Integrating external biological knowledge in the construction of regulatory networks from LINCS data

Xiao Liang

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science in Computer Science and Systems

University of Washington

2017

Committee:

Ka Yee Yeung

Martine De Cock

Ling-Hong Hung

Program Authorized to Offer Degree:
Computer Science and Systems

University of Washington

Abstract

Integrating external biological knowledge in the construction of regulatory networks from LINCS data

Xiao Liang

Chair of the Supervisory Committee:
Associate Professor Ka Yee Yeung
Institute of Technology

The inference of gene regulatory networks is of great interest and has various applications. The recent advances in high-throughout biological data collection have facilitated the construction and understanding of gene regulatory networks in many model organisms. However, the inference of gene networks from large-scale human genomic data could be challenging. Generally, it is difficult to identify the correct regulators for each gene in the large search space, given that the high dimensional gene expression data only provides small number of observations for each gene.

In this thesis, we propose a Bayesian approach integrating external data sources with knockdown data from human cell lines to infer regulatory gene networks. In particular, we assemble multiple data sources including gene expression data, genome-wide binding data, gene ontology and known pathways and employ a supervised learning framework to compute prior probabilities of regulatory relationships. We show that our integrated method improves the accuracy of inferred gene networks. We present our assessment results against benchmark method and data in different forms, figures, graphs and tables. We illustrate our results in two different human cell lines, and demonstrate the generality of our results.

TABLE OF CONTENTS

Page

List of Figures ii

List of Tables iv

Chapter 1: Introduction 1
1.1 Background 1
1.2 Our contribution 2

Chapter 2: Related Work 4
2.1 Data in Gene Network Inference 4
2.2 Models in Gene Network Inference 6
2.3 Data Integration 8

Chapter 3: Integration of multiple data sources for gene network inference using genetic perturbation data 9
3.1 Data 9
3.2 Methods 11
3.3 Assessment 20
3.4 Results and Discussion 21

Chapter 4: Conclusion and future work 31

Bibliography 32

LIST OF FIGURES

Figure Number Page

3.1 An overview of the approach. We first build a supervised network for a small set of gene pairs using external knowledge derived from literature and existing datasets. Then we apply machine learning method to predict the regulatory relationships across all gene pairs in LINCS L1000 data. The predicted regulatory relationships are used as the prior probabilities in our Bayesian approach to predict the posterior probabilities. 12

3.2 The histograms of expected number of regulators per target gene predicted using knockdown data in cell line A375. The prior adjustment is performed only in supervised step. Then the regulators are predicted using the priors combined with knockdown data. Threshold probability value is 0.5. A) shows the histogram of the expected number of regulators per target gene without adjustment to the prior. B) shows the histogram of the expected number of regulators per target gene with adjustment to the prior. 16

3.3 The histograms of expected number of regulators per target gene predicted using knockdown data in cell line A549. The prior adjustment is performed only in supervised step. Then the regulators are predicted using the priors combined with knockdown data. Threshold probability value is 0.5. A) shows the histogram of the expected number of regulators per target gene without adjustment to the prior. B) shows the histogram of the expected number of regulators per target gene with adjustment to the prior. 17

3.4 Precision-recall curves for cell line A375 using different data assessed with TRANSFAC and JASPAR. The results are improved by external knowledge integration with or without MCDC correction. 25

3.5 Our inferred gene network consisted of all the positive edges at a cutoff of 0.5. Each node represents a gene and each edge represents a regulatory interaction between the two genes. The arrows of edges show the direction of regulation. The width of each edge is in proportion to the inferred posterior probability that the regulatory relationship exists for the corresponding gene pair. . . . 26

3.6 Our inferred gene network consisted of all the true positive edges found in TRANSFAC and JASPAR database at a cutoff of 0.5. Each node represents a gene and each edge represents a regulatory interaction between the two genes. The arrows of edges show the direction of regulation. The width of each edge is in proportion to the inferred posterior probability that the regulatory relationship exists for the corresponding gene pair. 27

3.7 Precision-recall curves for cell line A549 using different data assessed with TRANSFAC and JASPAR. The results are improved by external knowledge integration with or without MCDC correction. 30

LIST OF TABLES

Table Number Page

3.1 This table shows the attributes in the supervised network. Different versions of the same data sources generated in different years are considered independently. For each regulator-target gene pair, we have an outcome value derived from PAZAR TF-G website. For each external knowledge source, we have an attribute storing the value derived from the source. For CCLE expression data and RNA-seq data, we calculate the correlation between the data of the paired two genes. For pathways, ChIP and Go data, we assign a positive binary value if the two genes appear to participate in the same biological process, otherwise a negative binary value. 18

3.2 This table shows the AUROC of different machine learning models in 10-fold cross validation. Models include logistic regression, SVM [1], 5-nn [2], Ada boost [3] and random forest [4]. In terms of AUROC, logistic regression and SVM are the best two models for our data. 19

3.3 Definition of contingency. 20

3.4 Assessment results compared to TRANSFAC and JASPAR. The contingency tables display the comparison results for cell line A375 before and after external knowledge integration at 0.5 and 0.95 cutoffs. MCDC correction and external knowledge integration helps improving both the p-value and precision. 22

3.5 Comparison of the rank of the first 25 edges found and match the TRANSFAC and JASPAR edgelist in cell line A375. Edges are ranked by posterior probability. The numbers represent the rankings of true positive edges (*i.e.* edges found both in our gene network and T&J edgelist) among positive edges (*i.e.* edges found in our network). The table shows the external knowledge integration helps improving results of middle-ranked edges, which makes the result more steady. 24

3.6 Assessment results compared to TRANSFAC and JASPAR. The contingency tables display the comparison results for cell line A549 before and after external knowledge integration at 0.5 and 0.95 cutoffs. P-values have been improved by external knowledge integration at both cutoffs. 28

3.7 Comparison of the rank of the first 25 edges found and match the TRANSFAC and JASPAR edgelist in cell line A549. Edges are ranked by posterior probability. The numbers represent the rankings of true positive edges (*i.e.* edges found both in our gene network and T&J edgelist) among positive edges (*i.e.* edges found in our network). The table shows the external knowledge integration helps improving results of middle-ranked edges, which makes the result more steady. 29

Chapter 1

INTRODUCTION

1.1 Background

The inference of gene regulatory networks is of great interest in recent years, especially gene network inference from large-scale data. Advances in technology have led to the generation of high-throughput biological data. Gene regulatory networks play an important role in understanding the interactions between genes and have a lot of applications. However, the inference of gene networks from high-dimensional genomic data can be challenging.

We define a gene regulatory network as a directed graph that represents the regulatory relationships between genes, in which each node represents a gene and each directed edge represents the regulatory relationship between a regulator (parent node) and a target gene (child node). Furthermore, these regulatory relationships or edges from regulators to target genes can be calibrated by probabilities representing the likelihood of such edges, especially in Bayesian approaches.

There is an extensive literature on methods for the inference of human gene regulatory networks and their applications. Some authors inferred gene network to uncover causal relationships between gene expression and disease, which could help drug discovery and development [5, 6, 7] as well as disease biomarkers [8]. Woo *et al.* [9] proposed a method to predict changes in gene expression level after drug perturbation, which offers insight into target prioritization of novel compounds. Also, gene network inference could advance the understanding of the mechanisms underlying various biological processes and figure out the interesting genes that play important roles in biological activities.

Bayesian networks are one of the most commonly used modeling approaches in gene network construction. For example, Friedman *et al.* [10] built a framework on Bayesian networks to infer interactions between genes based on multiple expression measurements. A Bayesian network is a directed acyclic graph that describes the joint probabilities of the

conditional independence between nodes. Bayesian network methods have been applied to yeast gene expression data and further extended using probabilistic graphical models [11]. Many other models have also been developed to infer gene regulatory networks, including ordinary differential equation methods (e.g. [12, 13]) and regression-based approaches (e.g. [14]). Ordinary differential equations have been used in both static and time-series gene expression data. These methods usually suffer from the curse of dimensionality, especially in the case of human data when the number of genes is large. Subsequently, dimension reduction techniques have been used, such as forward feature selection [15], singular value decompositions [16] and Principal Component Analysis [17].

Yeung and colleagues developed Bayesian regression-based network inference methods by integrating external data to yeast time series gene expression data [18, 19, 20]. Specifically, they developed a regression framework called Bayesian Model Averaging (BMA) to select predictive variables using time series gene expression in yeast. BMA methods sample each model in the ensemble to improve the accuracy of inference. In addition, they also developed a supervised learning framework to integrate multiple data sources, including genome-wide binding data, additional gene expression data, protein-protein interaction data and prior knowledge from the literature.

However, given that the high-dimensional gene expression data typically consist of limited numbers of experiments, it is difficult to identify the correct regulators for each gene in the large search space. The expected number of regulators for each target gene is relatively small compared to the whole gene set.

To help the inference of gene regulatory networks, many types of external data have been used. These external data include genome-wide binding data [21], genetic interactions data [22] and protein-protein interaction data [23, 24].

1.2 Our contribution

In this thesis, we present an approach integrating external data sources with knockdown data from human cell lines for predictive regulatory network inference. Our methods build on the previous work by Young and colleagues [25, 26] in which they developed a Bayesian regression framework to infer gene networks from knockdown expression data. Our key

contribution is to integrate multiple data sources in this extended regression framework. Specifically, we compute prior probabilities from external human data sources using a supervised learning framework. We incorporate these prior probabilities in the Bayesian regression approach to infer gene networks from knockdown data. Our results show improved accuracy of the inferred gene networks. In addition, we extend Young *et al.* [25] by applying our methods to more than one cell line (skin melanoma cell line A375 and lung cancer cell line A549).

The thesis is organized as follows. In Chapter 2, we review some previous related work and analyze concisely the advantages and disadvantages for some types of data and models which are utilized widely in gene network inference. Chapter 3 is the main part of the thesis which is devoted to the elaborating of our work on the integration of multiple data sources for gene network inference using genetic perturbation data. In the last chapter, we state our conclusions and outline some ideas for future work.

Chapter 2

RELATED WORK

2.1 Data in Gene Network Inference

2.1.1 Time-series Gene Expression Data

Various types of data have been used to infer gene networks, including both time-series and static gene expression data [27, 28]. Compared to static data, time-series data provide much additional information from sequential time points. Using time-series gene expression data, dynamic Bayesian networks considering gene expression levels from different time points allow self-loops [29, 30, 31, 32], which are not possible in Bayesian networks due to the directed acyclic graph assumption.

Although time-series gene expression data may provide useful information from which gene regulatory relationships could be derived, it could also introduce noise and redundant information which could subsequently result in a reduction of accuracy. It is also difficult to determine the optimal number of time points profiled in the experimental design and the intervals between consecutive time points, especially after taking into account the balance of data informativeness and experiment efforts [33, 34]. Interpolation approaches using measured time points have been employed to solve the problem when the number of time points are not sufficient to infer a gene network of high accuracy. Interpolation approaches not only makes the expression levels distributed more smoothly across the time points, but also handle the estimation of time derivatives of each time point which could be utilized to build ordinary differential equations (ODE) models [35, 34]. However, interpolation could not help to alleviate the curse of dimensionality of time-series data [36]. Due to the challenge of distinguishing signal from noise in time-series gene expression data and the increased dimensionality, gene network inference from time-series data is generally time and resource consuming.

Another limitation is that it is highly difficult to infer causality using time series gene ex-

pression data alone without additional data sources. Causality in gene regulatory networks is of great biological interest. In the context of gene networks, an inferred directed edge in the form of $(A \rightarrow B)$ means the following: gene A is the parent or regulator of gene B, and that if the expression level of gene A changes then we can expect the expression level of gene B will change as well. However, time-series data only provide information on statistical causality. Methods have been developed to infer directed edges by leveraging additional data sources and expert knowledge [19] or using graphical models [37]. Nevertheless, these models are limited to the inference of statistical causality.

2.1.2 Perturbation gene expression data

To derive directed gene regulatory networks, perturbation data such as over-expression and knockdown data has been used in many proposed methods [38, 39, 10]. As static expression data without any time points, perturbation data does not reflect any dynamic biological behavior over time but the experimental design could be used to derive a causal relationship. Specifically, after gene A is perturbed (*i.e.* by either knockdown or over-expression), the expression level of gene B is observed to change. Since the causal event (perturbation) is included in the experimental design, we can infer a directed edge $(A \rightarrow B)$. Knockdown data has been widely used in the literature in gene network inference. For example, Pinna *et al.* [40] showed the effectiveness of genetic perturbation expression data inferring gene networks using genetic perturbation expression data followed by graph analyses when applied to synthetic data from the DREAM4 *in silico* network challenge. Subsequently, Pinna *et al.* used non-linear ordinary differential equations (ODE) to generate additional synthetic data to optimize input parameters. Other methods applied to knockdown data include Bayesian networks [10, 41] and correlation-based Gaussian noise models [42]. In addition to accounting for causality in the experimental design, processing perturbation data is generally not too time and resource consuming.

However, the accuracy of gene network inference using knockdown data to some extent depends on the assumption that all the knocked down genes are fully suppressed in the experiments, which could potentially be difficult to accomplish in practice. The difficulties

do not only come from the limitations of experimental techniques, but also from the fact that many functional genes are not able to be completely removed or knocked out, although this problem could be partially solved by partially suppressing the target gene. Perturbation experiments also require advanced lab techniques and much more resources compared to time-series and static gene expression data. Therefore, the existing data sources generally contain relatively less knockdown data. For example, in the LINCS L1000 gene expression data most of the knockdown experiments are generated using only 8 main cell lines [43]. The lack of data sources could limit the usage of perturbation type data.

2.1.3 Combined Data

Bonneau *et al.* proposed combining time-series and knockdown gene expression data in which a regression model coupled with bi-clustering algorithms were developed and applied to infer gene networks using data from the archaeon Halobacterium. [44]. Other studies have combined static and knockdown gene expression data using ordinary differential equation (ODE) based methods with or without external knowledge integration [45, 46].

2.2 Models in Gene Network Inference

2.2.1 Ordinary Differential Equations (ODE)

As mentioned before, many different models have been used for gene network inference such as models based on ordinary differential equations (ODE) [12, 13]. By representing the model as linear or non-linear differential equations, the interaction of expression level measured from different genes could be described by variables in the equations. Linear differential equations are generally more highly abstract in terms of describing the gene interactions and could be handled by existing linear algebra methods. Non-linear differential equations are able to describe complex behaviors with the trade-off of more handling cost and strict constraints [12, 47]. Differential equations could be employed to model different types of data such as state gene expression data [48, 49] and time-series gene expression data [17]. Being approximated by other types of approaches such as Bayesian approach [25], differential equations could also allow flexible external data integration.

Although ODEs are relatively easy to handle and less time consuming, this type of methods still suffers from the curse of dimensionality given a large set of candidate genes such as human gene set. To handle this problem, dimension reduction techniques have been widely used such as forward feature selection [15], singular value decompositions [16] and principal component analysis [17].

2.2.2 Regression-based

Regression models are also widely used in gene network inference. The regression models could be built on linear differential equations [48] or other types of functions to select the variables using a regression approach. By converting the modeling process into a variable selection problem, regression-based methods could be relatively easy to handle using the existing tools and techniques but suffer from the curse of dimensionality. Commonly used regression-based methods include regularization methods and Bayesian Model Averaging (BMA). Regularization methods such as least absolute shrinkage and selection operator (LASSO) [50], least angle regression (LARS) [51] and elastic net [52] have been employed to model different types of data in gene network inference [53, 54, 55, 56]. Various types of advanced BMA (Bayesian model averaging) methods have been proposed to facilitate gene network inference. Examples include iBMA [57], ScanBMA [20], fastBMA [58] and NetworkBMA [59] for analyzing high dimensional gene expression data.

2.2.3 Bayesian Networks

Another common approach is to use Bayesian networks in gene network inference [31, 60, 11]. A Bayesian network is a flexible framework that could be applied to both continuous and discrete types of data. It allows incorporating prior knowledge [61]. Although only statistical causality is provided by a Bayesian network itself, biological causality could be implied with proper type of prior knowledge or expression levels [62, 29]. Bayesian network models usually convert the modeling process into a model selecting problem. Randomness, noise and hidden variables are generally well-handled in Bayesian network models [10]. Furthermore, the problem of over-fitting can be avoided if the Bayesian networks are inferred in proper

manners [28].

However, Bayesian network inference could be time consuming given a large set of candidate genes since it is a NP-complete problem [63, 64]. The computational complexity increases exponentially with the number of network nodes. A Bayesian network is a directed acyclic graph (DAG) providing statistical and even biological causality. But it contains no feedback loops of nodes which could have biological significance. Although as mentioned before, this constraint could be removed when the Bayesian networks are applied to time-series data.

2.3 Data Integration

Instead of using a single data source, many proposed works have incorporated external knowledge in construction of Bayesian networks and other models. For example, Le *et al.* [65] and Geier *et al.* [33] have applied Bayesian networks to synthetic data with prior knowledge. Imoto *et al.* [24] employed Bayesian networks on gene expression data with known regulatory interactions in yeast. Other works of data integration in Bayesian networks include James *et al.* [66] using gene expression data in Escherichia coli with literature knowledge, Djebbari and Quackenbush [23] integrating literature knowledge and protein-protein interaction (PPI), and Nariai *et al.* [67] integrating protein-protein interactions and pathway data. External knowledge integration methods, other than Bayesian networks, include linear differential equations [68] and non-linear differential equations [69]. Other than synthetic data and yeast gene expression data, data integration on human microarray data has also been conducted [68]. As well as single source of information integration, multiple source knowledge integration has also shown to lead to positive results in gene network inference [70, 71].

Integrating different sources of biological information can avoid the bias generated from a single data source. It also allows building on knowledge from previous literature and therefore conduct the study systematically. However, external knowledge not only provides external information but also introduces external noise. The effect of data integration depends on the quality, assumptions and biases of the external knowledge.

Chapter 3

INTEGRATION OF MULTIPLE DATA SOURCES FOR GENE NETWORK INFERENCE USING GENETIC PERTURBATION DATA

3.1 Data

The Library of Integrated Cellular Signatures (LINCS) `http://lincsproject.org` [43] is a National Institutes of Health (NIH) funded program that aims to develop comprehensive signatures of cellular states and related tools. Many types of large scale data were generated to profile changes induced by genetic and drug perturbations across human cell lines. In particular, the LINCS L1000 data generated by the Broad Institute measure the gene expression levels across approximately 1,000 landmark genes. These landmark genes were chosen to capture approximately 80% of the information for 20,000 genes in the human genome. The LINCS L1000 gene expression data are publicly available from the Gene Expression Omnibus (GEO) database with accession number GEO GSE70138 `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70138`.

The L1000 experiments were performed using the Luminex Bead technology [72] that generates high-throughput gene-expression assays using 384-well plates. To measure the expression level of specific genes, specific color-coded microspheres were coded to bind fluorophore and the corresponding RNA sequence. Therefore the expression level of each gene could be represented by the intensity of the fluorescence. For each two genes, two types of beads sharing one bead color were designed to measure the expression of the two different genes. In each perturbation experiment, about 35,000 to 50,000 beads sharing 500 bead colors were added to each well to measure the expression levels of approximately 1000 landmark genes.

The beads for a pair were mixed in approximately a 2:1 ratio. Therefore, two peaks are expected in a histogram of fluorescence levels, and these observed peaks could be deconvoluted to assign expression values to the appropriate pair of genes.

To reduce the noise from experimental conditions, there are several wells used for control on each plate. In addition, technical replicate data was generated in which the same perturbations were performed in the same wells across multiple plates.

The L1000 experiments data was generated and processed by the Broad Institute LINCS Center for Transcriptomics as part of their Connectivity Map project [73]. L1000 gene expression data processed by the Broad Institute was provided as levels 1 to 5 to the public [74]. Level 1 represents the raw unprocessed data from the Luminex Bead technology. In level 2, the gene expression values of the landmark genes were deconvoluted from the observed fluorescence levels and normalized to a set of internal standards. Subsequently, quantile normalization was performed on these landmark genes, and interpolated to all 20,000 human genes in level 3. The level 4 and level 5 data consist of the gene signatures comparing the perturbed experiments to the unperturbed experiments.

Young *et al.* [26] discovered that the deconvolution step would introduce artifacts in the data. There are three types of artifacts found and discussed in the expression data of two paired genes on the same bead color. Firstly, the two gene could be assigned the same expression value if their expression levels are not significant enough to be distinguished. Secondly, together with the quantile normalization step, the deconvolution step could generate incorrect additional clusters. Finally, sometimes the paired two genes could be assigned flipped expression values, which means gene A could be assigned the expression value of gene B and vice versa. A correction which could be applied to the data to eliminate the effect of artifacts would be discussed in a later section.

As a large-scale genomic data resource, LINCS L1000 data provides a rich set of human gene expression information. Specifically, in our work we used the knockdown experiments data. There are approximately 4500 knockdown experiments in L1000 dataset. Most of these data were generated using 8 cell lines: A375, A549, HA1E, HCC515, HEPG2, HT29, MCF7 and PC3. Data from these knockdown experiments is usually collected 96 hours after the perturbation [43].

3.2 Methods

3.2.1 Method Outline

We integrate external data in gene network inference using human gene perturbation data. The performance has been improved by the external data integration and MCDC (model-based clustering with data correction) method. The human gene expression data we use is derived from LINCS L1000 data. The details of data and MCDC are described in Methods section. Figure 3.1 shows the overview of our approach.

Our inferred gene network is directed since we infer causal relationship between gene pairs. This feature of our gene network does not come from Bayesian approach but the gene knockdown experimental design. With this biological context, we could know the observed changes in gene expression level origin from the knockdown gene. Therefore we could infer a directed edge from the knockdown gene to the affected gene.

3.2.2 BayesKnockdown

We use the `BayesKnockdown` Bioconductor package [75] to calculate posterior probabilities of regulatory relationships. This package is written in R by Dr. William Chad Young. In the previous work [25], this `BayesKnockdown` package was applied to L1000 gene expression data from a single cell line (A375). In this paper, we extend Young *et al.* [25] by integrating additional data sources and by applying the package to an additional cell line (A549).

To prepare the input data for the `BayesKnockdown` package, the LINCS L1000 knockdown experiments are first transformed by calculating z-scores to account for bias and noise among replicates:

$$x^*_{hi} = \frac{x_{hi} - \bar{x}_{hp}}{s_{hp}},$$

where x_{hi} represents the gene expression level of gene h and experiment (well) i on plate p, $\bar{x}_{hp}$ and s_{hp} are, respectively, the mean and standard deviation for gene h across all control experiments on plate p. A linear regression model is then applied to model the change in a target gene t as dependent on the change in the knockdown gene h, with $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2_\varepsilon)$ as

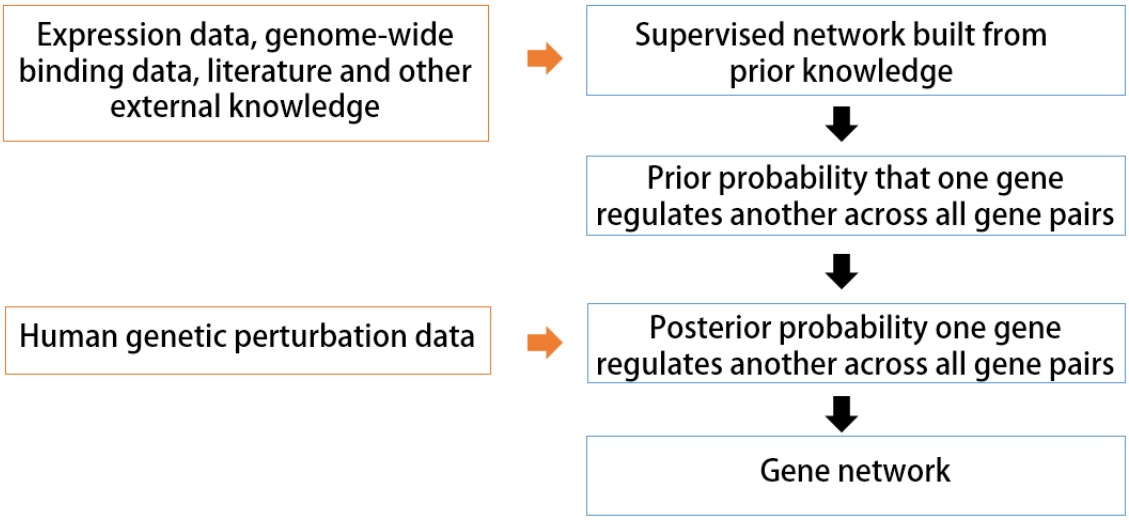

Figure 3.1: An overview of the approach. We first build a supervised network for a small set of gene pairs using external knowledge derived from literature and existing datasets. Then we apply machine learning method to predict the regulatory relationships across all gene pairs in LINCS L1000 data. The predicted regulatory relationships are used as the prior probabilities in our Bayesian approach to predict the posterior probabilities.

the error term:

$$x^*_{ti} = \beta_0 + \beta_1 x^*_{hi} + \varepsilon_i, \quad i = 1, ..., n_h.$$

In the `BayesKnockdown` package [25], the linear regression model is estimated with a Bayesian approach using Zellners g-prior [76] for the model parameters. The parameter g specifies the expected size of the regression coefficient β_1. The value of g can be estimated using an Expectation-Maximization algorithm [77, 20]. Then the regression model with g-prior is used to calculate the probability $Pr(h \to t|x)$ that gene h regulates gene t given the data x, versus the probability Pr_0 that there is no regulatory relationship:

$$\pi_{ht} = \frac{Pr(h \to t|x)}{Pr_0} = \frac{\pi_{ht}}{1 - \pi_{ht}} \exp[(n_h - 2) \log\left(1 + g\left(1 - R^2\right)\right)/2].$$

In the absence of external data sources, the prior probability of regulation π_{ht} is set to 0.0005 for all the gene pairs in Young *et al.* [25]. The value of 0.0005 is derived from prior knowledge for yeast data, reflecting the expected number of regulators per gene [78]. The coefficient of determination R^2 for the simple linear regression is calculated from the correlation of the expression data of gene h and gene t. Then we have

$$p_{ht} = \frac{Pr(h \to t|x)}{Pr(h \to t|x) + Pr_0} = \frac{\pi_{ht}}{1 + \pi_{ht}},$$

where p_{ht} is the posterior probability of the regulatory relationship between a given gene pair.

3.2.3 Data integration using supervised machine learning

We downloaded transcription factor and target gene pairs (TF-G pairs) from the PAZAR database, a public resource for transcription factor and regulatory sequence annotation [79, 80, 81]. Subsequently, we mapped the target genes in the format of Ensemble IDs to Entrez IDs using Biomart [82]. After the data processing, we kept the TF-gene pairs in which both the TFs and target genes are in the L1000 landmark genes. This resulted in a total of 232 TF-gene pairs that we labeled as positive training samples (Y=1) in our supervised framework. Due to a lack of documentation on non-regulatory TF-gene pairs, we randomly generated 240 negative training samples of TF-gene (Y=0) that are not documented in PAZAR.

After collecting the positive and negative training samples of TF-gene pairs, we derived the training data using external data sources to generate attributes in the supervised framework as described below. Table 3.1 summarizes the different types of attributes defined using external data sources in our supervised learning framework.

- Gene expression data across human cell lines. For each TF-gene pair, we compute the Pearson's correlation between TF and gene across 917 human cell lines from the Cancer Cell Line Encyclopedia (CCLE) [83]. The CCLE data is publicly available from `https://portals.broadinstitute.org/ccle/home`. As another attribute (or variable) in the training data, we also compute the Pearson's correlation between TF and gene across 675 common used human cell lines in the RNAseq data generated by Klijn *et al.* [84]. The Klijn *et al.* data is publicly available from ArrayExpress with accession number E-MTAB-2706 `http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2706/`.

- Gene ontology (GO). Gene Ontology (GO) defines a controlled vocabulary and descriptions of gene products across biological systems [85, 86]. Genes assigned to the same ontology terms are expected to share common functionalities. Intuitively, we expect regulatory TF-gene pairs to share common GO terms. Since GO terms are hierarchical in nature, we filter out large and hence, less informative GO terms. The upper boundary is set to 100. We define a binary attribute in our supervised framework: if a given TF-gene pair are both assigned to the same GO term, we define the binary variable to be 1, otherwise 0.

- Genome-wide binding data. We also use genome-wide binding data (ChIP-chip, ChIP-seq) from ENCODE [87]. The chromatin immunoprecipitation (ChIP) technology could be used to detect binding between proteins and DNA *in vivo*. Since trascriptional regulation is typically preceded by binding, we define a binary attribute for a (TF, gene) pair to be 1 if TF binds gene, and 0 otherwise. We derive these binary variables by parsing the processed ChIP data from the ENRICHR website [88].

- Pathways data. We hypothesize that a regulatory (TF, gene) pair is more likely to be assigned to the same biochemical pathways. Therefore, we define a binary variable for each of Wikipathways [89], KEGG [90], BioCarta [91] and Reactome [92, 93]. If TF and gene appear in the same pathway, we define the binary variable to be 1, otherwise 0. We derive these binary variables by parsing the processed library data from the ENRICHR website [88] `http://amp.pharm.mssm.edu/Enrichr/`.

Some attributes in our original supervised network have very similar values, which decreases the informativeness of the attributes. We remove the duplicates first by standard error values. The attributes with extremely high standard error in the logistic regression model are removed. Then the attributes are further filtered by applying machine learning models on the supervised training data, each time leaving one column out. The less informative attributes, without which there would be higher accuracy in the validation, are filtered out. We have also removed duplicates of pathways attributes using the same method. After the filtering, there are in total 14 attributes remain in our final supervised network.

After finalizing our supervised network, we perform 10-fold CV using different machine learning methods on our supervised network. Table 3.2 summarizes the results. Logistic regression and SVM are the best two models in our cross validation step. We use logistic regression to build our models in next steps.

3.2.4 Sampling bias correction

After getting the priors, we correct for sampling bias in the prior [19]. Specifically, we add an offset of $log(\pi_1/\pi_0)$ to the log odds in our logistic regression model. Here π_1 and π_0 are the sampling rates for positive and negative cases respectively in the training data. We use $\pi_1/\pi_0 = 7003.864$ from our 232 positive instances and 240 negative instances in the supervised network. Figure 3.2 shows the histograms before and after the correction in cell line A375. Figure 3.3 shows the histograms in cell line A549. Threshold probability values are set to 0.5.

The intuition of this correction is that our supervised training data is balanced to improve the prediction process. However, the positive cases are usually rare in real situation.

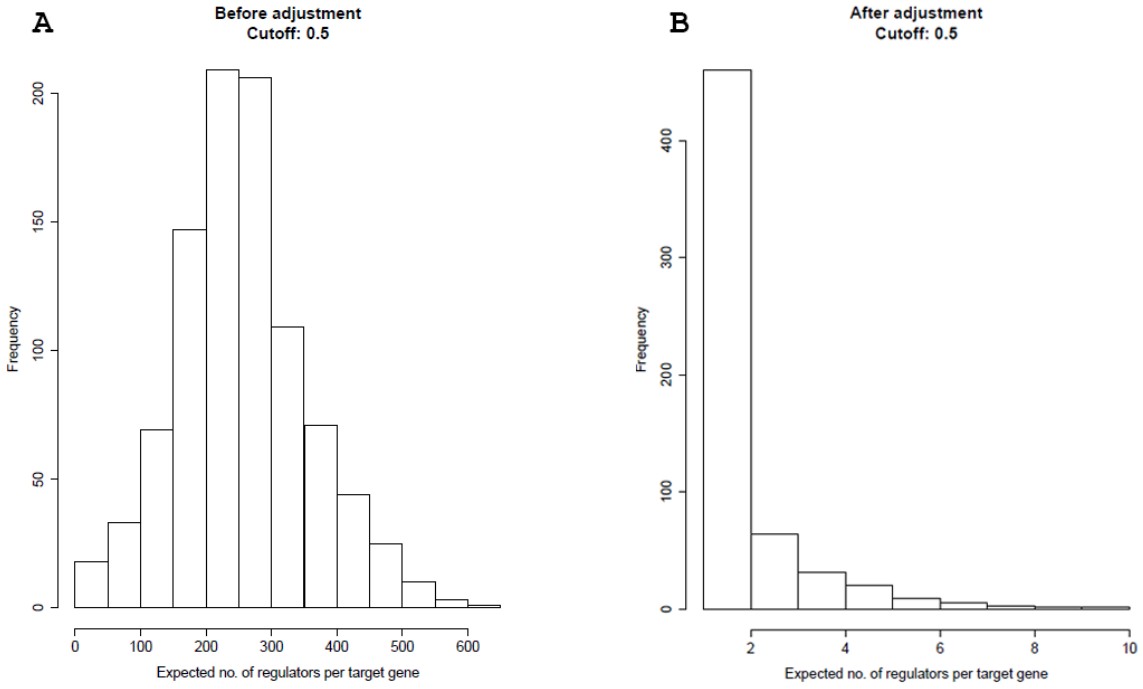

Figure 3.2: The histograms of expected number of regulators per target gene predicted using knockdown data in cell line A375. The prior adjustment is performed only in supervised step. Then the regulators are predicted using the priors combined with knockdown data. Threshold probability value is 0.5. A) shows the histogram of the expected number of regulators per target gene without adjustment to the prior. B) shows the histogram of the expected number of regulators per target gene with adjustment to the prior.

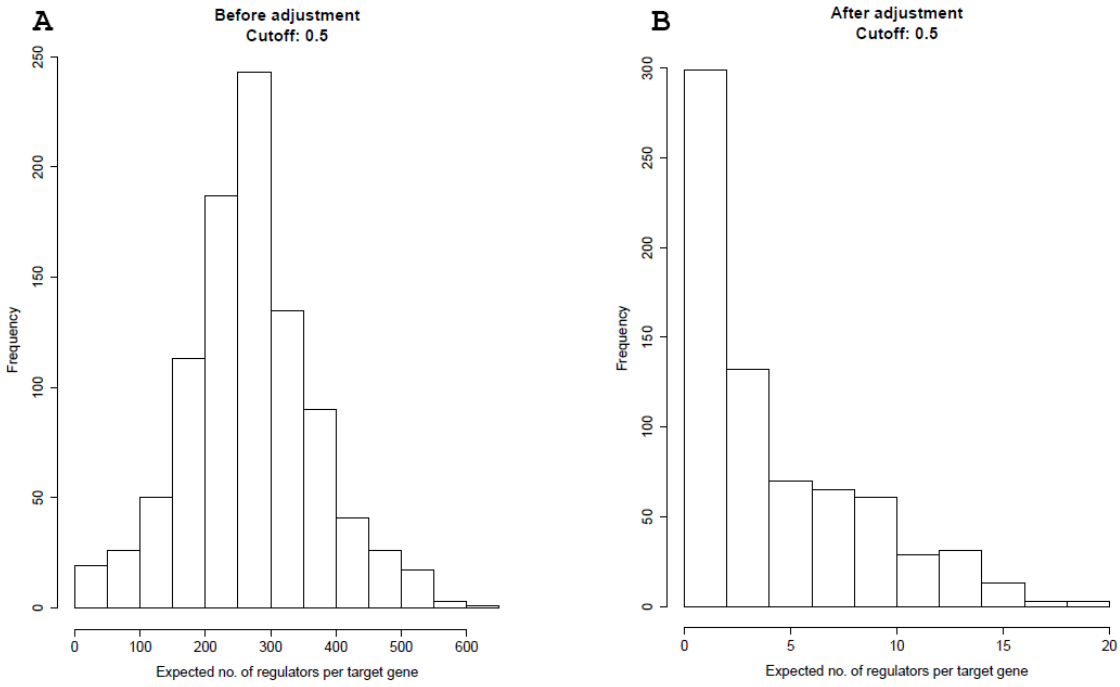

Figure 3.3: The histograms of expected number of regulators per target gene predicted using knockdown data in cell line A549. The prior adjustment is performed only in supervised step. Then the regulators are predicted using the priors combined with knockdown data. Threshold probability value is 0.5. A) shows the histogram of the expected number of regulators per target gene without adjustment to the prior. B) shows the histogram of the expected number of regulators per target gene with adjustment to the prior.

Therefore we perform this correction to better match the biological relationships between genes in practice.

3.2.5 MCDC

MCDC (Model-Based Clustering with Data Correction) is a correction intended to remove artifacts in gene expression data [26]. As mentioned before, the gene expression data was

DATA SOURCE	VARIABLE	NOTES	COEFFICIENT	PR(>|Z|)
(Intercept)	-		-4.0720	**0.000880**
CCLE	correlation	cor(TF, g)	-2.1920	**0.049747**
RNA-seq	correlation	cor(TF, g)	2.0429	0.071182
BioCarta 2013	binary	1 if common pathway 0 otherwise	-0.1954	0.927270
BioCarta 2015	binary	1 if common pathway 0 otherwise	0.7050	0.738940
KEGG 2015	binary	1 if common pathway 0 otherwise	-1.6199	0.111851
KEGG 2016	binary	1 if common pathway 0 otherwise	0.5397	0.342801
WikiPathways 2013	binary	1 if common pathway 0 otherwise	1.1687	0.210120
WikiPathways 2016	binary	1 if common pathway 0 otherwise	1.5055	**0.000244**
Reactome 2016	binary	1 if common pathway 0 otherwise	0.6304	**0.046281**
ENCODE and ChEA Consensus TFs from ChIP-X	binary	1 if known interaction 0 otherwise	1.0907	**2.02e-06**
ChEA 2013	binary	1 if known interaction 0 otherwise	1.4378	**0.000690**
ChEA 2015	binary	1 if known interaction 0 otherwise	1.6597	0.192905
ChIP chip	binary	1 if known interaction 0 otherwise	-0.4291	0.277970
GO terms	binary	1 if common GO term 0 otherwise	0.5366	0.141719

Table 3.1: This table shows the attributes in the supervised network. Different versions of the same data sources generated in different years are considered independently. For each regulator-target gene pair, we have an outcome value derived from PAZAR TF-G website. For each external knowledge source, we have an attribute storing the value derived from the source. For CCLE expression data and RNA-seq data, we calculate the correlation between the data of the paired two genes. For pathways, ChIP and Go data, we assign a positive binary value if the two genes appear to participate in the same biological process, otherwise a negative binary value.

originally paired on the bead when LINCS data was processed. In the deconvolution step during the processing, sometimes additional artifactual clusters could be generated and the expression values of the paired two genes could be reversed. This MCDC correction could help correcting such a situation by model-based clustering.

Generally, model-based clustering [94, 95, 96, 97] assumes the data comes from the distribution of a mixture of multiple components such as clusters. Each of the components could be modeled by a Gaussian distribution with parameters which could be estimated using EM (Expectation-Maximization) algorithm. MCDC extends the basic model-based clustering to detect flipped points in the gene expression data.

Instead of the original data, MCDC considers the distribution of transformed matrices. Each of the matrices represents the probability of a certain data point from the paired genes having been transformed. This method thus could be used to identify flipped data points. Furthermore, to eliminate the effect of artifactual clusters generated from expression level estimating process, the expression levels of the paired genes are estimated as the mean of the largest cluster after selecting the best model in MCDC.

The MCDC runs with different numbers of clusters ranging from 1 to an arbitrary maximum number. In this work the maximum number is set to 9. The best number of

Model	AUROC	Assessment Value
Logistic Regression	0.761564	probability
SVM	0.727529	probability
5-nn	0.7087759	probability
Ada Boost	0.668736	binary
Random Forest	0.500566	binary

Table 3.2: This table shows the AUROC of different machine learning models in 10-fold cross validation. Models include logistic regression, SVM [1], 5-nn [2], Ada boost [3] and random forest [4]. In terms of AUROC, logistic regression and SVM are the best two models for our data.

clusters as well as the best model are then selected using the BIC (Bayesian information criterion) values [97].

Our work applies MCDC to the untreated data in cell line A375 and A549. The untreated data is then used as the control to the knockdown data.

3.3 Assessment

The high-quality transcription factor binding profile (JASPAR) dataset [98] provides experimentally defined transcription factor (TF) DNA-binding sites for eukaryotes. We use the TF-gene targeting relationships in this data to assess the resulting gene networks. We use Fisher's exact test to calculate the p-values of contingency table.

A Pearsons chi-square test is applied to a 2 x 2 contingency table to represent the consistency of our constructed network and the known regulatory relationships. Table 3.3 shows an example contingency table with the definitions of TP, FP, TN and FN.

		Edge in T&J	
		Yes	No
Edge in Our Inferred network	Yes	TP	FP
	No	FN	TN

Table 3.3: Definition of contingency.

To assess our result, we use TRANSFAC and JASPAR [98] lists of edges as our reference standard. The TRANSFAC and JASPAR (T&J) edgelist contains approximately 4200 edges for 37 transcription factors that overlaps LINCS landmark genes. This is the same gold standard which has been used in Young *et al.* [25, 26]. Although T&J edgelist is limited to transcription factors that are previously well-studied, it is difficult to find a comprehensive standard for gene network assessment in mammalian systems at current stage.

3.4 Results and Discussion

3.4.1 Results: NIH LINCS Data A375

We first evaluated the performance of our network by comparing our results to T&J dataset using contingency tables. Table 3.4 shows the assessment. We used two cutoff posterior probability values 0.5 and 0.95 as thresholds for positive edges. The first two tables are computed from the network inferred using knockdown data only. The second two tables are from the network inferred using knockdown data and our external knowledge integration. The last two tables show the assessment result of network inferred from knockdown data using untreated data with MCDC correction as control, as well as external knowledge integration.

We could see the MCDC correction and integrated prior knowledge improved both the p-value and precision. By applying both MCDC correction and external knowledge integration, the p-values were improved from around 0.01 to around 0.001. Also, the precision increased from 0.14 to 0.17 at 0.5 cutoff, and 0.2 to 0.3 at 0.95 cutoff.

We then compared our predicted edges to T&J dataset by ranked lists. The assessment result is shown in Table 3.5. We first found all the edges in the intersection of our predicted edges and T&J edges. Then we ranked these found edges by posterior probabilities from our prediction in descending order. Finally we ranked all our predicted edges by posterior probabilities in descending order. For each edge also in T&J dataset, we noted down the corresponding ranking in our edgelist for the same edge. With the edge ranks we not only assessed our edgelist by the found edges, but also involved the values of posterior probabilities.

For example, for the first column Knockdown Data, the sixth edge in our edgelist is the first found edge in T&J. This number means that the top 5 edges in our edgelist are not found in T&J. These ranked lists could help us know the difference between T&J and our own edgelist. We could see the external knowledge integration helps improving results of middle-ranked edges. For example, in the third column from knockdown data and external knowledge integration, we could see the 25th found edge is ranked 127th in our edgelist. In the first column from knockdown data only, the 25th found edge is ranked 135th in our

1. KNOCKDOWN DATA

CUTOFF 0.5

T&J	YES	NO
YES	38	3100
NO	225	27488

P-VALUE: 0.01717

PRECISION: 0.14449

CUTOFF 0.95

T&J	YES	NO
YES	13	3125
NO	55	27658

P-VALUE: 0.01842

PRECISION: 0.19118

2. KNOCKDOWN DATA + SUPERVISED NETWORK

CUTOFF 0.5

T&J	YES	NO
YES	27	3111
NO	142	27571

P-VALUE: 0.01414

PRECISION: 0.15976

CUTOFF 0.95

T&J	YES	NO
YES	8	3130
NO	34	27679

P-VALUE: 0.05842

PRECISION: 0.19048

3. KNOCKDOWN DATA + MCDC UNTRT + SUPERVISED NETWORK

CUTOFF 0.5

T&J	YES	NO
YES	25	3113
NO	122	27542

P-VALUE: 0.00716

PRECISION: 0.17007

CUTOFF 0.95

T&J	YES	NO
YES	11	3127
NO	25	27688

P-VALUE: 0.0006371

PRECISION: 0.30556

Table 3.4: Assessment results compared to TRANSFAC and JASPAR. The contingency tables display the comparison results for cell line A375 before and after external knowledge integration at 0.5 and 0.95 cutoffs. MCDC correction and external knowledge integration helps improving both the p-value and precision.

edgelist. The larger difference in rankings indicates larger difference between our prediction and T&J dataset.

Figure 3.4 shows precision-recall curves under different combinations of data and prior. After external knowledge integration, the area under the curve has been improved. Furthermore, with the MCDC correction to the untreated data the area under the curve has been further improved.

Our gene network inferred from A375 gene expression data is shown in the form of directed graph in two figures. Figure 3.5 shows the positive edges at a cutoff of 0.5. Figure 3.6 shows all the true positive edges found in TRANSFAC and JASPAR database at a cutoff of 0.5. Some of our inferred edges are also found in literature such as $CREB1 \rightarrow JUN$ [99, 100, 101]. Compared to another cell line we worked on, A375 cell line has less noise and therefore constructs a more reliable gene network. This cell line was also used in the previous work [25] on which we built our work. Using the same cell line helped performing our baseline comparison.

3.4.2 Results: Lung Cancer A549

We used the same assessment dataset and methods for cell line A549 as well as for cell line A375. Table 3.6 shows the contingency table comparing T&J and our result. The two thresholds for positive edges were set to 0.5 and 0.95.

We could see the MCDC correction and integrated prior knowledge also improved both the p-value and precision in A549, although the effect applying MCDC was not as significant as in A375. By applying both MCDC correction and external knowledge integration, the p-values were improved from 0.001 level to 0.0001 level. Also, the precision increased from 0.13 to 0.15 at 0.5 cutoff, and 0.14 to 0.15 at 0.95 cutoff.

The assessment result of edge ranks is shown in Table 3.7. As in A375, the external knowledge integration and MCDC correction improved the results of middle-ranked edges. Figure 3.7 shows precision-recall curves for A549 cell line. After external knowledge integration, the area under the curve has been improved. The MCDC correction to the untreated data has not further improved the result significantly, which is different from A375.

Found Edge	Knockdown data	Supervised Network	KD + Supervised Network	KD + MCDC Untrt	KD + MCDC Untrt + Supervised Network
1	6	2	7	2	2
2	9	3	10	6	6
3	11	7	11	8	8
4	15	12	15	9	9
5	22	26	33	12	13
6	34	40	34	20	16
7	39	54	37	22	23
8	40	63	42	29	27
9	41	64	43	42	28
10	48	66	45	45	30
11	56	88	65	49	35
12	58	91	67	53	41
13	65	106	73	56	56
14	77	109	74	57	57
15	78	110	80	66	58
16	86	112	82	68	63
17	98	114	83	71	80
18	100	117	95	73	87
19	102	118	100	83	90
20	111	132	109	92	96
21	119	136	113	93	97
22	122	139	119	95	99
23	130	151	122	102	102
24	131	154	124	108	116
25	135	156	127	112	123

Table 3.5: Comparison of the rank of the first 25 edges found and match the TRANSFAC and JASPAR edgelist in cell line A375. Edges are ranked by posterior probability. The numbers represent the rankings of true positive edges (*i.e.* edges found both in our gene network and T&J edgelist) among positive edges (*i.e.* edges found in our network). The table shows the external knowledge integration helps improving results of middle-ranked edges, which makes the result more steady.

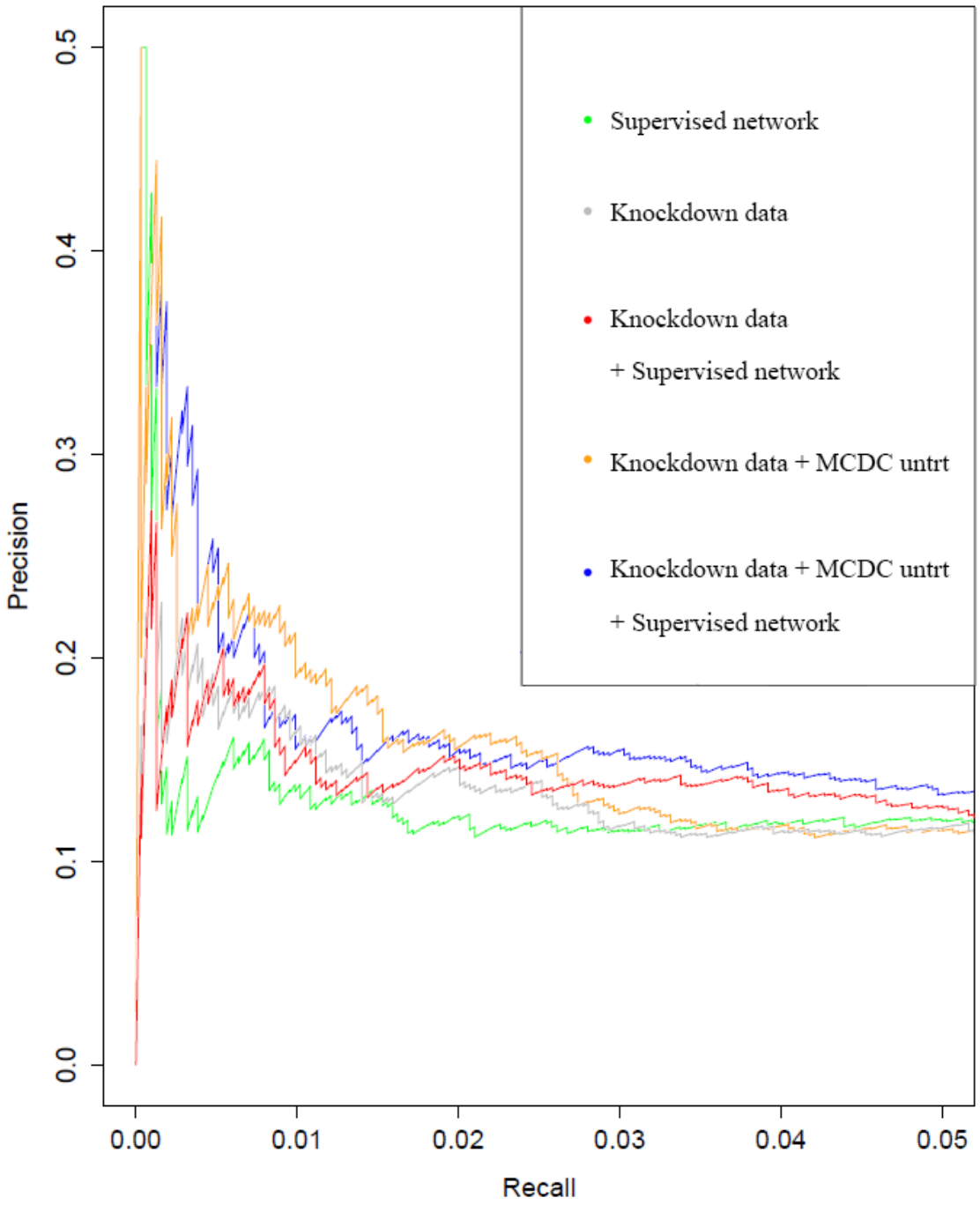

Figure 3.4: Precision-recall curves for cell line A375 using different data assessed with TRANSFAC and JASPAR. The results are improved by external knowledge integration with or without MCDC correction.

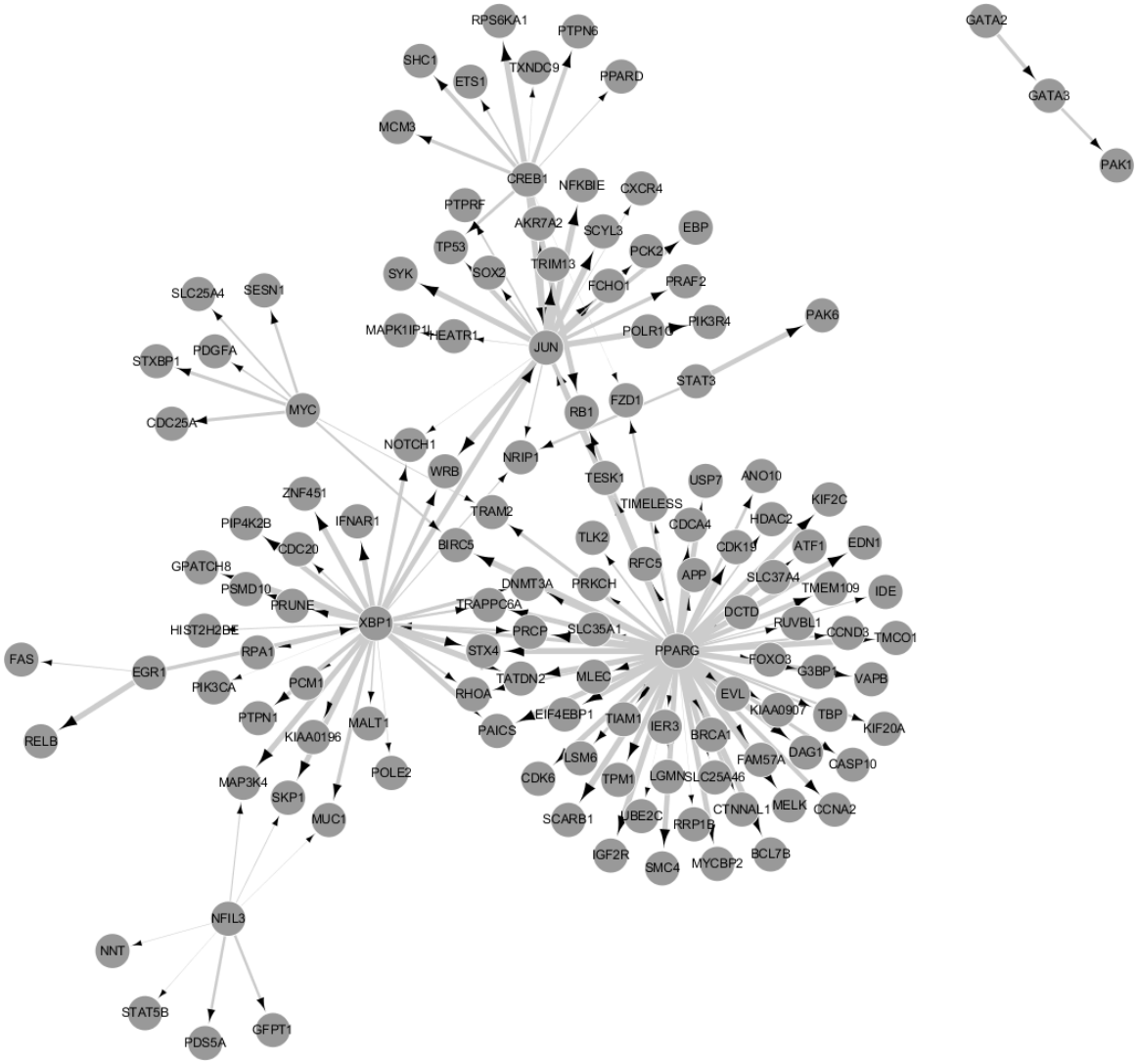

Figure 3.5: Our inferred gene network consisted of all the positive edges at a cutoff of 0.5. Each node represents a gene and each edge represents a regulatory interaction between the two genes. The arrows of edges show the direction of regulation. The width of each edge is in proportion to the inferred posterior probability that the regulatory relationship exists for the corresponding gene pair.

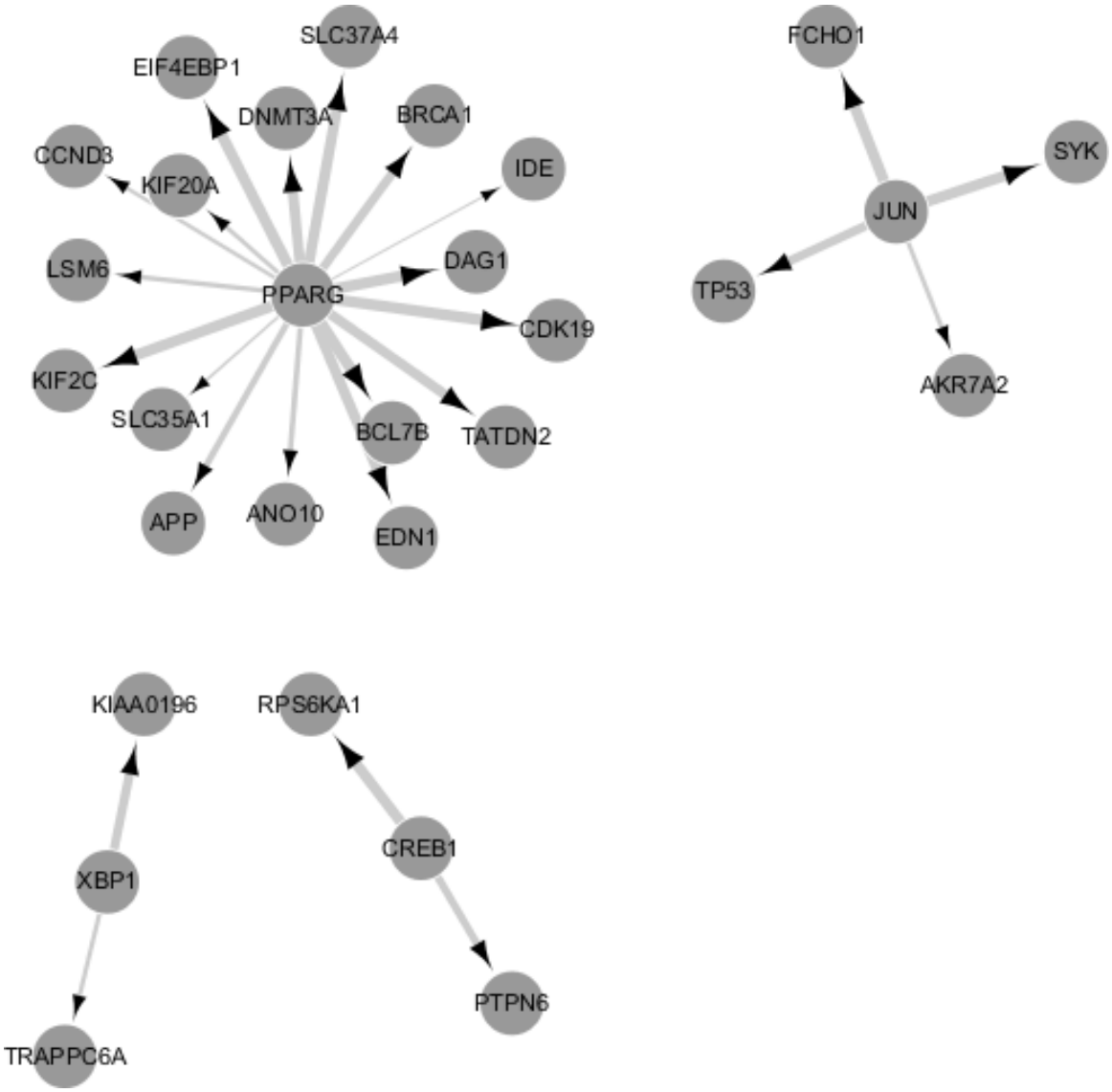

Figure 3.6: Our inferred gene network consisted of all the true positive edges found in TRANSFAC and JASPAR database at a cutoff of 0.5. Each node represents a gene and each edge represents a regulatory interaction between the two genes. The arrows of edges show the direction of regulation. The width of each edge is in proportion to the inferred posterior probability that the regulatory relationship exists for the corresponding gene pair.

1. KNOCKDOWN DATA

CUTOFF 0.5

T&J	YES	NO
YES	75	2853
NO	487	26472

P-VALUE: 0.00371

PRECISION: 0.13345

CUTOFF 0.95

T&J	YES	NO
YES	62	2866
NO	380	26579

P-VALUE: 0.002561

PRECISION: 0.14027

2. KNOCKDOWN DATA + SUPERVISED NETWORK

CUTOFF 0.5

T&J	YES	NO
YES	82	2641
NO	466	26490

P-VALUE: 0.0001138

PRECISION: 0.149805

CUTOFF 0.95

T&J	YES	NO
YES	67	2861
NO	272	26586

P-VALUE: 0.0004039

PRECISION: 0.149883

3. KNOCKDOWN DATA + MCDC UNTRT + SUPERVISED NETWORK

CUTOFF 0.5

T&J	YES	NO
YES	79	2849
NO	450	26509

P-VALUE: 0.0001037

PRECISION: 0.149338

CUTOFF 0.95

T&J	YES	NO
YES	64	2864
NO	347	26612

P-VALUE: 0.0001383

PRECISION: 0.155718

Table 3.6: Assessment results compared to TRANSFAC and JASPAR. The contingency tables display the comparison results for cell line A549 before and after external knowledge integration at 0.5 and 0.95 cutoffs. P-values have been improved by external knowledge integration at both cutoffs.

Found Edge	Knockdown data	Supervised Network	KD + Supervised Network	KD + MCDC Untrt	KD + MCDC Untrt + Supervised Network
1	8	9	2	5	6
2	11	14	25	8	9
3	16	21	30	12	14
4	35	31	36	26	28
5	40	32	43	27	29
6	41	40	46	34	36
7	43	45	47	52	56
8	55	57	62	53	57
9	81	70	70	64	70
10	82	84	71	75	73
11	107	87	83	85	75
12	111	97	94	87	79
13	131	101	100	89	88
14	136	110	107	98	104
15	139	115	115	100	111
16	164	117	123	103	118
17	175	119	124	125	127
18	177	121	128	137	139
19	178	124	129	149	141
20	180	143	130	157	161
21	183	147	142	170	163
22	190	149	156	171	175
23	196	162	163	180	176
24	219	165	170	184	177
25	220	183	178	194	180

Table 3.7: Comparison of the rank of the first 25 edges found and match the TRANSFAC and JASPAR edgelist in cell line A549. Edges are ranked by posterior probability. The numbers represent the rankings of true positive edges (*i.e.* edges found both in our gene network and T&J edgelist) among positive edges (*i.e.* edges found in our network). The table shows the external knowledge integration helps improving results of middle-ranked edges, which makes the result more steady.

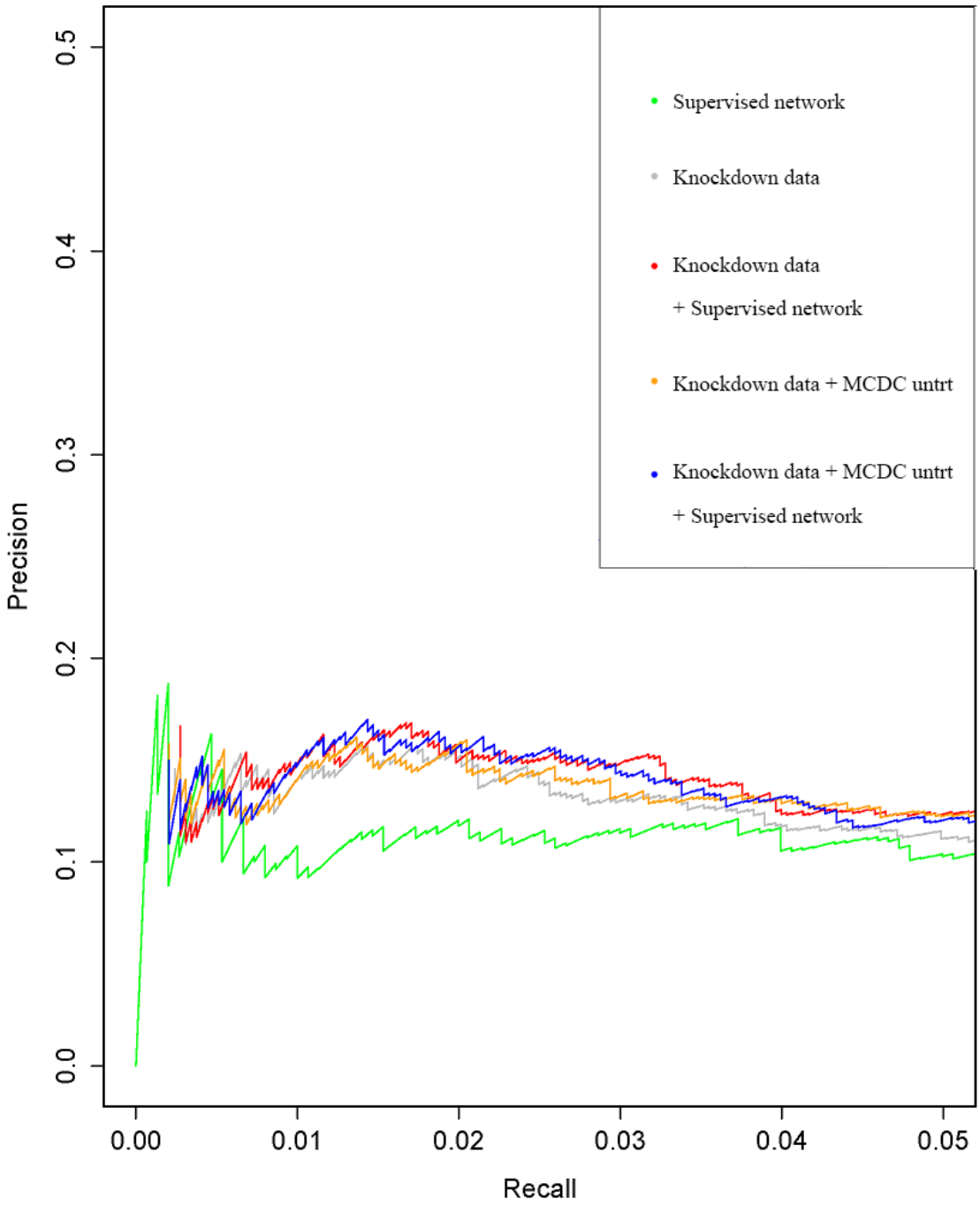

Figure 3.7: Precision-recall curves for cell line A549 using different data assessed with TRANSFAC and JASPAR. The results are improved by external knowledge integration with or without MCDC correction.

Chapter 4

CONCLUSION AND FUTURE WORK

In this thesis, we have presented an approach that integrates external data sources with knockdown data from human cell lines for predictive regulatory network inference. Our key contribution is to integrate multiple data sources in this extended regression framework. In addition, we extend Young *et al.* [25] by applying our methods to more than one cell line (skin melanoma cell line A375 and lung cancer cell line A549). In our method, the causality of our inferred gene network is given by the biological feature of knockdown data. Our model which is essentially a linear regression model is relatively less time consuming and efficient. The Bayesian approach we used to approximately estimate our linear regression model allows flexible external data integration.

As shown in this work, the accuracy of the inferred gene network has been improved with the help of prior knowledge integration and MCDC correction. One point notable from our results is that the extent of performance increasing varies from cell line to cell line. Studying the different responds to the data integration and MCDC of different cell lines could remain as a meaningful future work.

MCDC has only been applied to untreated data which is used as control to the knockdown data in this work. One direction of the future work could be applying MCDC to knockdown expression data instead of only control data to observe further results. Another interesting topic which has not been studied in this work is to incorporate feedback cycles in our Bayesian approach. The results presented in this paper has only incorporated 14 external data sources and derived from 2 cell lines. We would expect better performance integrating more reliable prior knowledge and extend our work on different cell lines.

BIBLIOGRAPHY

[1] e1071 package. https://cran.r-project.org/package=e1071. Last accessed February, 2017.

[2] class package. https://cran.r-project.org/package=class. Last accessed February, 2017.

[3] ada package. https://cran.r-project.org/package=ada. Last accessed February, 2017.

[4] randomForest package. https://cran.r-project.org/package=randomForest. Last accessed February, 2017.

[5] Y. Chen, J. Zhu, P. Y. Lum, X. Yang, S. Pinto, D. J. MacNeil, C. Zhang, J. Lamb, S. Edwards, S. K. Sieberts, *et al.*, "Variations in dna elucidate molecular networks that cause disease," *Nature*, vol. 452, no. 7186, pp. 429–435, 2008.

[6] V. Emilsson, G. Thorleifsson, B. Zhang, A. S. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G. B. Walters, S. Gunnarsdottir, *et al.*, "Genetics of gene expression and its effect on disease," *Nature*, vol. 452, no. 7186, pp. 423–428, 2008.

[7] E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, *et al.*, "An integrative genomics approach to infer causal associations between gene expression and disease," *Nature genetics*, vol. 37, no. 7, pp. 710–717, 2005.

[8] E. E. Schadt, A. Sachs, and S. Friend, "Embracing complexity, inching closer to reality," *Sci STkE*, vol. 295, p. 40, 2005.

[9] J. H. Woo, Y. Shimoni, W. S. Yang, P. Subramaniam, A. Iyer, P. Nicoletti, M. R. Martínez, G. López, M. Mattioli, R. Realubit, *et al.*, "Elucidating compound mechanism of action by network perturbation analysis," *Cell*, vol. 162, no. 2, pp. 441–451, 2015.

[10] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using bayesian networks to analyze expression data," *Journal of computational biology*, vol. 7, no. 3-4, pp. 601–620, 2000.

[11] N. Friedman, "Inferring cellular networks using probabilistic graphical models," *Science*, vol. 303, no. 5659, pp. 799–805, 2004.

[12] E. O. Voit, *Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists*. Cambridge University Press, 2000.

[13] M. Kærn, T. C. Elston, W. J. Blake, and J. J. Collins, "Stochasticity in gene expression: from theories to phenotypes," *Nature Reviews Genetics*, vol. 6, no. 6, pp. 451–464, 2005.

[14] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nature genetics*, vol. 34, no. 2, pp. 166–176, 2003.

[15] T. Huang, L. Liu, Z. Qian, K. Tu, Y. Li, and L. Xie, "Using genereg to construct time delay gene regulatory networks," *BMC research notes*, vol. 3, no. 1, p. 142, 2010.

[16] S.-Q. Zhang, W.-K. Ching, N.-K. Tsing, H.-Y. Leung, and D. Guo, "A new multiple regression approach for the construction of genetic regulatory networks," *Artificial Intelligence in Medicine*, vol. 48, no. 2, pp. 153–160, 2010.

[17] M. Bansal, G. Della Gatta, and D. Di Bernardo, "Inference of gene regulatory networks and compound mode of action from time course gene expression profiles," *Bioinformatics*, vol. 22, no. 7, pp. 815–822, 2006.

[18] Yeung, Ka Yee, et al., "Construction of regulatory networks using expression time-series data of a genotyped population," *Proceedings of the National Academy of Sciences*, vol. 108, no. 48, pp. 19436–19441, 2011.

[19] K. Lo, A. E. Raftery, K. M. Dombek, J. Zhu, E. E. Schadt, R. E. Bumgarner, and K. Y. Yeung, "Integrating external biological knowledge in the construction of regulatory networks from time-series expression data," *BMC systems biology*, vol. 6, no. 1, p. 1, 2012.

[20] W. C. Young, A. E. Raftery, and K. Y. Yeung, "Fast bayesian inference for gene regulatory networks using scanbma," *BMC systems biology*, vol. 8, no. 1, p. 47, 2014.

[21] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-dna interactions," *Science*, vol. 316, no. 5830, pp. 1497–1502, 2007.

[22] M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Koh, K. Toufighi, S. Mostafavi, *et al.*, "The genetic landscape of a cell," *science*, vol. 327, no. 5964, pp. 425–431, 2010.

[23] A. Djebbari and J. Quackenbush, "Seeded bayesian networks: constructing genetic networks from microarray data," *BMC systems biology*, vol. 2, no. 1, p. 57, 2008.

[24] S. Imoto, S. Kim, T. Goto, S. Aburatani, K. Tashiro, S. Kuhara, and S. Miyano, "Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network," *Journal of bioinformatics and computational biology*, vol. 1, no. 02, pp. 231–252, 2003.

[25] W. C. Young, A. E. Raftery, and K. Y. Yeung, "A posterior probability approach for gene regulatory network inference in genetic perturbation data," *Mathematical Biosciences and Engineering*, vol. 13, no. 6, pp. 1241–1251, 2016.

[26] W. C. Young, K. Y. Yeung, and A. E. Raftery, "Model-based clustering with data correction for removing artifacts in gene expression data," *arXiv preprint arXiv:1602.06316*, 2016.

[27] P. D'haeseleer, X. Wen, S. Fuhrman, R. Somogyi, *et al.*, "Linear modeling of mrna expression levels during cns development and injury.," in *Pacific symposium on biocomputing*, vol. 4, pp. 41–52, Citeseer, 1999.

[28] R. Guthke, U. Möller, M. Hoffmann, F. Thies, and S. Töpfer, "Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection," *Bioinformatics*, vol. 21, no. 8, pp. 1626–1634, 2005.

[29] J. Zhu, Y. Chen, A. S. Leonardson, K. Wang, J. R. Lamb, V. Emilsson, and E. E. Schadt, "Characterizing dynamic changes in the human blood transcriptional network," *PLoS Comput Biol*, vol. 6, no. 2, p. e1000671, 2010.

[30] S. Y. Kim, S. Imoto, and S. Miyano, "Inferring gene networks from time series microarray data using dynamic bayesian networks," *Briefings in bioinformatics*, vol. 4, no. 3, pp. 228–235, 2003.

[31] K. Murphy, S. Mian, *et al.*, "Modelling gene expression data using dynamic bayesian networks," tech. rep., Technical report, Computer Science Division, University of California, Berkeley, CA, 1999.

[32] J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink, and E. D. Jarvis, "Advances to bayesian network inference for generating causal networks from observational biological data," *Bioinformatics*, vol. 20, no. 18, pp. 3594–3603, 2004.

[33] F. Geier, J. Timmer, and C. Fleck, "Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge," *BMC systems biology*, vol. 1, no. 1, p. 11, 2007.

[34] M. S. Yeung, J. Tegnér, and J. J. Collins, "Reverse engineering gene networks using singular value decomposition and robust regression," *Proceedings of the National Academy of Sciences*, vol. 99, no. 9, pp. 6163–6168, 2002.

[35] T. Chen, H. L. He, G. M. Church, *et al.*, "Modeling gene expression with differential equations.," in *Pacific symposium on biocomputing*, vol. 4, p. 40, 1999.

[36] L. F. Wessels, E. P. van Someren, M. J. Reinders, *et al.*, "A comparison of genetic network models.," in *pacific Symposium on Biocomputing*, vol. 6, pp. 508–519, 2001.

[37] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.

[38] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello, "Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans," *nature*, vol. 391, no. 6669, pp. 806–811, 1998.

[39] C. C. Mello and D. Conte, "Revealing the world of rna interference," *Nature*, vol. 431, no. 7006, pp. 338–342, 2004.

[40] A. Pinna, N. Soranzo, and A. De La Fuente, "From knockouts to networks: establishing direct cause-effect relationships through graph analysis," *PloS one*, vol. 5, no. 10, p. e12912, 2010.

[41] S. Rogers and M. Girolami, "A bayesian regression approach to the inference of regulatory networks from gene expression data," *Bioinformatics*, vol. 21, no. 14, pp. 3131–3137, 2005.

[42] F. H. M. Salleh, S. M. Arif, S. Zainudin, and M. Firdaus-Raih, "Reconstructing gene regulatory networks from knock-out data using gaussian noise model and pearson correlation coefficient," *Computational biology and chemistry*, vol. 59, pp. 3–14, 2015.

[43] Q. Duan, C. Flynn, M. Niepel, M. Hafner, J. L. Muhlich, N. F. Fernandez, A. D. Rouillard, C. M. Tan, E. Y. Chen, T. R. Golub, *et al.*, "Lincs canvas browser: interactive web app to query, browse and interrogate lincs l1000 gene expression signatures," *Nucleic acids research*, p. gku476, 2014.

[44] R. Bonneau, D. J. Reiss, P. Shannon, M. Facciotti, L. Hood, N. S. Baliga, and V. Thorsson, "The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo," *Genome biology*, vol. 7, no. 5, p. R36, 2006.

[45] A. Shojaie, A. Jauhiainen, M. Kallitsis, and G. Michailidis, "Inferring regulatory networks by combining perturbation screens and steady state gene expression profiles," *PloS one*, vol. 9, no. 2, p. e82393, 2014.

[46] S. Christley, Q. Nie, and X. Xie, "Incorporating existing network information into gene network inference," *PloS one*, vol. 4, no. 8, p. e6799, 2009.

[47] H. De Jong, "Modeling and simulation of genetic regulatory systems: a literature review," *Journal of computational biology*, vol. 9, no. 1, pp. 67–103, 2002.

[48] T. S. Gardner, D. Di Bernardo, D. Lorenz, and J. J. Collins, "Inferring genetic networks and identifying compound mode of action via expression profiling," *Science*, vol. 301, no. 5629, pp. 102–105, 2003.

[49] D. di Bernardo, M. J. Thompson, T. S. Gardner, S. E. Chobot, E. L. Eastwood, A. P. Wojtovich, S. J. Elliott, S. E. Schaus, and J. J. Collins, "Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks," *Nature biotechnology*, vol. 23, no. 3, pp. 377–383, 2005.

[50] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[51] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, *et al.*, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[52] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[53] C. Charbonnier, J. Chiquet, and C. Ambroise, "Weighted-lasso for structured network inference from time course data," *Statistical applications in genetics and molecular biology*, vol. 9, no. 1, p. 15, 2010.

[54] E. P. van Someren, B. L. Vaes, W. T. Steegenga, A. M. Sijbers, K. J. Dechering, and M. J. Reinders, "Least absolute regression network analysis of the murine osteoblast differentiation network," *Bioinformatics*, vol. 22, no. 4, pp. 477–484, 2006.

[55] M. Gustafsson and M. Hörnquist, "Gene expression prediction by soft integration and the elastic netbest performance of the dream3 gene expression challenge," *PLoS One*, vol. 5, no. 2, p. e9134, 2010.

[56] J. Peng, J. Zhu, A. Bergamaschi, W. Han, D.-Y. Noh, J. R. Pollack, and P. Wang, "Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer," *The annals of applied statistics*, vol. 4, no. 1, p. 53, 2010.

[57] K. Y. Yeung, R. E. Bumgarner, and A. E. Raftery, "Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data," *Bioinformatics*, vol. 21, no. 10, pp. 2394–2402, 2005.

[58] Ling-Hong Hung, Kaiyuan Shi, Migao Wu, William Chad Young, Adrian E. Raftery and Ka Yee Yeung, "fastbma: Scalable network inference and transitive reduction," in preparation.

[59] M. Fronczuk, A. E. Raftery, and K. Y. Yeung, "Cynetworkbma: a cytoscape app for inferring gene regulatory networks," *Source code for biology and medicine*, vol. 10, no. 1, p. 11, 2015.

[60] T. D. Nielsen and F. V. Jensen, *Bayesian networks and decision graphs*. Springer Science & Business Media, 2009.

[61] A. V. Werhli, D. Husmeier, *et al.*, "Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge," *Stat Appl Genet Mol Biol*, vol. 6, no. 1, p. 15, 2007.

[62] E. E. Schadt, "Molecular networks as sensors and drivers of common human diseases," *Nature*, vol. 461, no. 7261, pp. 218–223, 2009.

[63] D. M. Chickering, "Learning bayesian networks is np-complete," in *Learning from data*, pp. 121–130, Springer, 1996.

[64] D. M. Chickering, D. Heckerman, and C. Meek, "Large-sample learning of bayesian networks is np-hard," *Journal of Machine Learning Research*, vol. 5, no. Oct, pp. 1287–1330, 2004.

[65] P. P. Le, A. Bahl, and L. H. Ungar, "Using prior knowledge to improve genetic network reconstruction from microarray data," *In silico biology*, vol. 4, no. 3, pp. 335–353, 2004.

[66] G. M. James, C. Sabatti, N. Zhou, and J. Zhu, "Sparse regulatory networks," *The annals of applied statistics*, vol. 4, no. 2, p. 663, 2010.

[67] N. Nariai, S. Kim, S. Imoto, and S. Miyano, "Using protein-protein interactions for refining gene networks estimated from microarray data by bayesian networks," in *Pacific Symposium on Biocomputing (PSB03)*, pp. 336–347, 2003.

[68] D. Koczan, S. Drynda, M. Hecker, A. Drynda, R. Guthke, J. Kekow, and H.-J. Thiesen, "Molecular discrimination of responders and nonresponders to anti-tnfalpha therapy in rheumatoid arthritis by etanercept," *Arthritis research & therapy*, vol. 10, no. 3, p. R50, 2008.

[69] C. Spieth, F. Streichert, N. Speer, A. Zell, *et al.*, "Inferring regulatory systems with noisy pathway information.," in *German Conference on Bioinformatics*, pp. 193–203, Citeseer, 2005.

[70] J. Zhu, B. Zhang, E. N. Smith, B. Drees, R. B. Brem, L. Kruglyak, R. E. Bumgarner, and E. E. Schadt, "Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks," *Nature genetics*, vol. 40, no. 7, pp. 854–861, 2008.

[71] K. Y. Yip, R. P. Alexander, K.-K. Yan, and M. Gerstein, "Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data," *PloS one*, vol. 5, no. 1, p. e8121, 2010.

[72] S. A. Dunbar, "Applications of luminex® xmap technology for rapid, high-throughput multiplexed nucleic acid detection," *Clinica Chimica Acta*, vol. 363, no. 1, pp. 71–82, 2006.

[73] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, *et al.*, "The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease," *science*, vol. 313, no. 5795, pp. 1929–1935, 2006.

[74] LINCS Workflow: L1000 data processing. http://lincsproject.org/LINCS/tools/workflows/find-the-best-place-to-obtain-the-lincs-l1000-data. Last accessed April, 2017.

[75] BayesKnockdown package. https://bioconductor.org/packages/release/bioc/html/BayesKnockdown.h Last accessed February, 2017.

[76] A. Zellner, "On assessing prior distributions and bayesian regression analysis with g-prior distributions," *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, vol. 6, pp. 233–243, 1986.

[77] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[78] N. Guelzim, S. Bottani, P. Bourgine, and F. Képès, "Topological and causal structure of the yeast transcriptional regulatory network," *Nature genetics*, vol. 31, no. 1, pp. 60–63, 2002.

[79] E. Portales-Casamar, S. Kirov, J. Lim, S. Lithwick, M. I. Swanson, A. Ticoll, J. Snoddy, and W. W. Wasserman, "Pazar: a framework for collection and dissemination of cis-regulatory sequence annotation," *Genome biology*, vol. 8, no. 10, p. R207, 2007.

[80] E. Portales-Casamar, D. Arenillas, J. Lim, M. I. Swanson, S. Jiang, A. McCallum, S. Kirov, and W. W. Wasserman, "The pazar database of gene regulatory information coupled to the orca toolkit for the study of regulatory sequences," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D54–D60, 2009.

[81] PAZAR, public database of transcription factors and regulatory sequence annotation. http://www.pazar.info/. Last accessed February, 2017.

[82] BioMart. http://www.biomart.org/. Last accessed February, 2017.

[83] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, *et al.*, "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, 2012.

[84] C. Klijn, S. Durinck, E. W. Stawiski, P. M. Haverty, Z. Jiang, H. Liu, J. Degenhardt, O. Mayba, F. Gnad, J. Liu, *et al.*, "A comprehensive transcriptional portrait of human cancer cell lines," *Nature biotechnology*, vol. 33, no. 3, pp. 306–312, 2015.

[85] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[86] G. O. Consortium *et al.*, "Gene ontology consortium: going forward," *Nucleic acids research*, vol. 43, no. D1, pp. D1049–D1056, 2015.

[87] E. P. Consortium *et al.*, "An integrated encyclopedia of dna elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012.

[88] E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Maayan, "Enrichr: interactive and collaborative html5 gene list enrichment analysis tool," *BMC bioinformatics*, vol. 14, no. 1, p. 128, 2013.

[89] T. Kelder, M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo, and A. R. Pico, "Wikipathways: building research communities on biological pathways," *Nucleic acids research*, vol. 40, no. D1, pp. D1301–D1307, 2012.

[90] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[91] D. Nishimura, "Biocarta," *Biotech Software & Internet Report: The Computer Software Journal for Scient*, vol. 2, no. 3, pp. 117–120, 2001.

[92] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, *et al.*, "The reactome pathway knowledgebase," *Nucleic acids research*, vol. 42, no. D1, pp. D472–D477, 2014.

[93] A. Fabregat, K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann, R. Haw, B. Jassal, S. Jupe, F. Korninger, S. McKay, *et al.*, "The reactome pathway knowledgebase," *Nucleic acids research*, vol. 44, no. D1, pp. D481–D487, 2016.

[94] J. H. Wolfe, "Pattern clustering by multivariate mixture analysis," *Multivariate Behavioral Research*, vol. 5, no. 3, pp. 329–350, 1970.

[95] J. D. Banfield and A. E. Raftery, "Model-based gaussian and non-gaussian clustering," *Biometrics*, pp. 803–821, 1993.

[96] G. McLachlan and D. Peel, *Finite mixture models*. John Wiley & Sons, 2004.

[97] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.

[98] X. Zhao and et al., "Jaspar 2013: An extensively expanded and updated open-access database of transcription factor binding profiles," *TBA*, vol. TBA, no. TBA, p. TBA, 2013.

[99] D. M. Benbrook and N. C. Jones, "Heterodimer formation between creb and jun proteins," *Oncogene*, vol. 5, no. 3, pp. 295–302, 1990.

[100] D. J. Spring and E. G. Krebs, "Deletion of 11 amino acids in p90 rsk-mo-1abolishes kinase activity," *Molecular and cellular biology*, vol. 19, no. 1, pp. 317–320, 1999.

[101] N. Liu, E. Cigola, C. Tinti, B. K. Jin, B. Conti, B. T. Volpe, and H. Baker, "Unique regulation of immediate early gene and tyrosine hydroxylase expression in the odor-deprived mouse olfactory bulb," *Journal of Biological Chemistry*, vol. 274, no. 5, pp. 3042–3047, 1999.

[102] TRANSFAC and JASPAR. http://jaspar.genereg.net/. Last accessed February, 2017.

[103] Gene Ontology Consortium. http://geneontology.org/. Last accessed February, 2017.

[104] Cancer Cell Line Encyclopedia. http://www.broadinstitute.org/ccle/. Last accessed February, 2017.

[105] RNA-seq data. http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2706/. Last accessed February, 2017.

[106] Encyclopedia of DNA Elements (ENCODE) Consortium. http://genome.ucsc.edu/ENCODE/. Last accessed February, 2017.

[107] S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano, "Combining microarrays and biological knowledge for estimating gene networks via bayesian networks," *Journal of bioinformatics and computational biology*, vol. 2, no. 01, pp. 77–98, 2004.

[108] Enrichr. http://amp.pharm.mssm.edu/Enrichr/. Last accessed February, 2017.

[109] The LINCS program. http://www.lincsproject.org/. Last accessed February, 2017.

[110] T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood, "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network," *Science*, vol. 292, no. 5518, pp. 929–934, 2001.

[111] T. Ideker, T. Galitski, and L. Hood, "A new approach to decoding life: systems biology," *Annual review of genomics and human genetics*, vol. 2, no. 1, pp. 343–372, 2001.

[112] M. Hecker, S. Lambeck, S. Toepfer, E. Van Someren, and R. Guthke, "Gene regulatory network inference: data integration in dynamic modelsa review," *Biosystems*, vol. 96, no. 1, pp. 86–103, 2009.

[113] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, R. A. Young, *et al.*, "Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks.," in *Pacific symposium on biocomputing*, vol. 6, p. 266, 2001.

[114] J. Zhu, P. Y. Lum, J. Lamb, D. GuhaThakurta, S. W. Edwards, R. Thieringer, J. P. Berger, M. S. Wu, J. Thompson, A. B. Sachs, *et al.*, "An integrative genomics approach to the reconstruction of gene networks in segregating populations," *Cytogenetic and genome research*, vol. 105, no. 2-4, pp. 363–374, 2004.

[115] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

[116] K. Y. Yeung, K. M. Dombek, K. Lo, J. E. Mittler, J. Zhu, E. E. Schadt, R. E. Bumgarner, and A. E. Raftery, "Construction of regulatory networks using expression time-series data of a genotyped population," *Proceedings of the National Academy of Sciences*, vol. 108, no. 48, pp. 19436–19441, 2011.

[117] S.-I. Lee, A. M. Dudley, D. Drubin, P. A. Silver, N. J. Krogan, D. Pe'er, and D. Koller, "Learning a prior on regulatory potential from eqtl data," *PLoS Genet*, vol. 5, no. 1, p. e1000358, 2009.

[118] J. Rung, T. Schlitt, A. Brazma, K. Freivalds, and J. Vilo, "Building and analysing genome-wide gene disruption networks," *Bioinformatics*, vol. 18, no. suppl 2, pp. S202–S210, 2002.