Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models

Anders Skrondal
&
Sophia Rabe-Hesketh

Contents

CHAPTER 1

A Misclassification Model for Inferring Transcriptional Regulatory Networks

Ning Sun, Hongyu Zhao
Yale University

1.1 Introduction

Understanding gene regulations through the underlying transcriptional regulatory networks (referred as TRNs in the following) is a central topic in biology. A TRN can be thought of as consisting of a set of proteins (transcription factors), genes, small modules, and their mutual regulatory interactions. The potentially large number of components, the high connectivity among various components, and the transient stimulation in the network result in great complexity of TRNs. With the rapid advances of molecular technologies and enormous amounts of data being collected, intensive efforts have been made to dissect TRNs using data generated from the state-of-the-art technologies, including gene expression data and other data types (e.g. Chu *et al.*, 1998; Ren *et al.*, 2000; Davison *et al.*, 2002; Lee *et al.*, 2002; Bar-Joseph *et al.*, 2003; Zhang and Gerstein, 2003). The computational methods include gene clustering (e.g. Eisen *et al.*, 1998; Roberts *et al.*, 2000), Boolean network modeling (e.g. Liang *et al.*, 1998; Akutsu *et al.*, 1999, 2000; Shmulevich *et al.*, 2002), Bayesian network modeling (e.g. Friedman *et al.*, 2000, Hartemink *et al.*, 2001, 2002), differential equation systems (e.g. Gardner *et al.*, 2003; Tegnr *et al.*, 2003), information integration methods (e.g. Gao *et al.*, 2004), and other approaches. For recent reviews, see de Jong *et al.*(2002) and Sun and Zhao (2004). As discussed in our review (Sun and Zhao, 2004), although a large number of studies are devoted to infer TRNs from gene expression data alone, such data only provide very limited amount of information. On the other hand, other data types, such as protein-DNA interaction data (which measure the binding targets of each transcription factor, denoted by TF in our following discussion, through direct biological experiments), may be much more informative and should be combined together for network inference.

In this article, we describe a Bayesian framework for TRN inference based on the combined analysis of gene expression data and protein-DNA interaction data. The

statistical properties of our approach are investigated through extensive simulations, and our method is then applied to study TRNs in the yeast cell cycle.

1.2 METHODS

In this article, we model a TRN as a bipartite graph: a one-layer network where a set of genes are regulated by a set of TFs. The TFs bind to the regulatory regions of their target genes to regulate (activate or inhibit) the transcription initiation of these genes. Transcription initiation is a principal mode of regulating the expression levels of many, if not most, genes (Carey and Smale, 1999). Because the number of the genes largely exceeds the number of TFs in any organism (e.g. there are 374 TF entries in the updated Transfac database (http://www.gene-regulation.com/pub/databases.html) and more than 6000 genes in yeast), there is combinatorial control of the TFs on genes. That is, for a given gene, its expression level is controlled by the joint actions of its regulators. Two well-known facts on the joint actions of TFs include cooperativity, which in the context of protein-DNA interaction refers to two or more TFs engaging in protein-protein interaction stabilize each other's binding to DNA sequences, and transcriptional synergy, which refers to the interacting effects among the Polymerase II general transcriptional machinery and the multiple TFs on controlling transcription levels. In our previous work (Zhao *et al.*, 2003), we assumed that the expression level of a specific gene is controlled through the additive effects of its regulators. Liao *et al.* (2003) applied Hill's equation for the cooperative TF bindings on the regulatory regions of their target genes and the first order kinetics for the rate of gene transcription. Under a quasi-steady state assumption, they proved that the relative gene expression level has a linear relationship with the relative activities of the TFs that bind on the gene's regulatory region. In order to obtain a unique solution of the regulation matrix, they required the full column-rank of the regulation and its reduced matrices. In this article, we extend our previous work (Zhao *et al.*, 2003) to fully incorporate gene expression data and protein-DNA binding data to infer TRNs. Before the discussion of our model, we first give a brief overview of the protein-DNA binding data used in our method.

As the primary goal of TRN inference is to identify the regulation targets of each TF, the most direct biological approach for this goal is to experimentally identify the targets of various TFs. Many different biological methodologies are available to serve this purpose. The large-scale chromatin immunoprecipitation microarray data (ChIP-chip data) provide the in vivo measurements on TFs and DNA binding in yeast (Ren *et al.*, 2000; Lee *et al.*, 2002). In our study, the protein-DNA binding data thus collected are viewed as one measurement of the TRN with certain level of measurement errors due to biological and experimental variations, e.g. physical binding is not equivalent to regulation. We use the ChIP-chip data collected by Lee *et al.* (2002) as the data source for protein-DNA binding. These data represent a continuous measurement of the binding strength between each TF and its potential targets, and a p-value is derived based on replicated experiments to assess the statistical significance of binding. In our following work, the inferred binding p-values between

a TF and its potential target genes are transformed into binary observations using a significance level cut off of 0.05. That is, for all TF-gene pairs whose p-value is below 0.05, we denote the observation as 1, representing evidence for binding, and for those pairs whose p-value is larger than 0.05, we denote the observation as 0, representing not sufficient evidence for binding. The reason that we utilize protein-DNA binding data is because we believe that the information from such data serves as a close measurement for the true underlying TRN.

In our previous work (Zhao *et al.*, 2003), we treated protein-DNA binding data as representing the true underlying network, and used a simple linear model to describe the relationship between the transcript amounts of the genes considered and their regulators' activities. In our current work, we extend this linear model to incorporate potential errors associated with protein-DNA binding data to integrate three components that are biologically important in transcription regulation, namely, the TRN as characterized by the covariate (or design) matrix in the linear model, protein regulation activities as defined by the predictors in the model, and gene expression levels as defined by the response variables. We propose a misclassification model to simultaneously extract information from protein-DNA binding data and gene expression data to reconstruct TRNs.

1.2.1 Model Specification

Our model relating gene expression levels, TRNs, and TF activities can be described through three sub-models:

- A linear regression model relating gene expression levels with the true underlying TRNs and regulators' activities;
- A misclassification model relating the true underlying networks and the observed protein-DNA binding data;
- Prior distribution on the TRNs.

The information on the measurement error can be built in a flexible way into a graphical model (Richardson and Gilks, 1993; Richardson, 1999). The hierarchical structure of our graphical model is summarized in Figure 1.1 and we describe each component in detail in the following.

The first sub-model: the linear regression model

Let N denote the number of genes and M denote the number of TFs related to the regulation of these genes. We consider a total number T of gene expression experiments, where these experiments may represent a time-course study, e.g. yeast cell cycle studies, or different knock out experiments. We focus on time-course experiments in our following discussion. In this case, we use t represents a specific time point. The observed gene expression levels at time t, $\mathbf{Y}_t$, are the vector of N expression levels normalized over all time points for each gene i and serve as the response

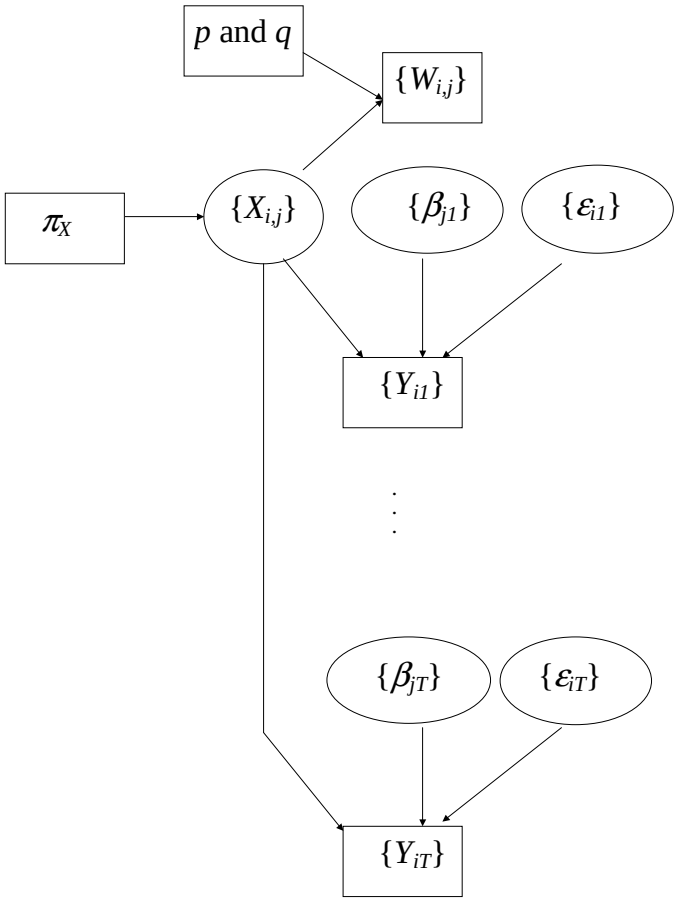

Figure 1.1 *The hierarchical structure of the misclassification model discussed in this paper. The unknown parameters are in the ovals, and the known parameters are in the rectangles.*

in the linear model (1.1) with the following form:

$$Y_t = X\beta_t + \epsilon_t \tag{1.1}$$

$$\epsilon_{it} \sim N(0, \sigma_t^2) \tag{1.2}$$

where **X** represents the true TRN, β represents the time dependent regulator activities of the *M* TFs, and ϵ_t represents the errors that are associated with gene expression measurements. In matrix **X**, each row corresponds to a gene and each column corresponds to a TF. Therefore, the (i,j)th entry in this matrix represents the regulation pattern of the *j*th TF to the *i*th gene. The value of this entry is 1 if the *j*th TF affects the transcription level of the *i*th gene, and the value is 0 otherwise. Therefore, if our primary interest is to infer the TRN, the overall objective is to infer the values in this matrix, either 0 or 1.

This model states that (1) the expression level of a gene is largely controlled by the additive regulation activities of its regulators, (2) the same regulator has the same relative effect on all its targets, (3) the TRN is identical across all time points, and (4) the errors associated with gene expression measurements have the same distribution across all the genes. We note that these assumptions are simplistic and may only provide a first order approximation to reality. This model has nevertheless (implicitly or explicitly) been used in the analysis of TRNs by many research groups and found good success. The limitations and modifications of these assumptions are further discussed in the Summary section.

Because protein-DNA binding data are often obtained from a mixture of biological samples across all the time points, e.g. the asynchronized cells, they measure an averaged protein-DNA binding over the whole cell cycle. Although we may use the time-course gene expression data to investigate the fluctuation of the network over time, the information at one time point may not be sufficient for statistical inference (see results in the simulation study in the following). Therefore, we make the assumption that the network is time independent and combine the information across time points. Consequently, the variation of the response variable, gene expression, across time points is accredited to the change in activities of the TFs, β_t. With the given activities of the predictors, the TRN of gene i ($\mathbf{X}_i$) is independent of the network of any other gene $\mathbf{X}_{i'}$, where $i' = 1, 2, ..., (i-1), (i+1), ..., N$.

The second sub-model: the misclassification model

In our model set-up, both the true and observed covariates are binary, where 0 corresponds to no regulation and 1 corresponds to regulation. We assume the following model (1.3-1.6):

$$P(W_{ij} = 1|X_{ij} = 1) = 1 - p \tag{1.3}$$
$$P(W_{ij} = 0|X_{ij} = 1) = p \tag{1.4}$$
$$P(W_{ij} = 0|X_{ij} = 0) = 1 - q \tag{1.5}$$
$$P(W_{ij} = 0|X_{ij} = 1) = q \tag{1.6}$$

where the values of p and q are the false-negative and false-positive rates of the protein-DNA data. In practice, these values may be directly estimated from some control experiments, thus we treat these parameters as known or prior information in the misclassification model and specify their values. In the case these values may not be precisely known, we also study the robustness of their misspecifications on statistical inference. Note that the false-positive and false-negative rates may be gene-TF specific, therefore, our assumption here represents a first-order approximation to reality that may need further extension in future studies. The binary binding matrix $\mathbf{W}$ serves as the measurement for the true TRN $\mathbf{X}$.

The third sub-model: the exposure model

For this submodel, we need to specify the prior distribution of the regulatory matrix **X**. The prior distribution of **X** (π_X) describes the probability of X_{ij} being 1, where X_{ij} represents the regulation between TF j and gene i. We assume that the X_{ij} are independent and have an identical distribution π_X. For a given true network **X**, the value of π_X can be calculated from the data. When X is unknown and W serves as the surrogate of **X**, π_X is a model parameter to be specified.

1.2.2 MCMC algorithm for statistical inference

In our model set-up, a large number of unknown parameters $\{\mathbf{X}, \beta_t, \sigma_t^2\}$ need to be inferred based on the observations $\mathbf{Y}_t$, t=1, , T, and **W**. We propose to use the Gibbs sampler for statistical inference. The Gibbs sampler is alternated between two steps: (1) sample $\{\beta_t, \sigma_t^2\}$ conditional on **X**; and (2) sample **X** conditional on $\{\beta_t, \sigma_t^2\}$. These two steps are described in detail in the following.

Given current estimate of **X**, the model reduces to a standard linear regression model. The parameters $\{\beta_t, \sigma_t^2\}$ are sampled through (1.7 and 1.8)

$$\sigma_t^2 \sim Inv - \chi^2(df, s_t^2) \tag{1.7}$$

$$\beta_t \sim N(\widehat{\beta}_t, \mathbf{V}_\beta \sigma_t^2) \tag{1.8}$$

where $df = N - M$, $\mathbf{V}_\beta = (\widehat{\mathbf{X}}^T \widehat{\mathbf{X}})^{-1}$, $\widehat{\beta}_t = \mathbf{V}_\beta \widehat{\mathbf{X}}^T Y_t$, and s_t is the sample standard deviation. The matrix is the current estimate for the TRN.

Given current estimates of $\{\beta_t, \sigma_t^2\}$, we individually update the TRN for each gene. If there are M TFs, there are a total of $K = 2^M$ possible combined patterns among the TFs to jointly regulate a specific gene. The likelihood L_{ik} for each pattern k can be evaluated as

$$L_{ik} = L_{ik}^X + L_{ik}^Y \tag{1.9}$$

where

$$L_{ik}^X = n_1 \log \pi_X + n_{11} \log (1-p) + n_{10} \log p + n_0 \log (1-\pi_X) + n_{01} \log q + n_{00} \log (1-q) \tag{1.10}$$

$$L_{ik}^Y = -\sum_{t=1}^{T} \frac{(Y_{it} - \widehat{Y_{ikt}})^2}{2\sigma_t^2} \tag{1.11}$$

In the above expression, L_{ik}^X and L_{ik}^Y represent the likelihood contributions from the protein-DNA binding data and the expression data, respectively. In the expression for L_{ik}^X, n_{so} represents the number of TF-gene pairs whose true regulation is s and the observed binding is o, where the values of s and o are 0 or 1. For example, n_{11} corresponds to the number of pairs whose true regulation and observed binding are both 1, $n_1 = n_{10} + n_{11}$, and $n_0 = n_{00} + n_{01}$. The expression for L_{ik}^Y represents the likelihood component derived from gene expression data across all time points. After evaluating the log-likelihood for all the patterns, we sample one pattern based

on the following multinomial distribution:

$$L_{ik}^{Y} \sim multinomial(1, \frac{\exp(L_{ik})}{\sum_{k=1}^{K} \exp(L_{ik})}) \tag{1.12}$$

Therefore, in the updating of the TRN, our algorithm does an exhaustive search over all possible network patterns for each gene, and sample a specific network based the relative likelihood of all possible networks. We repeat this for each of the N genes to obtain the updated $\widehat{\mathbf{X}}$ for the next iteration.

Based on the sampled parameter values, we can derive the posterior distributions for all the unknown parameters. For example, we can obtain the inferred TRN describing the binding between the jth TF and the ith gene through the marginal posterior distribution, i.e. the proportion of samples that the value of X_{ij} is 1. These posterior probabilities can then be used to infer the presence or absence of regulation through specifying a cut-off value, e.g. 0.5, such that all the entries below this cut-off are inferred not to have regulation effect, whereas all the entries having values above this cutoff are inferred to have regulation.

1.2.3 Data analysis and simulation set-up

As our simulation model is based on the real data to be analyzed, we describe the data sources first. According to the literature, we select eight important cell cycle TFs, namely Fkh1, Fkh2, Ndd1, Mcm1, Ace2, Swi5, Mbp1, and Swi4, and based on protein-DNA interaction data reported in Lee *et al.* (2002), we obtain a binary binding matrix for these regulators and all yeast genes. The binary observation is obtained by applying a 0.05 p-value cut-off to the p-values reported by Lee and colleagues. We then remove those genes with no *in vivo* binding evidence with any of the eight TFs from the binding matrix, and further focus only on yeast cell cycle genes defined by Spellman *et al.* (1998). These steps result in a total of 295 genes to be analyzed, and the observed protein-DNA binding matrix has a dimension of 295 (genes) by 8 (TFs). For gene expression data, we use the α arrest cell cycle data with 18 time points collected by Spellman *et al.* (1998).

Now we describe our set-up used to conduct simulation experiments to evaluate the performance of our proposed procedure. In our simulation model, we need to specify (1) the true TRN, (2) true protein regulation activities, (3) false-positive and false-negative rates in the observed binding matrix, and (4) measurement errors associated with microarray data. We consider all 295 genes used in the real data analysis, and select five TFs (Fkh2, Mcm1, Ace2, Mbp1, and Swi4, which are reported to control the gene expression at the four cell cycle stages) out of the total eight in our simulations to simplify the analysis and summary. For the specification of the "true" TRN in our simulations, we use the observed binding data to represent the true TRN. As for TF activity specifications, we estimate the activities of the chosen five TFs from the linear regression model using the above "true" TRN and the expression levels of all 295 genes at each time point. The activity levels of the five TFs over 18 time

points are shown in Figure 1.2. As for false-positive and false-negative rates, we vary their levels from 0.1 to 0.9 to examine their effects on statistical inference. Finally, we assess the effect of the measurement variation associated with microarray data on statistical inference. For the majority of simulations, we assume that the microarray data are collected from 18 time points as in Spellman *et al.* (2002). In one case, we vary the number of time points available to investigate the effect of the number of time points on statistical inference.

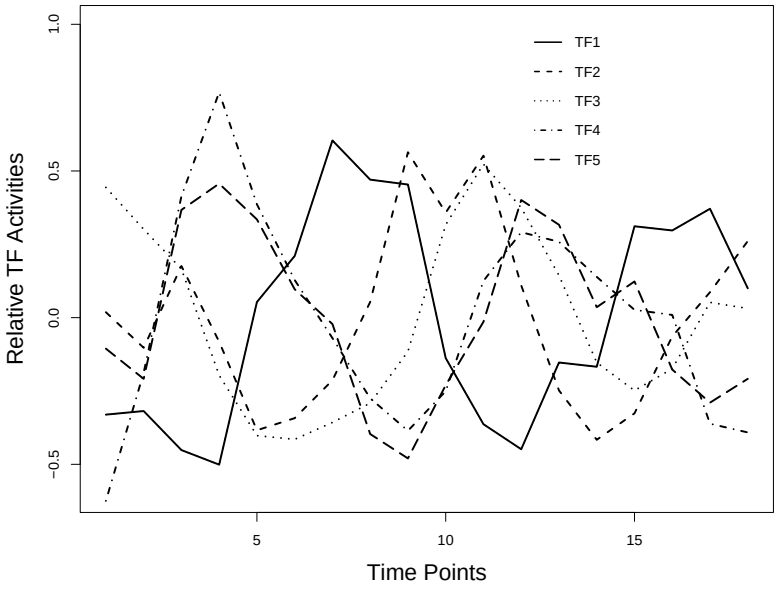

Figure 1.2 *The activities of five transcription factors vary over 18 time points. Two of the five transcription factors share similar variation, which may lead to identifiable problem of the model. However, our results show that the slight difference between the TF activities prevents the problem.*

1.3 Simulation Results

1.3.1 Convergence diagnosis of the MCMC procedure

Based on our simulation runs, we generally find good mixing of the proposed MCMC procedure. Both the traces of the parameter values and the autocorrelation of the parameter curves indicate that a burn-in run of 1,000 iterations out of 10,000 iterations

is stable enough to obtain reliable posterior distributions. The posterior distributions of the five TF activities (β_t) and measure variations from microarrays σ_t^2 at a time point from a randomly chosen simulated data set are shown in Figure 1.3. We also investigate the effect of the initial network (covariate matrix) on MCMC results. When the measurement errors in gene expression data are low, the MCMC procedure has good convergence regardless of the initial network. In general, the observed protein-DNA binding data provide a good starting point for statistical inference.

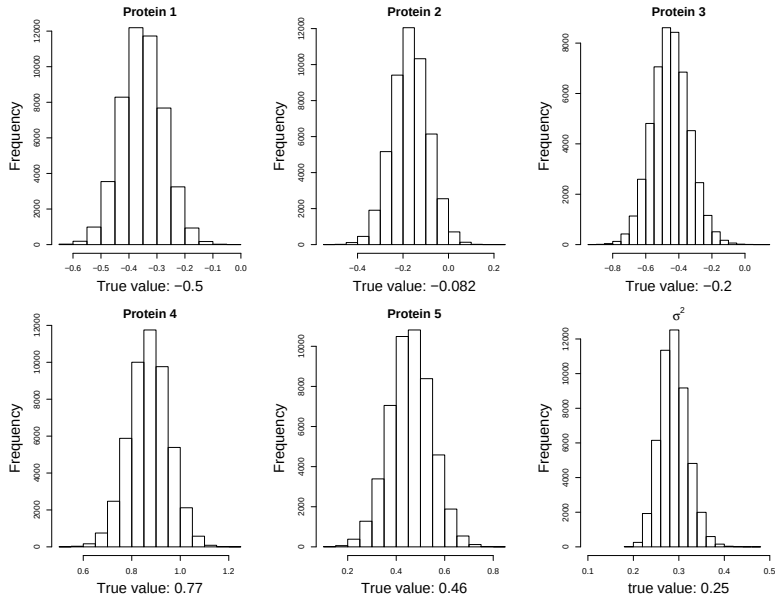

Figure 1.3 *The posterior distributions for the model parameters β_t and σ_t^2 at t = 4. The standard deviations of these posterior distributions are 0.075, 0.078, 0.092, 0.077, 0.091, and 0.027, respectively.*

In our model specification, there are two types of errors: the errors associated with the measured gene expression levels (responses, denoted by σ) and those associated with the observed protein-DNA binding data (denoted by p and q). In order to systematically investigate the effect of both types of errors, we consider seven pairs of p and q as (0.1,0.1), (0.2,0.2), (0.2,0.4), (0.4,0.2), (0.3,0.3), (0.4,0.4), and (0.5,0.5). For each pair of p and q values, we simulate the observed protein-DNA binding data as well as gene expression data under 22 different σ values, ranging from 0.001 to 1.5. For each specification of the $22 \times 7 = 154$ sets of parameter values, we simulate

data sets consisting of protein-DNA interaction data and gene expression data. Each data set is analyzed through our proposed MCMC approach with a burn-in of 1,000 iterations and a further run of 5,000 iterations. The posterior distribution for each unknown parameter is summarized and compared to the true underlying network. We use a cut-off of 0.5 to infer the presence or absence of interactions between TFs and genes. The inferred network is then compared to the true network to calculate the proportion of false-positive and false-negative inferences for each TF-gene pair. The overall false-positive and false-negative rates are then estimated through the average of all TF-gene pairs across all the simulated data sets. The results are summarized in Figure 1.4. In Figure 1.4(a), we plot the false-positive rates for the inferred network. As can be seen from this figure, the false-positive rates for the inferred network increase as σ, p, and q increase. The false-negative rates for the inferred networks show a similar pattern. The major feature is that the information from gene expression data may significantly improve the estimation on $\mathbf{X}$. When s is small and p and q are not too high, there is a very good chance that the true network can be recovered from the joint analysis of gene expression data and protein-DNA binding data. For example, with a 30% false-positive and 30% false-negative rates, when σ is less than 0.2, the whole network may be fully recovered. Even when σ is large, the false-positive rates in the inferred network using both binding data and gene expression data still outperform the false-positive rates in the observed protein-DNA expression data, i.e. gene expression data are not considered in the inference. The results for the false-negative rates as shown in Figure 1.4(b) show similar patterns.

1.3.2 Misspecification of the model parameters p, q, and π_X

In the results summarized above, we assume that the true values of p and q are precisely known to us. However, their exact values may not be accurately inferred. Therefore, we conduct simulation experiments to examine the performance of the proposed procedure when the values of p and q are misspecified. In this set of simulations, we simulate data from three sets of p and q values: (0.1,0.1), (0.3,0.3), and (0.2,0.4). For each simulated data set under a given set of parameter values, we perform statistical analysis under different sets of specifications for p and q, including (0.9,0.9), (0.8,0.8), (0.7,0.7), (0.6,0.6), (0.5,0.5), (0.4,0.4), (0.3,0.3), (0.2,0.2), (0.1,0.1), (0.05, 0.05), (0.01,0.01), and (0.05, 0.4). Throughout these simulations, we assume $\sigma = 0.2$. The performance of our procedure in terms of false-positive and false-negative rates is summarized in Figures 1.5(a) to 1.5(c). These results suggest that the statistical inference is robust to the misspecification of the parameters p and q when the specified values are not too distinct from the true parameter values. We observe similar patterns for other values of σ.

As another parameter that needs to be specified in our approach is the prior probability, π_X, that there is an interaction between a TF and a gene, we further investigate the performance of our approach when π_X is misspecified. The true value of π_X is about 0.46 $(683/(295 \times 5))$, where there are 683 regulation pairs in the protein-DNA binding data) in the given true network $\mathbf{X}$, but we consider 0.1, 0.2, 0.3, 0.4, 0.46,

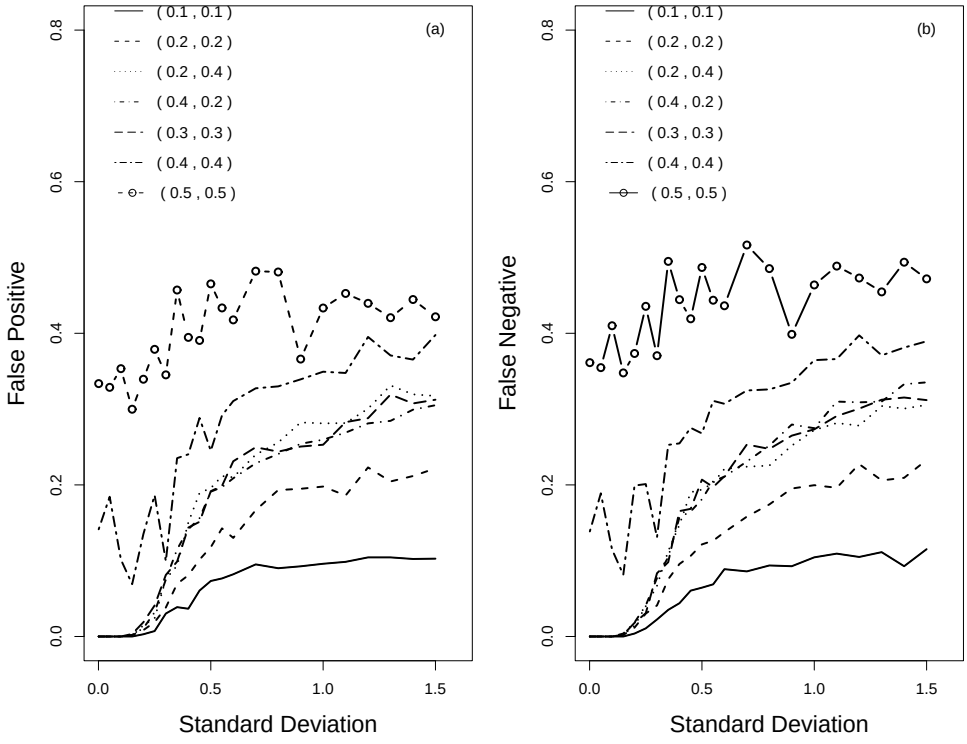

Figure 1.4 *The false positive and false negative rates of the inferred network. The X-axis is the standard deviation in the gene expression data, while the Y-axis is either the false positive rate or false negative of the posterior network with respect to the true regulatory network in the cell cycle. Different lines correspond to different levels of quality of the protein-DNA binding data.*

0.5, 0.6, 0.7, 0.8, and 0.9 in the specification of π_X in our analysis. The results are summarized in Figure 1.5(d). Compared to the results for p and q, the statistical inference is more sensitive to the value of π_X. However, when the specified parameter value is reasonably close to the true value, our approach yields generally robust estimates.

Overall, our simulation studies suggest that misspecifications of model parameters p, q, and π_X within a reasonable range will not substantially affect the statistical inference of the true network.

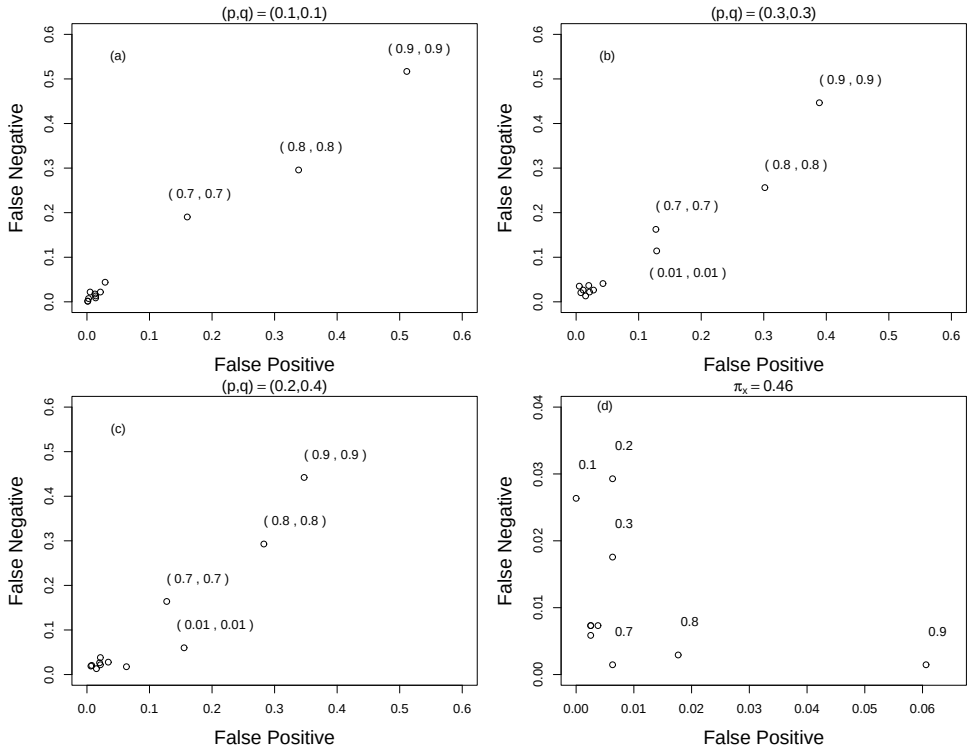

Figure 1.5 *The effects of the misspecification of the model parameters p, q, and π_X on the inferred network. The standard deviation of the simulated gene expression data is 0.2. The real values of parameters (p,q) or π_X are indicated in the title of each plot. In the first three plots, the true value of π_X is 0.46, but (p,q) are specified as (0.9,0.9), (0.8,0.8), (0.7,0.7), (0.6,0.6), (0.5,0.5), (0.4,0.4), (0.3,0.3), (0.2,0.2), (0.1,0.1), (0.05, 0.05), (0.01,0.01), and (0.05, 0.4). For the last plot, the values of (p, q) are (0.1,0.1), but π_X is specified at various levels: 0.1, 0.2, 0.3, 0.4, 0.46, 0.5, 0.6, 0.7, 0.8, and 0.9.*

1.3.3 Effect of the number of experiments used in the inference

In the above simulations, we simulate data from 18 time points and use all of them in the inference of the underlying network. In this subsection, we consider the effect of the number of time points on the inference. For this set of simulations, we simulate the protein-DNA binding data by fixing the values of p and q at 0.1, select the value of σ at 0.001, 0.2, and 0.5, and vary the number of time points used in the analysis from 1 to 18. When there is little error associated with gene expression data, i.e. $\sigma = 0.001$, the data at one time point can carry enough information to fully recover the true network. With increasing σ values, the number of time points affects the results on the inferred network (Figure 1.6). When σ is 0.2, our previous results show that there is a significant improvement of the inferred network from the binding data. As more time points are included in the analysis, we observe a more accurate inference of the underlying network. When σ is 0.5, the improvement of the inferred network from the binding data is still obvious but limited by too much noise in gene expression data.

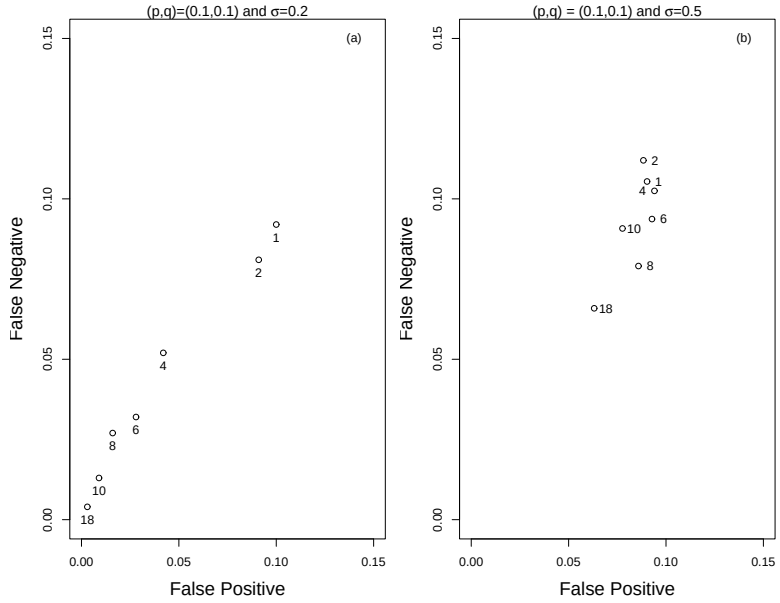

Figure 1.6 *The effect of sample size on the inferred network. The number besides each symbol indicates the number of the time points used in the simulated gene expression data. The value of π_X is 0.46, and the values of other parameters are indicated in the title of each plot.*

1.4 Application to Yeast Cell Cycle Data

In this section, we apply our method to jointly analyze gene expression data from 295 genes over 18 time points (Spellman *et al.* 2002) and protein-DNA binding data of Fkh1, Fkh2, Ndd1, Mcm1, Swi5, Ace2, Mbp1, and Swi4 (Lee *et al.* 2002). We consider eight sets of model parameters for $\{p, q, \pi_X\}$: $\{0.1, 0.1, 0.5\}$, $\{0.2, 0.2, 0.5\}$, $\{0.2, 0.1, 0.5\}$, $\{0.1, 0.2, 0.5\}$, $\{0.2, 0.2, 0.4\}$, $\{0.2, 0.2, 0.6\}$, $\{0.1, 0.1, 0.4\}$, and $\{0.1, 0.1, 0.6\}$. For each set of parameter specifications, we run MCMC with a burn-in of 1,000 runs and an additional 5,000 runs to obtain the posterior distributions for the parameters of interest. The overall inference is based on the average posterior probabilities over the eight model parameter settings, which yield similar results among different settings.

The posterior distributions of the protein activities for the eight TFs and the σ at every time point are summarized in Table 1.1. The average value of σ across 18 time points is about 0.55. Based on our simulation studies, at this level of expression errors, the incorporation of gene expression data should improve the inference of TRNs.

1.5 Summary

In this article, we have developed a misclassification model to integrate gene expression data and protein-DNA binding data to infer TRNs. Compared to other models, our model (1) integrates gene expression data and protein-DNA binding data through a consistent framework, (2) considers the misclassification associated with protein-DNA binding data explicitly, and (3) consists of a flexible model structure. The systematic simulation results indicate that this model performs well in the reconstruction of the underlying networks when the misclassification associated with gene expression data and (more importantly) protein-DNA binding data are within reasonable ranges. For example, in the case of less than 30% to 40% false-positive and false-negative rates in the observed binding data, our method may significantly reduce both types of error rates in the inferred network when the standard deviation in gene expression measurements is around 0.5 or less. In all the cases, the inclusion of gene expression data leads to improved inference of the underlying network compared to that solely based on the binding data even when the measurement error in gene expression data is very high.

In this article, we have considered five TFs in simulation studies and eight TFs in the application to the yeast cell cycle data. Because there are 133 TFs in yeast protein-DNA binding data, the inclusion of all TFs in the same model will create both statistical and computation challenges. In the context of yeast cell cycle data, protein-DNA binding data suggest that close to 20 TFs may be involved in the regulation of cell cycle genes (data not shown). The results of the application of our method to a more complete TF set and biological interpretations of the results will be reported in a separate article. From this study, we have found that (1) protein-DNA binding data can serve as a good starting point in the proposed MCMC procedure, and (2) the

Table 1.1 *The estimates of the regulation activities of the transcription factors and σ based on our model.*

Time Point	Fkh1	Fkh2	Ndd1	Mcm1	Ace2	Swi5	Mbp1	Swi4	σ
1	0.09	-0.81	-0.55	0.54	1.84	-0.29	-0.79	-0.27	0.88
	±0.13	±0.12	±0.13	±0.13	±0.14	±0.13	±0.12	±0.12	
2	-0.36	-1.00	0.24	0.28	1.18	-0.46	-0.18	-0.01	0.75
	±0.11	±0.11	±0.11	±0.11	±0.13	±0.12	±0.10	±0.11	
3	-0.53	-0.63	0.14	0.09	0.98	-0.35	1.43	0.06	0.66
	±0.10	±0.10	±0.10	±0.10	±0.14	±0.11	±0.09	±0.10	
4	-0.34	-0.31	-0.25	-0.29	0.17	-0.42	1.86	0.27	0.58
	±0.08	±0.09	±0.09	±0.08	±0.13	±0.10	±0.07	±0.08	
5	0.73	0.12	-0.62	-0.63	0.26	-0.67	0.79	0.13	0.54
	±0.07	±0.08	±0.08	±0.07	±0.09	±0.08	±0.07	±0.08	
6	0.72	0.20	-0.42	-0.49	-0.17	-0.49	0.28	-0.04	0.6
	±0.08	±0.08	±0.09	±0.08	±0.10	±0.09	±0.08	±0.08	
7	1.31	0.16	0.41	-0.61	-0.07	-0.55	-0.28	-0.28	0.53
	±0.08	±0.09	±0.08	±0.08	±0.10	±0.09	±0.08	±0.08	
8	0.44	0.18	0.61	0.01	-0.47	-0.31	-0.43	-0.57	0.44
	±0.06	±0.06	±0.06	±0.06	±0.08	±0.07	±0.06	±0.06	
9	0.17	0.09	1.03	0.58	-0.46	-0.00	-0.57	-0.74	0.5
	±0.07	±0.07	±0.07	±0.07	±0.09	±0.08	±0.07	±0.07	
10	-0.27	-0.48	0.81	0.47	-0.54	1.11	-0.39	-0.42	0.57
	±0.07	±0.08	±0.07	±0.07	±0.10	±0.08	±0.07	±0.07	
11	-0.90	0.02	-0.01	0.79	-0.32	1.23	0.13	0.08	0.75
	±0.10	±0.11	±0.11	±0.10	±0.13	±0.12	±0.10	±0.11	
12	-1.07	0.22	-0.29	0.14	-0.45	0.93	0.56	0.65	0.44
	±0.07	±0.06	±0.07	±0.06	±0.08	±0.07	±0.07	±0.06	
13	-0.20	0.44	-0.82	-0.28	-0.15	0.35	0.16	0.63	0.45
	±0.07	±0.07	±0.07	±0.06	±0.08	±0.07	±0.06	±0.06	
14	-0.35	0.42	-0.68	-0.37	-0.31	-0.08	-0.31	0.52	0.45
	±0.06	±0.07	±0.07	±0.07	±0.08	±0.07	±0.06	±0.06	
15	0.44	0.68	-0.61	-0.51	-0.08	-0.32	-0.44	0.38	0.44
	±0.06	±0.07	±0.07	±0.06	±0.08	±0.07	±0.06	±0.07	
16	0.09	0.59	-0.10	-0.16	-0.58	-0.04	-0.45	0.13	0.6
	±0.08	±0.08	±0.08	±0.08	±0.10	±0.09	±0.07	±0.08	
17	0.26	0.26	0.46	-0.02	-0.27	-0.08	-0.71	-0.26	0.62
	±0.08	±0.09	±0.09	±0.08	±0.10	±0.09	±0.07	±0.08	
18	-0.20	-0.15	0.66	0.48	-0.57	0.44	-0.63	-0.26	0.57
	±0.08	±0.09	±0.09	±0.08	±0.10	±0.10	±0.07	±0.08	

larger the number of gene expression data sets used, the more accurate we expect our procedure performs, especially when the gene expression data have low to moderate measurement errors. Therefore, in general, when the number of TFs increases, we hope to collect more samples on relevant gene expressions. More samples can be achieved by increasing the number of experimental conditions or the number of replicates per experimental condition or both. The advantage of increasing the number of experimental conditions is to introduce more variations of TF activity profiles so as to better infer the underlying network. However, more parameters are needed to specify the model for the additional conditions. We also need to be cautious on how to pool the experiments to infer the TRN. In this work, we have assumed a time independent TRN throughout the yeast cell cycle. This assumption may be true in this context and it allows us to pool information from across all time points. However, the TRN may differ under different conditions, and the transient behavior of the TRN needs to be taken into account when using all the microarray data. The advantage of increasing the number of replicates per condition is to reduce errors associated with measured gene expression levels at each point without introducing more model parameters. In this study, the replicates were not included in the model set-up, however, the flexible structure of our model allows an easy incorporation of such information into the model.

In our simulation studies, we have investigated the sensitivity of our method when some of the model parameters are misspecified, including the prior distribution on the network connections and our belief (measured by p and q) on the quality of protein-DNA binding data. We found that the method is not sensitive to the misspecifications of these model parameters unless the specified model parameters are drastically different from the true model parameters. In the analysis of yeast cell cycle data, we considered eight sets of model parameters and observed general agreements among results from different parameter specifications. In practice, we may take a full Bayesian approach to inferring the network through averaging inferred networks under certain prior distributions for the model parameters.

As discussed above, although we have treated the observed protein-DNA binding data as a 0-1 variable, the observed data are, in fact, continuous. In this case, our model can be modified within the measurement model framework so that the measured and true covariate values are continuous. To specify the prior distribution for the covariate values, we may use normal mixtures or more sophisticated models for the binding intensity. However, the interpretation of the model parameters will be somewhat different if the intensity levels are used because the parameter β_t cannot be simply interpreted as TF activities.

In our model set-up, we assume that all the TFs act additively to affect the transcription levels of their target genes and this linear relationship between TF activities and the normalized expression levels is a key assumption for this model. Because of the complexity in transcription regulation, such as synergistic effects among TFs, a linear model can serve as an approximation at best. Nevertheless, linear models have been used in this context by various authors (Bussemaker *et al.*, 2001; Liu *et al.*, 2002; Wang *et al.*, 2002; Liao *et al.*, 2003; Gao *et al.*, 2004). The potential depar-

ture from linearity may result from synergistic regulation effects of TFs bound to the upstream region of the same gene, and we are in the process of developing statistical approaches for analyzing nonlinear models.

To conclude, we note that our model can be extended in different ways to be more comprehensive and better represent the underlying biological mechanisms. For example, the linear form of the model can be extended to incorporate nonlinear interactions among different TFs as discussed above; the replicates per experiment can be considered into the model to improve the data quality; more prior information or more sophisticated statistical models can be used to construct the prior distribution of the network (π_X). In addition, our general framework has the potential to integrate more data types into the model, such as sequence data and mRNA decay data to further infer the transcriptional regulatory networks.

1.6 Acknowledgment

This work was supported in part by NSF grant DMS-0241160.

1.7 References

T. Akutsu, S. Miyano, S. Kuhara, "Identification of genetic networks from a small number of gene expression patterns under the Boolean network model", The Pacific Symposium on Biocomputing, 4: 17-28, 1999.

T. Akutsu, S. Miyano, S. Kuhara, "Inferring qualitative relations in genetic networks and metabolic pathways", Bioinformatics, 16(8): 727-734, 2000.

Z. Bar-Joseph, G.K. Gerber, T.I. Lee, N.J. Rinaldi, J.Y. Yoo, F. Robert, D.B. Gordon, E. Fraenkel, T.S. Jaakkola, R.A. Young, D.K. Gifford, "Computational discovery of gene modules and regulatory networks", Nature Biotechnology, 21(11): 1337-1342, 2003.

H.J. Bussemaker, H. Li, E.D. Siggia, "Regulatory element detection using correlation with expression", Nature Genetics, 27(2): 167-171, 2001.

J.S. Buzas, T.D. Tosteson, L.A. Stefanski, Measurement Error, Institute of Statistics Mimeo Series No. 2544, 2003.

M. Carey, S.T. Smale, Transcriptional Regulation in Eukaryotes, Cold Spring Harbor Laboratory Press, 1999.

R.J. Carroll, "Measurement error in epidemiologic studies", in Encyclopedia of Biostatistics 2491-2519, 1998.

S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P.O. Brown, I. Herskowitz, "The transcriptional program of sporulation in budding yeast", Science, 282(5389): 699-705, 1998.

E.H. Davidson, J.P. Rast, P. Oliveri, A. Ransick, C. Calestani, C.H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C.T. Brown, C.B. Livi, P.Y. Lee, R. Revilla, A.G. Rust, Z. Pan, M.J. Schilstra, P.J. Clarke, M.I. Arnone, L. Rowen, R.A. Cameron, D.R. McClay, L. Hood, H. Bolouri, "A genomic regulatory network for development", Science, 295(5560): 1669-78, 2002.

H. de Jong, "Modeling and simulation of genetic regulatory systems: a literature review", Journal of Computational Biology, 9(1): 67-103, 2002.

M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, "Cluster analysis and display of genome-wide expression patterns", Proceedings of the National Academy of Sciences USA, 95(25): 14863-14868, 1998.

N. Friedman, M. Linial, I. Nachman, D. Pe'er, "Using Bayesian networks to analyze expression data", Journal of Computational Biology, 7(3-4): 601-620, 2000.

W.A. Fuller, Measurement Error Models, Wiley, New York, 1987.

F. Gao, B.C. Foat, H.J. Bussemaker HJ, "Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data", BioMed Central Bioinformatics. 5(1): 31, 2004.

T.S. Gardner, D. di Bernardo, D. Lorenz, J.J. Collins JJ, "Inferring genetic networks and identifying compound mode of action via expression profiling", Science, 301(5629):102-105, 2003.

A.J. Hartemink, D.K. Gifford, T.S. Jaakkola, R.A. Young, "Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks", The Pacific Symposium on Biocomputing, 6: 422-33, 2001.

A.J. Hartemink, D.K. Gifford, T.S. Jaakkola, R.A. Young, "Combining location and expression data for principled discovery of genetic regulatory network models", The Pacific Symposium on Biocomputing, 7: 437-49, 2002.

T.I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Thompson, I. Simon, J. Zeitlinger, E.G. Jennings, H.L. Murray, D.B. Gordon, B. Ren, J.J. Wyrick, J.B. Tagne, T.L. Volkert, E. Fraenkel, D.K. Gifford, R.A. Young RA, "ranscriptional regulatory networks in Saccharomyces cerevisiae", Science, 298(5594): 799-804, 2002.

S. Liang, S. Fuhrman, R. Somogyi R, "Reveal, a general reverse engineering algorithm for inference of genetic network architectures", The Pacific Symposium on Biocomputing, 3: 18-29, 1998.

J.C. Liao, R. Boscolo, Y.L. Yang, L.M. Tran, C. Sabatti, V.P. Roychowdhury VP, "Network component analysis: reconstruction of regulatory signals in biological systems", Proceedings of the National Academy of Sciences USA, 100(26): 15522-15527, 2003.

X.S. Liu, D.L. Brutlag, J.S. Liu JS, "An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments", Nature Biotechnology, 20(8): 835-839, 2002.

B. Ren, F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T.L. Volkert, C.J. Wilson, S.P. Bell, R.A. Young RA, "Genome-wide location and function of DNA binding proteins", Science, 290(5500): 2306-2309, 2000.

S. Richardson, "Measurent error modelling from a Bayesian perspective", Bulletin of the International Statistical Institute, 1999.

S. Richardson, and W.R. Gilks, "Conditional independence models for epidemiological studies with covariate measurement error", Statistics in Medicine, 12: 1703-1722, 1993.

C.J. Roberts, B. Nelson, M.J. Marton, R. Stoughton, M.R. Meyer, H.A. Bennett, Y.D. He, H. Dai, W.L. Walker, T.R. Hughes, M. Tyers, C. Boone, S.H. Friend, "Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles", Science, 287(5454): 873-880, 2000.

I. Shmulevich, E.R. Dougherty, S. Kim, W. Zhang, "Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks", Bioinformatics, 18(2): 261-274, 2002.

P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization", Molecular Biology of the

Cell, 9(12): 3273-3297, 1998.

N. Sun, H. Zhao, "Genomic approaches in dissecting complex biological pathways", Pharmacogenomics, 5(2): 163-179, 2004.

J. Tegnr, M.K. Yeung, J. Hasty, J.J. Collins, "Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling", Proceedings of the National Academy of Sciences USA, 100(10): 5944-5949, 2003.

W. Wang, J.M. Cherry, D. Botstein, H. Li, "A systematic approach to reconstructing transcription networks in Saccharomycescerevisiae", Proceedings of the National Academy of Sciences USA, 99(26): 16893-16898, 2002.

H. Zhao, B. Wu, N. Sun, "DNA-Protein binding and gene expression patterns", Science and Statistics: A festschrift for Terry Speed, Institutue of Mathematical Statistics Lecture Notes-Monograph Series, Vol 40:259-274, 2003.

Z. Zhang, M. Gerstein, "Reconstructing genetic networks in yeast", Nature Biotechnology, 21(11): 1295-1297, 2003.

Index