Berger, Taro E. Akiyama, and Peter T. Meinke. Ppars: therapeutic targets for metabolic disease. *Trends in Pharmacological Sciences*, 26(5):244–251, May 2005.

[165] Natalie C Duarte, Scott A Becker, Neema Jamshidi, Ines Thiele, Monica L Mo, Thuy D Vo, Rohith Srivas, and Bernhard Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U.S.A.*, 104(6):1777–1782, February 2007.

[166] Hongwu Ma, Anatoly Sorokin, Alexander Mazein, Alex Selkov, Evgeni Selkov, Oleg Demin, and Igor Goryanin. The edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol*, 3:135–135, September 2007.

[167] A. Bordbar and B. O Palsson. Using the reconstructed genome?scale human metabolic network to study physiology and pathology. *Journal of Internal Medicine*, 271(2):131–141, February 2012.

[168] Tong Hao, Hong-Wu Ma, Xue-Ming Zhao, and Igor Goryanin. Compartmentalization of the edinburgh human metabolic network. *BMC Bioinformatics*, 11:393–393, 2010.

[169] Ines Thiele and Bernhard Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc*, 5(1):93–121, 2010.

[170] Gregory N. Stephanopoulos, Aristos A. Aristidou, and Jens Nielsen. *Metabolic Engineering: Principles and Methodologies*. Academic Press, October 1998.

[171] Jeffrey D Orth, Ines Thiele, and Bernhard O Palsson. What is flux balance analysis? *Nat Biotech*, 28(3):245–248, March 2010.

[172] Yoshihiro Toya, Nobuaki Kono, Kazuharu Arakawa, and Masaru Tomita. Metabolic flux analysis and visualization. *J. Proteome Res.*, 10(8):3313–3323, August 2011.

[173] Javier Delgado and James C Liao. Inverse flux analysis for reduction of acetate excretion in escherichia coli. *Biotechnology Progress*, 13(4):361–367, January 1997.

[174] Jong Myoung Park, Tae Yong Kim, and Sang Yup Lee. Constraints-based genome-scale metabolic simulation for systems metabolic engineering. *Biotechnology Advances*, 27(6):979–988, December 2009.

[175] Christoph Gille, Christian Blling, Andreas Hoppe, Sascha Bulik, Sabrina Hoffmann, Katrin Hbner, Anja Karlstdt, Ramanan Ganeshan, Matthias Knig, Kristian Rother, Michael Weidlich, Jrn Behre, and Herrmann-Georg Holzhtter. Hepatonet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Mol. Syst. Biol.*, 6:411–411, September 2010.

[176] Nathan E Lewis, Gunnar Schramm, Aarash Bordbar, Jan Schellenberger, Michael P Andersen, Jeffrey K Cheng, Nilam Patel, Alex Yee, Randall A Lewis, Roland Eils, Rainer Knig, and Bernhard Palsson. Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nat. Biotechnol.*, 28(12):1279–1285, December 2010.

[177] Scott A. Becker and Bernhard O. Palsson. Context-specific metabolic networks are consistent with experiments. *PLoS Comput Biol*, 4(5):e1000082–e1000082, May 2008.

[178] Tomer Shlomi, Moran N Cabili, Markus J Herrgrd, Bernhard Palsson, and Eytan Ruppin. Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.*, 26(9):1003–1010, September 2008.

[179] Livnat Jerby, Tomer Shlomi, and Eytan Ruppin. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol Syst Biol*, 6:401–401, September 2010.

[180] Matthew A Oberhardt, Bernhard Palsson, and Jason A Papin. Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.*, 5:320–320, 2009.

[181] Roger L Chang, Li Xie, Lei Xie, Philip E Bourne, and Bernhard Palsson. Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS Comput. Biol.*, 6(9):e1000938–e1000938, 2010.

[182] Ori Folger, Livnat Jerby, Christian Frezza, Eyal Gottlieb, Eytan Ruppin, and Tomer Shlomi. Predicting selective drug targets in cancer through metabolic networks. *Mol Syst Biol*, 7:–, June 2011.

[183] Tomer Shlomi, Tomer Benyamini, Eyal Gottlieb, Roded Sharan, and Eytan Ruppin. Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the warburg effect. *PLoS Comput Biol*, 7(3):e1002018–e1002018, March 2011.

[184] Caroline Colijn, Aaron Brandes, Jeremy Zucker, Desmond S. Lun, Brian Weiner, Maha R. Farhat, Tan-Yun Cheng, D. Branch Moody, Megan Murray, and James E. Galagan. Interpreting expression data with metabolic flux models: Predicting mycobacterium tuberculosis mycolic acid production. *PLoS Comput Biol*, 5(8):e1000489–e1000489, 2009.

[185] Stefanie Sassen, Eric A. Miska, and Carlos Caldas. Micrornaimplications for cancer. *Virchows Arch*, 452(1):1–10, January 2008.

[186] George Adrian Calin, Calin Dan Dumitru, Masayoshi Shimizu, Roberta Bichi, Simona Zupo, Evan Noch, Hansjuerg Aldler, Sashi Rattan, Michael Keating, Kanti Rai, Laura Rassenti, Thomas Kipps, Massimo Negrini, Florencia Bullrich, and Carlo M Croce. Frequent deletions and down-regulation of micro- rna genes mir15 and mir16 at 13q14 in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci. U.S.A.*, 99(24):15524–15529, November 2002.

[187] Lin He, J. Michael Thomson, Michael T. Hemann, Eva Hernando-Monge, David Mu, Summer Goodson, Scott Powers, Carlos Cordon-Cardo, Scott W. Lowe, Gregory J. Hannon, and Scott M. Hammond. A microrna polycistron as a potential human oncogene. *Nature*, 435(7043):828–833, June 2005.

[188] Steven M Johnson, Helge Grosshans, Jaclyn Shingara, Mike Byrom, Rich Jarvis, Angie Cheng, Emmanuel Labourier, Kristy L Reinert, David Brown, and Frank J Slack. Ras is regulated by the let-7 microrna family. *Cell*, 120(5):635–647, March 2005.

[189] Jun Lu, Gad Getz, Eric A Miska, Ezequiel Alvarez-Saavedra, Justin Lamb, David Peck, Alejandro Sweet-Cordero, Benjamin L Ebert, Raymond H Mak, Adolfo A Ferrando, James R Downing, Tyler Jacks, H Robert Horvitz, and Todd R Golub. Microrna expression profiles classify human cancers. *Nature*, 435(7043):834–838, June 2005.

[190] Arti Gaur, David A Jewell, Yu Liang, Dana Ridzon, Jason H Moore, Caifu Chen, Victor R Ambros, and Mark A Israel. Characterization of microrna expression levels and their biological correlates in human cancer cell lines. *Cancer Res.*, 67(6):2456–2468, March 2007.

[191] Madhu S Kumar, Jun Lu, Kim L Mercer, Todd R Golub, and Tyler Jacks. Impaired microrna processing enhances cellular transformation and tumorigenesis. *Nat. Genet.*, 39(5):673–677, May 2007.

[192] Naotake Tsuda, Kouichiro Kawano, Clay L Efferson, and Constantin G Ioannides. Synthetic microrna and double-stranded rna targeting the 3'-untranslated region of her-2/neu mrna inhibit her-2 protein expression in ovarian cancer cells. *Int. J. Oncol.*, 27(5):1299–1306, November 2005.

[193] Janaiah Kota, Raghu R. Chivukula, Kathryn A. O'Donnell, Erik A. Wentzel, Chrystal L. Montgomery, Hun-Way Hwang, Tsung-Cheng Chang, Perumal Vivekanandan, Michael Torbenson, K. Reed Clark, Jerry R. Mendell, and Joshua T. Mendell. Therapeutic microrna delivery suppresses tumorigenesis in a murine liver cancer model. *Cell*, 137(6):1005–1017, June 2009.

[194] Kathryn A O'Donnell, Erik A Wentzel, Karen I Zeller, Chi V Dang, and Joshua T Mendell. c-myc-regulated microrna modulate e2f1 expression. *Nature*, 435(7043):839–843, June 2005.

[195] Yu Liang. An expression meta-analysis of predicted microrna targets identifies a diagnostic signature for lung cancer. *BMC Med Genomics*, 1:61–61, 2008.

[196] Xionghui Zhou, Juan Liu, Changning Liu, S. Rayner, Fengji Liang, Jingfang Ju, Yinghui Li, Shanguang Chen, and Jianghui Xiong. Context-specific mirna regulation network predicts cancer prognosis. pages 225–243. IEEE, September 2011.

[197] Robin C. Friedman, Kyle Kai-How Farh, Christopher B. Burge, and David P. Bartel. Most mammalian mrnas are conserved targets of microrna. *Genome Research*, 19(1):92–105, January 2009.

[198] David P Bartel. Microrna: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, January 2004.

[199] S. Huang and X. He. The role of microrna in liver cancer progression. *British Journal of Cancer*, 104(2):235–240, 2011.

[200] Judice L. Y. Koh, Kevin R. Brown, Azin Sayad, Dahlia Kasimer, Troy Ketela, and Jason Moffat. Colt-cancer: functional genetic screening resource for essential genes in human cancer cell lines. *Nucl. Acids Res.*, pages –, November 2011.

[201] Langho Lee, Kai Wang, Gang Li, Zhi Xie, Yuli Wang, Jiangchun Xu, Shaoxian Sun, David Pocalyko, Jong Bhak, Chulhong Kim, Kee-Ho Lee, Ye Jang, Young Yeom, Hyang-Sook Yoo, and Seungwoo Hwang. Liverome: a curated database of liver cancer-related gene signatures with self-contained context information. *BMC Genomics*, 12(Suppl 3):S3–S3, 2011.

[202] Heather R. Christofk, Matthew G. Vander Heiden, Marian H. Harris, Arvind Ramanathan, Robert E. Gerszten, Ru Wei, Mark D. Fleming, Stuart L. Schreiber, and Lewis C. Cantley. The m2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature*, 452(7184):230–233, March 2008.

[203] Shih-Hsin Tu, Chih-Chiang Chang, Ching-Shyang Chen, Ka-Wai Tam, Ying-Jan Wang, Chia-Hwa Lee, Hsiao-Wei Lin, Tzu-Chun Cheng, Ching-Shui Huang, Jan-Show Chu, Neng-Yao Shih, Li-Ching Chen, Sy-Jye Leu, Yuan-Soon Ho, and Chih-Hsiung Wu. Increased expression of enolase alpha in human breast cancer confers tamoxifen resistance in human breast cancer cells. *Breast Cancer Res. Treat.*, 121(3):539–553, June 2010.

[204] Michela Capello, Sammy Ferri-Borgogno, Paola Cappello, and Francesco Novelli. -enolase: a promising therapeutic and diagnostic tumor target. *FEBS Journal*, 278(7):1064–1074, 2011.

[205] Haoqiang Ying, Alec C Kimmelman, Costas A Lyssiotis, Sujun Hua, Gerald C Chu, Eliot Fletcher-Sananikone, Jason W Locasale, Jaekyoung Son, Hailei Zhang, Jonathan L Coloff, Haiyan Yan, Wei Wang, Shujuan Chen, Andrea Viale, Hongwu Zheng, Ji-hye Paik, Carol Lim, Alexander R Guimaraes, Eric S Martin, Jeffery Chang, Aram F Hezel, Samuel R Perry, Jian Hu, Boyi Gan, Yonghong Xiao, John M Asara, Ralph Weissleder, Y Alan Wang, Lynda Chin, Lewis C Cantley, and Ronald A DePinho. Oncogenic kras maintains pancreatic tumors through regulation of anabolic glucose metabolism. *Cell*, 149(3):656–670, April 2012.

[206] Vanessa Fritz, Zohra Benfodda, Genevive Rodier, Corinne Henriquet, Franois Iborra, Christophe Avancs, Yves Allory, Alexandre de la Taille, Stphane Culine, Hubert Blancou, Jean Paul Cristol, Franoise Michel, Claude Sardet, and Lluis Fajas. Abrogation of de novo lipogenesis by stearoyl-coa desaturase 1 inhibition interferes with oncogenic signaling and blocks prostate cancer progression in mice. *Mol. Cancer Ther.*, 9(6):1740–1754, June 2010.

[207] Daniel Hess, Jeffrey W. Chisholm, and R. Ariel Igal. Inhibition of stearoylcoa desaturase activity blocks cell cycle progression and induces programmed cell death in lung cancer cells. *PLoS ONE*, 5(6):e11394–e11394, June 2010.

[208] R. Ariel Igal. Stearoyl-coa desaturase-1: a novel key player in the mechanisms of cell proliferation, programmed cell death and transformation to cancer. *Carcinogenesis*, 31(9):1509–1515, September 2010.

[209] R. Ariel Igal. Roles of stearoylcoa desaturase-1 in the regulation of cancer cell growth, survival and tumorigenesis. *Cancers*, 3(2):2462–2477, May 2011.

[210] Pankaj K. Singh, Kamiya Mehla, Michael A. Hollingsworth, and Keith R. Johnson. Regulation of aerobic glycolysis by micrornas in cancer. *Molecular and Cellular Pharmacology*, 3(3):125–134, December 2011.

[211] Boris Kholodenko, Michael B Yaffe, and Walter Kolch. Computational approaches for analyzing information flow in biological networks. *Sci Signal*, 5(220):re1–re1, April 2012.