University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2004 *Paper* 144

A Statistical Method for Constructing Transcriptional Regulatory Networks Using Gene Expression and Sequence Data

Biao Xing* Mark J. van der Laan†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, xing.biao@gene.com

†Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper144

A Statistical Method for Constructing Transcriptional Regulatory Networks Using Gene Expression and Sequence Data

Biao Xing and Mark J. van der Laan

Abstract

Transcriptional regulation is one of the most important means of gene regulation. Uncovering transcriptional regulatory network helps us to understand the complex cellular process. In this paper, we describe a comprehensive statistical approach for constructing the transcriptional regulatory network using data of gene expression, promoter sequence, and transcription factor binding sites. Our simulation studies show that the overall and false positive error rates in the estimated transcriptional regulatory network are expected to be small if the systematic noise in the constructed feature matrix is small. Our analysis based on 658 microarray experiments on yeast gene expression programs and 46 transcription factors suggests that the method is capable of identifying important transcriptional regulatory interactions and uncovering the corresponding regulatory network structures.

1. INTRODUCTION

Transcriptional regulatory network is an important part of the gene interaction networks. It specifies the interactions among regulatory genes and between regulatory genes and their target genes. Transcriptional regulatory genes produce transcription factors (TF), which are regulatory proteins that regulate the expression levels of target genes by recognizing and binding to specific non-coding DNA segments (so called TF binding motifs) of target genes and initiating the transcription process. Transcriptional regulation is one of the most important means for gene regulations. Uncovering transcriptional regulatory network, therefore, helps us to understand the underlying mechanism of complex cellular process.

Methods have been proposed for discovering transcriptional regulatory networks systematically. Lee et al. (2002) used genome-wide location analysis (Ren et al., 2000) to investigate how yeast transcription factors bind to promoter sequences across the genome, then used the gene-specific TF binding information to identify the transcriptional regulatory network motifs and network structure. Their approach is mainly experiment based. It provides more convincing evidence of TF binding activities. However, evidence of physical binding does not directly imply transcriptional functional activity. Moreover, location analysis is typically based on a particular growth condition (e.g., rich medium). As a result, TF binding patterns specific to other growth conditions may not be observed.

Bar-Joseph et al. (2003) described the GRAM (Genetic Regulatory Modules) algorithm for discovering regulatory networks of gene module, which employs location analysis to identify initial gene modules, then expands them by searching genes with similar expression profiles. The method is essentially the same as Lee et al. (2002), but it recognizes the importance of using gene expression data in finding the transcriptional regulatory networks.

Wang et al. (2002) proposed a more computational approach for decomposing the transcriptional regulatory networks into functional modules and making inference on the activation of these modules or interaction between them based on correlation analysis. The construction of the transcriptional functional modules, however, depends on the so called transcription factor perturbation experiment (TFPE), in which the only perturbation is dele-

tion, mutation or over-expression of a transcription factor. One limitation of the use of TFPEs is that the availability of TFPEs is limited. Only 28 modules were constructed based on 28 available TFPEs. Moreover, due to the complex nature of transcriptional regulatory network and the discrete nature of the measurements on time scale, a TFPE does not guarantee that the gene expression changes are attributable solely and directly to the TF being perturbated. The authors themselves also noted that "the most significant motif identified in a TFPE might not necessarily be the motif directly bound by the factor (perturbated)". We think that it is more appropriate to view a microarray experiment as a realization of a certain part of the whole transcriptional regulatory network, which is activated under the experiment condition. Even with a perturbation experiment in which a TF is over-expressed, the activated part of the transcriptional regulatory network should consist of a bunch of regulatory genes functioning via a network structure rather than just the one being perturbated. Therefore, we think that constructing modules for individual TF based on a perturbation experiment may not be effective for the purpose of uncovering the underlying network structure.

Other available methods include reverse engineering approach (Somogyi et al., 1997; Liang et al., 1998; D'Haeseleer et al., 2000), differential equations (Chen et al., 1999; D'Haeseleer et al., 1999), Bayesian networks (Friedman et al., 2000; Yoo et al., 2002), machine learning by SVM (Qian et al., 2003), etc. These methods may work for certain problems or situations, but they usually require large number of time-course data or lack of computational stability. We do not discuss them in details due to the space limit.

In the next section we describe a purely statistical method for uncovering the transcriptional regulatory network based on gene expression data, promoter sequences, and knowledge of TF binding sites. The method identifies active TFs and estimates the corresponding active part of the transcriptional regulatory network under each experiment condition, then average over different experiments to infer the overall network structure. We conduct simulation studies to demonstrate the performance of the proposed method. The results are summarized in Section 3. In Section 4, we apply the method to the yeast data to study the yeast transcriptional regulatory network. We

conclude with a discussion of the advantages and limitations of the proposed method in Section 5.

2. METHOD

2.1 *Input data*

The input data include gene expression data, denoted by Y, DNA sequence data for the transcriptional control region (TCR) of the genes, denoted by S, and TF binding motifs data, denoted by W.

More specifically, $Y = \{Y_{ji} : j = 1, \ldots, J; \quad i = 1, \ldots, I.\}$ is a J by I matrix, where Y_{ji} is the gene expression measurement for the j^{th} gene under the i^{th} experiment condition. In other words, Y is a collection of experiments under various conditions and not necessarily time-course data.

$S = \{S_{jl} : j = 1, \ldots, J; \quad l = 1, \ldots, L.\}$ is a J by L matrix, where S_j is the DNA sequence extracted from the TCR of the j^{th} gene and L is the length of the sequences. (For simplicity, we let the sequences to be of the same length L.)

$W = \{W_t : t = 1, \ldots, T\}$ is a vector containing binding motifs specific to T distinct TFs.

2.2 *Feature matrix X*

The feature matrix X, which measures gene-specific oligomer motif abundance, is created by matching TF binding motifs W to the sequence data S. The most simplest way to construct X is to define X_{jt} as the count of the occurrences of the t^{th} motif in the j^{th} sequence, i.e.,

$$X_{jt} = \sum_{l=1}^{L-w(t)+1} I[S_{j,\,l\,:\,l+w(t)-1} = h(W_t)] \tag{1}$$

where $w(t) = |W_t|$ is the length of the t^{th} motif, $h(W_t)$ allows for degenerated representation of W_t and its reverse complement, and l is updated by $l + w(t) - 1$ if the indicator function $I(\cdot)$ returns 1.

Alternatively, X may be constructed by incorporating information on both motif counts and motif locations (Keles et al., 2002) or using a position weight matrix and a background model with Markov dependency (Conlon et al., 2003). For the purpose of motif detecting, using a position weight

matrix and a Markov background model may improve the sensitivity and specificity. However, for the purpose of scoring a known motif, defining X as in Equation (1) is time-wise more efficient.

2.3 *Identifying active transcription factors*

Bussemaker et al. (2001), Keles et al. (2002) and Conlon et al. (2003) have shown that by regressing genome-wide gene expression measures over gene-specific oligomer motif abundance measures, one can identify the motifs (and thereby the corresponding transcription factors) that are likely to be active and responsible for the dramatic changes in the expression levels of their target genes under the current experiment condition. We adopt the same idea and describe two approaches for identifying active transcription factors under an experiment condition.

2.3.1 Multiple linear regression model selected by a loss-based V-fold cross-validation model selector The basic idea of this approach is to build a multiple linear regression model as follows using a single gene expression experiment and the motif abundance measure matrix X to identify the most significant motifs and the corresponding transcription factors under the given experiment condition:

$$y_j = \beta_0 + \sum_{t \in \tau(i)} \beta_t X_{jt} + \epsilon_j \tag{2}$$

where y_j is the absolute value of the expression level for the j^{th} gene, X_{jt} is the binding motif abundance measure for the t^{th} transcription factor in the promoter region of the j^{th} gene, β's are the regression coefficients, ϵ_j is gene-specific random error, $\tau(i) \subseteq \{1, \ldots, T\}$ is the set of transcription factors that are active under the i^{th} experiment condition, and $j = 1, \ldots, J$.

Note that using the absolute value of gene expression measure enables us to model the situation in which a transcription factors serves both as an activator to some genes and a repressor to some other genes under the same experiment condition.

An explicit assumption is that a transcription factor is active under the current experiment condition if its binding motif is significantly associated with the changes in the genome-wide gene expressions.

We use a loss-based V-fold cross-validation selector to find the best model for a given experiment. A natural choice of the loss function for the conditional mean model as Equation (2) is the squared error loss function given by

$$L(X, Y, \psi) = [Y - \psi(X)]^2 = [Y - E(Y|X)]^2,$$

where ψ is a function mapping from covariate space into outcome space. For the model selection purpose, we wish to estimate the true model, ψ_0, which minimizes the expected loss (i.e., risk)

$$E_{P_0} L(X, Y, \psi) = \int L(x, y, \psi) dP_0$$

with respect to the unknown true data generating distribution $P_0 = P_0(X, Y)$.

The basic idea of the V-fold cross-validation model selection is that the data is randomly divided into V mutually exclusive and exhaustive sets, each used in turn as the validation set and the remaining sets used as the training set. Denote the random split vector by $S_n = \{S_{n,i} : i = 1, \ldots, n\}$, where $S_{n,i} = 0$ if the i^{th} observation is in the training set and $S_{n,i} = 1$ if it is in the validation set. For the V-fold cross-validation, we have V realizations of S_n which satisfies that $\sum_i S^v_{n,i} \approx n/V$ and $\sum_v S^v_{n,i} = 1$, and each of the V split vector has a probability mass of $1/V$.

Let P^0_{n,S_n} and P^1_{n,S_n} denote the empirical distributions of the training and validation sets, respectively. For the conditional mean model as defined by Equation (2) with the squared error loss function used, the loss-based V-fold cross-validation model selector can be explicitly written as

$$\hat{k} = \text{argmin}_k \ \frac{1}{V} \sum_{v=1}^{V} \frac{1}{\sum_i S^v_{n,i}} \sum_{\{i: S^v_{n,i}=1\}} [y_i - \psi_k(x_i | P^0_{n,S^v_n})]^2, \tag{3}$$

where V is the number of splits, S^v_n is the v^{th} split vector, $\psi_k(\cdot|P^0_n) \in \Psi$, $k = 1, \ldots, K$, is a collection of candidate estimators of $\psi_0(\cdot)$ that are obtained based on only the training set. The expected loss is evaluated using only the validation set.

There are many ways to generate a set of candidate estimators for ψ_0. Here we describe a forward selection algorithm to generate a sequence of

nested candidate models. The procedure go as follows: Begin with the null model (with only intercept). First identify the variable that, if added to the model, contributes the most to the reduction in mean square error (MSE). Keep the variable and obtain a nested supper-model. Repeat this procedure until reaching the user-specified model size K. In this way, we generate a set of nested models with increasing dimensions. Index the candidate models by $k = 1, \ldots, K$.

We then use the cross-validation procedure to select from the candidate models the best one that minimizes the expected loss. The selected model identifies the set of TFs (denoted by $\tau(i)$) that are significantly associated with the gene expression changes and thus assumed to be active under the given experiment condition.

The loss-based cross validation model selector is asymptotically optimal and unbiased for estimation of the expected loss. We refer to Breiman et al. (1984) and van der Laan and Dudoit (2003) for more detailed theoretical dicsussions.

2.3.2 Simple linear regression model followed by a multiple testing procedure Alternatively, we can identify active TFs by fitting simple linear models and using a multiple testing procedure. The idea is first to fit a simple linear model as follows for every TF for a single gene expression experiment:

$$y_j = \beta_0 + \beta_1 X_{jt} + \epsilon_{jt}, \tag{4}$$

where y_j is the absolute value of the expression level of the j^{th} gene in the i^{th} experiment, X_{jt} is the motif abundance measure for the t^{th} transcription factor in the promoter of the j^{th} gene, β's are the regression coefficients, ϵ_{jt} is the random error, and $j = 1, \ldots, J$ and $t = 1, \ldots, T$.

We take the p-value of a model as a statistic indicating the significance of the association between the TF and the gene expression changes. For computational convenience, we may simply assume a normal model to calculate the p-value. In this way, we obtain a vector of p-values for all the TFs for a given experiment. Denote it by $\vec{p} = \{p_t : t = 1, \ldots, T\}$.

Next we take $\vec{p}$ as an input and employ a multiple testing procedure to select a subset of the TFs that are significantly associated with the gene

expression changes by a specified criterion. Candidate multiple testing procedures include those single- or multiple-step procedures controlling for the (generalized) family-wise error rate, false discovery rate (FDR), etc (Dudoit et al., 2003; Storey, 2003; van der Laan et al., 2004).

Here we describe a simple procedure to control the false discovery rate (defined as the expected proportion of false rejections) proposed by Benjamini and Hochberg (1995). Suppose we wish to test simultaneously null hypotheses $H_1, \ldots, H_T$ based on p-values $p_1, \ldots, p_T$. Let $\{r_1, \ldots, r_T\}$ be a mapping to $\{1, \ldots, T\}$ such that $p_{r_1} \leq p_{r_2} \leq \ldots \leq p_{r_T}$. Let q be the FDR level we wish to control. Solve

$$k = \text{argmax}_{\{t=1,\ldots,T\}} \; p_{r_t} \leq \frac{t}{T} q. \tag{5}$$

Define

$$\tau(i) = \begin{cases} r_t : t = 1, \ldots, k, & \text{if k is defined,} \\ \emptyset, & \text{otherwise.} \end{cases} \tag{6}$$

$\tau(i)$ is the rejection set for the i^{th} experiment. The transcription factors in $\tau(i)$ are significantly associated with gene expression changes by the specified multiple testing control criterion and thus assumed to be active under the experiment condition.

Benjamini and Hochberg (1995) have shown that the procedure controls the FDR at level q for any configuration of false null hypotheses and independent test statistics.

2.3.3 Remarks Since the genome-wide gene expression measures are used in building the regression model, both of the two approaches can only identify those transcription factors that potentially cause dramatic changes in the expression levels in target genes and therefore result in significant changes in the genome-wide gene expression profile. Both methods may fail to identify those transcription factors that have only subtle effects on the changes of genome-wide gene expression profile.

The results from the two approaches tend to be similar, but may not be exactly the same. Since the second approach is time-wise much more efficient, we recommend to use the first approach only when the number of

experiments and the numbers of TFs involved in the analysis are small or moderate. Otherwise, we recommend the use of the second approach.

2.4 *Identify target genes of active TFs*

The necessary conditions that a gene is significantly regulated by a transcription factor under a certain experiment condition include at least: (1) the upper stream region of the gene must be abundant with the transcription factor specific binding motif(s), for example, containing at least one copy of the binding motif; (2) the transcription factor is active under the experiment condition; and (3) the expression level of the gene is significantly different from zero, e.g., a 2-fold change.

Motivated by this reasoning, we propose the following procedures to identify the target genes of active transcription factors for a given experiment condition.

2.4.1 Gene expression data transformation Denote the gene expression data for the i^{th} experiment by $\vec{Y_i} = \{Y_{ji} : j = 1, \ldots, J.\}$. Transform the vector $\vec{Y_i}$ into a matrix $Z^{(i)} = \{Z^{(i)}_{jt} : j = 1, \ldots, J; t = 1, \ldots, T.\}$ according to

$$Z^{(i)}_{jt} = \begin{cases} Y_{ji}, & \text{if } t \in \tau(i) \text{ and } X_{jt} \geq 1, \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where $\tau(i) \subseteq \{1, \ldots, T\}$ is the set of transcription factors that are active under the i^{th} experiment condition (see Section 2.3.1 for definition).

As a result, the non-zero entries of the t^{th} column of matrix $Z^{(i)}$ are the potential target genes regulated by the t^{th} transcription factor under the i^{th} experiment condition.

2.4.2 Classification using a normal mixture model Since not all the potential target genes are significantly regulated by an active transcription factor under a particular experiment condition, we propose a classification procedure to identify those genes that are likely to be significantly regulated by the TF under the experiment condition using a 3-component normal mixture model. The normal mixture model is used because of computational convenience and the fact that the microarray data of gene expression are all

normalized such that the data are centered about the null.

The basic idea is that the non-zero entries of a column of $Z^{(i)}$ are seen as generated from a mixture of three normal distributions, which characterize the model for the target genes that are repressed, not significantly regulated, and induced under the experiment condition, respectively. Let $M \in \{1, 2, 3\}$ denotes these three situations (classes). M is not observed at all and treated as a missing variable.

Denote the t^{th} column of $Z^{(i)}$ by $\vec{Z}_t^{(i)}$. For simplicity we write it as $\vec{Z}_t$. If the t^{th} transcription factor is not active under the i^{th} experiment condition, the elements in $\vec{Z}_t$ are all zero and no further classification procedure is needed. Otherwise, let $\vec{Z}_t^*$ be the vector of non-zero elements of $\vec{Z}_t$.

Let $\theta = \{\pi_m, \mu_m, \sigma_m^2 : m = 1, 2, 3\}$ be the parameter of the mixture model, where π_m, μ_m and σ_m^2 are the mixing proportion, mean and variance for the m^{th} component distribution, respectively, subjected to the constraint that $\sum_{m=1}^{3} \pi_m = 1$. For convenience, we assume that the observations are independent and the true class labels are missing at random. (Although the actual gene expression data are not independent, we believe the independence assumption will not compromise the classification accuracy too much.) Then we can write the density of the marginal distribution of Z_{jt}^* given θ as follows

$$f(Z_{jt}^*|\theta) = \sum_{m=1}^{3} \pi_m \phi(Z_{jt}^*|\mu_m, \sigma_m^2),$$

where $\phi(\cdot)$ denotes the density function of the normal distribution, and $m \in \{1, 2, 3\}$ indexes the three components of the mixture.

The observed data log-likelihood is given by

$$\ell(\theta|\vec{Z}_t^*) = \sum_{j \in \vec{Z}_t^*} \log(\sum_{m=1}^{3} \pi_m \phi(Z_{jt}^*|\mu_m, \sigma_m^2)),$$

and the complete data log-likelihood is given by

$$\ell_c(\theta|\vec{Z}_t^*, \vec{M}_t^*) = \sum_{j \in \vec{Z}_t^*} \sum_{m=1}^{3} I(M_{jt}^* = m) \log(\pi_m \phi(Z_{jt}^*|\mu_m, \sigma_m^2)).$$

An EM algorithm (Dempster et al., 1977) can be used to estimate the model parameters iteratively. The algorithm iterates by alternately repeating the so-called E-step and M-step. In E-step, the expected complete data log-likelihood given the current parameter $\theta^{(k)}$ is computed as follows

$$Q(\theta|\theta^{(k)}) = E[\ell_c(\theta|\vec{Z}_t^*, \vec{M}_t^*)|\theta^{(k)}] = \sum_{j\in\vec{Z}_t^*}\sum_{m=1}^{3}\gamma_{jm}^{(k)}\log(\pi_m^{(k)}\phi(Z_{jt}^*|\mu_m^{(k)}, \sigma_m^2(k))),$$

where

$$\gamma_{jm}^{(k)} = P(M_{jt}^* = m|Z_{jt}^*, \theta^{(k)}) = \frac{\pi_m^{(k)}\phi(Z_{jt}^*|\mu_m^{(k)}, \sigma_m^2(k))}{\sum_{l=1}^{3}\pi_l^{(k)}\phi(Z_{jt}^*|\mu_l^{(k)}, \sigma_l^2(k))}$$

is the conditional expectation of $M = m$ given data and the current parameter. In M-step, the parameter is updated by

$$\theta^{(k+1)} = \text{argmax}_{\theta\in\Theta}\ \ Q(\theta|\theta^{(k)}).$$

The iteration stops when convergence is reached or when some other stopping rule is satisfied. We then classify the potential target genes of an active transcription factor into three classes: 'repressed', 'induced', and 'not significantly regulated', based on the posterior probabilities $\gamma_{jm} = P(M_{jt}^* = m|Z_{jt}^*, \hat{\theta})$, where $\hat{\theta}$ is the estimated model parameter.

If a potential target gene of an active transcription factor is classified as either 'repressed' or 'induced', we say that there is a transcriptional regulatory interaction between the TF and the gene under the given experiment condition. This definition assumes that a TF serves as both an inducer and a repressor in the same experiment. If we assume that a TF plays primarily a single role as an inducer or a repressor but not both in one experiment, we can first determine whether a TF is primarily an inducer or a repressor by fitting a multiple linear model using the selected TFs with the dependent variable being the original expression value, then looking at the sign of the regression coefficient corresponding to the TF of interest. If the coefficient is positive, we say that the TF is an inducer and we infer that the genes in the 'induced' class are transcriptionally regulated by the TF. If the coefficient is negative, we say that the TF is a repressor and we infer that the genes in the 'repressed' class are transcriptionally regulated by the TF.

After implementing this classification procedure to every column of the matrix $Z^{(i)}$, we obtain an experiment-specific transcriptional regulatory interaction matrix (TRIM), denoted by $B^{(i)} = \{B^{(i)}_{jt} : j = 1, \ldots, J; t = 1 \ldots, T.\}$, where $B^{(i)}_{jt}$ is the posterior probability that the j^{th} gene is transcriptionally regulated by the t^{th} transcription factor under the i^{th} experiment condition. By applying a cut-off (e.g., 0.50), we can convert the probability matrix to a binary matrix whose elements indicate whether a TF transcriptionally regulates a gene.

2.5 *Constructing the all-condition transcriptional regulatory interaction matrix*

Consider a hypothetical situation in which all transcription factors are active. Denote the corresponding all-condition transcriptional regulatory interaction matrix by B. The scientific question of uncovering transcriptional regulatory network is statistically equivalent to constructing the hypothetical transcriptional regulatory interaction matrix B.

Due to the complexity of the transcriptional regulatory network and the discrete nature of the gene expression experiments on time scale, a single microarray experiment carries only partial information on a particular part of the transcriptional regulatory network, which involves only a subset of TFs that are active under the experiment condition. Accordingly, we view the experiment-specific TRIM $B^{(i)}$ as a partial realization of the all-condition TRIM B. More specifically, we view $B^{(i)}$ as a realization of a particular set of columns of B, which correspond to the transcription factors that are active under the i^{th} experiment condition.

Suppose we have a collection of I experiments. We perform above procedures to obtain experiment-specific TRIM $B^{(i)}, \ldots, B^{(I)}$, and experiment-specific set of active transcription factors $\tau(1), \ldots, \tau(I)$ (see Section 2.3.1 for definition of $\tau(i)$).

Define $h(t) = \sum_{i=1}^{I} I(t \in \tau(i))$ for $t = 1, \ldots, T$. $h(t)$ is a count of how many times the t^{th} transcription factor is active among the I experiments. We then estimate B as follows

$$B_{jt} = \begin{cases} \frac{1}{h(t)} \sum_{i=1}^{I} B^{(i)}_{jt}, & \text{if } h(t) > 0, \\ 0, & \text{if } h(t) = 0. \end{cases} \tag{8}$$

Note that B_{jt} estimated using formula (8) is the experiment-weighted probability that the t^{th} transcription factor transcriptionally regulates the j^{th} gene. We can further transform the matrix into an indicator matrix by letting

$$B_{jt} = I(B_{jt} \geq c) \qquad \text{for j=1,...,J and t=1,..., T,} \tag{9}$$

where $I(\cdot)$ is an indicator function and $c \in (0, 1)$ is a user-specified cut-off. The bigger c is, the more conservative we are in characterizing the transcriptional regulatory interactions.

The binary version of the TRIM B is a convenient form for representing the transcriptional regulatory network, which can be translated into graphical network structure using the algorithm described in Section 2.6.

2.6 *Finding network motifs*

Network motifs are the simplest units of the network architecture, which suggest models for regulatory mechanism that can be tested. Lee et al. (2002) described six regulatory network motifs in terms of binding (see Figure 1) and algorithms to find them. We redefine the network motifs in terms of transcriptional regulatory interaction as follows: (a) autoregulation motif, in which a regulator gene regulates its own expression; (b) feedforward loop motif, in which a master regulator regulates the second regulator and both regulate a common target gene; (c) multi-component loop motif, in which regulator(1) regulates regulator(2), ..., regulator(n-1) regulates regulator(n), and regulator(n) regulates regulator(1), where $n \geq 2$; (d) single input motif, in which a single regulator uniquely regulates a set of target genes; (e) multi-input motif, in which a set of regulators regulate a set of target genes together; and (f) regulator chain motif, in which regulator(1) regulates regulator(2), ..., regulator(n-1) regulates regulator(n), where $n \geq 2$ and the chain ends if regulator(n) does not directly regulate any other regulator that is not on the chain.

We adopted the same idea as Lee et al. (2002) and developed R/S-plus based programs to find the network motifs. The input data is the binary version of the all-condition transcriptional regulatory interaction matrix B, which can be obtained using method described in Section 2.5. A square matrix R, also referred as to the regulator matrix, is extracted from B in

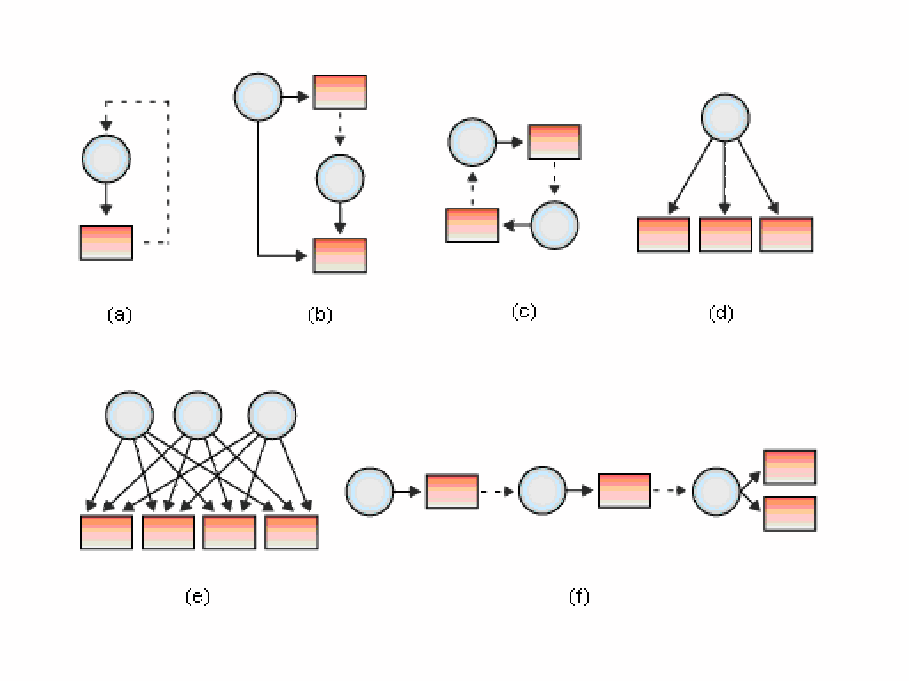

Figure 1. Transcriptional regulatory network motifs: (a) Auto-regulation, (b) Feed-forward loop, (c) Multi-component loop, (d) Single-input motif, (e) Multi-input motif, and (f) Regulator chain motif. Transcription factors are indicated by blue circles and genes by orange boxes. Solid arrows indicate regulatory interaction between TFs and their target genes. Dashed arrows link TFs and their producer genes. The diagram is modified from Lee et al. (2002).

a way such that the rows of R correspond to the set of genes that produce the transcription factors in the columns of R, listed in the same order. So $R \subset B$.

The algorithms to find the transcriptional regulatory network motifs are as follows:

1. Auto-regulation motif:
   Find all t such that $t \in \{1, \ldots, T\}$ and $R_{tt} = 1$. In other words, find all non-zero entries on the diagonal of matrix R. Each of them is an auto-regulatory motif.

2. Feed-forward loop motif:
   Find all (t_1, t_2, j) such that $R_{t_2,t_1} = 1$, $B_{j,t_1} = 1$ and $B_{j,t_2} = 1$, where $t_1, t_2 \in \{1, \ldots, T\}, t_1 \neq t_2$ and $j \in \{1, \ldots, J\}$. In other words, for each column of R (master regulator t_1), find all rows of R (secondary regulators) that t_1 regulates. For each master and secondary regulator pair (t_1, t_2), find all rows (i.e., genes indexed by j) in matrix B regulated by both regulators.

3. Multi-component loop motif:
   Find all $(t_1, \ldots, t_n)$ such that $R_{t_2,t_1} = 1, \ldots, R_{t_n,t_{n-1}} = 1$ and $R_{t_1,t_n} = 1$, where $t_1, \ldots, t_n \in \{1, \ldots, T\}$ and $t_1 \neq \ldots \neq t_n$. In other words, for each regulator (column of R), find its target regulators (rows of R). For each of the target regulators (corresponding column of R), find the target regulators (rows of R) of the target regulator. Repeat this until the target regulator is the same as the original.

4. Single input motif (SIM):
   Step 1, find the set $\omega = \{j : j \in \{1, \ldots, J\} \text{ and } (\sum_{t=1}^{T} B_{jt}) = 1\}$, which are genes that are uniquely regulated by a regulator. This is equivalent to taking the subset of rows of B such that the row sum is 1. Step 2, find the set $\omega_{(t)} = \{j : j \in \omega \text{ and } B_{jt} = 1\}$, which are genes that are uniquely regulated by regulator t. If the size $|\omega_{(t)}| \geq 1$, then $(t, \omega_{(t)})$ is a single input motif. Repeat Step 2 and find single input motifs for all $t \in \{1, \ldots, T\}$.

5. Multi-input motif (MIM):
   Step 1, find the set $\nu = \{j : j \in \{1, \ldots, J\} \text{ and } (\sum_{t=1}^{T} B_{jt}) > 1\}$, which are genes that are regulated by more than one regulator. This is equivalent to taking the subset of rows of B such that the row sum is > 1. Step 2, find the set $\nu_{(\vec{t})} \subset \nu$ such that $\vec{B}_l = \vec{B}_m$ for any row $l, m \in \nu_{(\vec{t})}$ and $l \neq m$. Then $(\vec{t}, \nu_{(\vec{t})})$ is a multi-input motif. This is equivalent to finding the genes (rows) in ν that are regulated by the same set regulators $(\vec{t})$. After identifying an MIM, let $\nu = \nu - \nu_{(\vec{t})}$, then repeat Step 2 until finding all possible MIMs.

6. Regulator chain motif:
   Find all $(t_1, \ldots, t_n)$ such that $R_{t_2,t_1} = 1$, ..., $R_{t_n,t_{n-1}} = 1$, and $R_{l,t_1} = 0$ for all $l \in \{1, \ldots, T\}$ except for $l = t_1$, $R_{m,t_n} = 0$ for all $m \in \{\{1, \ldots, T\} - \{t_1, \ldots, t_n\}\}$, where $t_1, \ldots, t_n \in \{1, \ldots, T\}$ and $t_1 \neq \ldots \neq t_n$. The algorithm involves the following steps: Step 1, find a possible starting regulator (t_1) of the chain such that it is regulated by no other regulators in the list except for itself. Step 2, find the target regulator t_k for regulator t_{k-1}. The recursive procedure stops when the regulator at the end of the chain does not directly regulate any other regulator that is not on the chain except for itself or some earlier regulators on the chain.

3. SIMULATION STUDIES

We conduct simulations to show how the proposed computational approach performs in re-constructing the underlying regulatory network structure. The parameter of interest is the transcriptional regulatory interaction matrix B, which may be regarded as a 2-dimensional representation of the underlying network. In practice, we don't know B. But in simulations, suppose we know B. We wish to estimate B using the above described approach, and assess the error in the estimation.

3.1 *Construct a fictitious regulatory network*

We consider a fictitious transcriptional regulatory network consisting of 10 TFs and 150 genes. For simplicity, suppose that five of the TFs are inducers and the other five are repressors. Also suppose that 50 genes are regulated

by at least one inducer but no repressors, another 50 regulated by at least one repressor but no repressors, and the remaining 50 genes regulated by none of the 10 TFs. We randomly construct a binary-valued transcriptional regulatory interaction matrix B, which satisfies the above condition.

3.2 *Construct a fictitious feature matrix*

Next we construct a fictitious feature matrix X, which measures the abundance of binding sites of the 10 fictitious TFs. A necessary condition for the t^{th} TF transcriptionally regulates the j^{th} gene is that the j^{th} gene must have at least one binding site for the t^{th} TF. In other words, $B_{jt} = 1$ implies that $X_{jt} > 0$. It is also true that $X_{jt} = 0$ implies that $B_{jt} = 0$. Assuming that transcriptional regulatory interaction between a TF and a gene is positively related to the abundance of the TF-specific binding sites, we then use the following rules to construct the feature matrix X:

- If $B_{jt} = 1$, then $X_{jt} \sim$ Uniform $\{2, 3, 4, 5\}$;
- If $B_{jt} = 0$, then $X_{jt} \sim$ Bernoulli $\{0, 1\}$ with $P(X_{jt} = 1) = \delta$.

Note the situation that $B_{jt} = 0$ and $X_{jt} > 0$ (i.e. a TF does not regulate a gene even though the gene promoter is abundant with binding sites of the TF) is regarded as systematic error. We consider three values for δ, i.e., $\delta = 0.10, 0.30, 0.50$, representing small, moderate and large systematic error in the feature matrix X, respectively.

3.3 *Estimation with a single experiment*

We first wish to see how the method estimates the partial transcriptional regulatory network under one experiment condition. We randomly choose a subset of TFs, denoted by τ^*, assuming the size of τ^* is $|\tau^*| \sim$ Uniform $\{3, \ldots, 7\}$. τ^* represents a particular experiment condition in which only the TFs in τ^* are active. The true transcriptional regulatory interaction matrix corresponding to τ^*, denoted by B^*, is a partial realization of the overall true transcriptional regulatory interaction matrix B, which satisfies that $B^*_{jt} = B_{jt}$ if $t \in \tau^*$ and $B^*_{jt} = 0$ otherwise.

We generate one set of fictitious gene expression data using a multiple

linear model as follows

$$Y_j = \beta_0 + \sum_{t \in \tau^*} \beta_t X_{jt} + \epsilon_j,$$

where j indexes genes, t indexes TFs, β's are coefficients and ϵ_j is the gene-specific random error.

For simplicity, we assume $\beta_0 = 0$, $\vec{\beta_t}$ = (0.25,0.30, 0.35, 0.40, 0.45,-0.25, -0.30, -0.35, -0.40, -045), and $\epsilon_j = \epsilon \sim N(0, \sigma^2)$. We consider three values for σ, i.e., $\sigma = 0.25, 0.50, 0.75$, representing small, medium and large random errors in microarray measurements.

We estimate $\hat{B}^*$ based on the generated data and compute the overall error rate and false positive rate as defined in Section 3.5. We repeat the procedures 100 times and get average estimates of the error rates.

3.4 *Estimation with a collection of experiments*

Next we generate data that resemble the situation that we have a collection of $I = 50$ experiments. Each experiment is seen as a realization of certain part of the true underlying regulatory network. Thus by averaging over all the experiments, we expect to uncover the underlying transcriptional regulatory interaction matrix B.

To do so, for each $i = 1, \ldots, I$, we draw a random subset $\tau(i) \subseteq \{1, \ldots, T\}$, with a random size $|\tau(i)| \sim$ Uniform $\{3, \ldots, 7\}$.

The fictitious gene expression data are generated using a multiple linear model as follows

$$Y_{ji} = \beta_0 + \sum_{t \in \tau(i)} \beta_t X_{jt} + \epsilon_{ji},$$

where i indexes experiments, $\tau(i)$ is the set of TFs that are active under the i^{th} experiment, and other notations are the same as before.

We estimate $\hat{B}$ by averaging over all experiments, and compute the error rates. We repeat the procedures 100 times and obtain the average error rates.

3.5 *Error in estimation*

To assess the error in estimation, we define the overall error rate as

$$err_1 = \frac{1}{J \times T} \sum_{j,t} I(B_{jt} \neq \hat{B}_{jt}),$$

and the false positive rate (FPR) as

$$err_2 = \sum_{j,t} I((B_{jt} = 0) \text{ and } (\hat{B}_{jt} = 1))/\sum_{j,t} I(\hat{B}_{jt} = 1).$$

A small false positive rate implies less error nodes in the constructed network. A small false positive rate plus a small overall error rate imply that the constructed network is more complete and has less error nodes.

3.6 *Simulation results*

The simulation results are shown in Table 1, where $\epsilon \sim N(0, \sigma^2)$ denotes random error in gene expression measurements, with $\sigma = 0.25, 0.5, 0.75$, for small, moderate and large random error, respectively and $\delta = 0.1, 0.3, 0.5$, for small, moderate and large systematic error in the feature matrix X, respectively.

In the first case, we are trying to estimate certain part of the underlying transcriptional regulatory network that is active under one experiment condition. In the second case, we are trying to estimate the overall transcriptional regulatory network based on a collection of 50 experiments. In both cases we see that both the overall error rate and the false positive rates increase as the systematic error increases. They also tend to increase as the random error increases for the estimations based on a single experiment, but seem not to change much for the estimations based on a set of different experiments. The overall error rate is pretty small even when the systematic and/or random error is large. The false positive rate is also small when the systematic and random errors are small. But it can be moderately big when the systematic and random errors become large. The false positive error rate also tends to be smaller for the estimation based on a collection of experiments than that that based on a single experiment.

In the simulation, we used $c = 0.5$ as a cut-off to convert the estimated regulatory interaction probability matrix into an indicator matrix. We noted that the choice of cut-off value plays a very important role in both the direction and magnitude of the error rates. A conservative choice of the cut-off value tends to result in small false positive rates, and may increase the overall error rates if the proportion of genes that are significantly regulated by the TFs is relatively big. A less conservative cut-off value tends to result in

Table 1. Average Error Rates in the Estimated Transcriptional Regulatory Interaction Matrices

	Sys. Error	$\epsilon \sim N(0, 0.25^2)$		$\epsilon \sim N(0, 0.50^2)$		$\epsilon \sim N(0, 0.75^2)$	
	δ	Overall	FPR	Overall	FPR	Overall	FPR
Single	0.10	0.0182	0.0940	0.0185	0.0912	0.0303	0.1486
experi-	0.30	0.0360	0.2105	0.0406	0.2276	0.0563	0.2973
ment	0.50	0.0397	0.2139	0.0503	0.2663	0.0637	0.3291
A set of	0.10	0.0101	0.0276	0.0097	0.0276	0.0104	0.0279
50 experi-	0.30	0.0408	0.1372	0.0398	0.1310	0.0394	0.1253
ments	0.50	0.0548	0.1872	0.0559	0.1875	0.0561	0.1831

increased false positive rates, and may reduce the overall error rates if the proportion of genes that are significantly regulated by the TFs is relatively big.

In real world, we do not know the magnitude of the systematic error in the feature matrix with respect to the relationship between motif abundance and TF binding. If the systematic error is very large, we would not expect the regression approach (Bussemaker et al., 2001, Keles et al., 2002, Conlon et al., 2003) to work well in detecting motifs. These studies imply that the assumption of a small or moderate systematic error is realistic in real data analysis.

For each TF, if the genes that are not significantly regulated by the TF dominates the experiment, the overall error rate in the estimated binding matrices tends to be small since the genes without necessary binding conditions and not significantly expressed are more accurately classified during the estimation procedure and they dominate the error rates. It is often true that a large proportion of genes are not significantly expressed in an actual DNA microarray experiment. This implies that the overall error rate should usually be small or moderate in real data analysis.

4. DATA ANALYSIS: TRANSCRIPTIONAL REGULATORY NETWORK IN YEAST

We apply our method to study the transcriptional regulatory network in *S. Cerevisiae* (yeast) based on analysis of a large collection of DNA microarray experiments.

4.1 *Data*

4.1.1 DNA microarray experiments We collect 658 DNA microarray experiments on yeast gene expression programs under various conditions: 7 on diauxic shift (DeRisi et al., 1997), 10 on sporulation (Chu et al., 1998), 60 on cell cycle (Spellman et al., 1998), 4 on adaptive evolution (Ferea et al., 1999), 173 on environmental stress (Gasch et al., 2000), 6 on Copper regulation (Gross et al., 2000), 300 on diverse mutations and chemical treatments (Hughes et al., 2000), 8 on Pho metabolism (Ogawa et al., 2000), 12 on SNF/SWI mutants (Sudarsanam et al., 2000), 26 on FKH1 and FKH2 roles during cell cycle (Zhu et al., 2000), and 52 on DNA damage (Gasch et al., 2001).

Prior to analysis, the data are normalized by subtracting the genome-wise median for every experiment. In addition, the log2-ratios are truncated by $\pm \log_2(20)$.

4.1.2 Promoter sequences We extract promoter sequences of 700 bps in length in the upper stream non-coding region [-700, -1] for 6136 ORFs using the SCPD database (Zhu and Zhang, 1999).

4.1.3 TF Binding Motifs We collect 46 yeast TFs with known binding sites from SCPD (Zhu and Zhang, 1999), TRANSFAC (Wingender et al., 1996), and YPD of Incyte Proteome BioKnowledge Library (Hodges et al., 1999) (see Table 2).

4.1.4 Constructing the feature matrix X The feature matrix X is constructed as described in Section 2.2 using the promoter sequence data and TF binding motif data. Note that a transcription factor may bind to a family of similar but distinct motifs. For example, the yeast transcription factor HSF1p binds the heat-shock dependent element which has at least four

Table 2. Some Yeast Transcription Factors and Their Specific Binding Motifs

TF	Binding Site	Site Name
ABF1	TCRNNNNNNACG	ABF1
ACE2	GCTGGT	ACE2
ADR1	TCTCC	ADR1
ATF1	ACGTCA	ATF
BAS2	TAATRA, TAANTAA	BAS2
CBF1	TCACGTG	CPF1
FKH2	GTMAACAA	SFF
FKH1	GTMAACAA	SFF
GAL4	CGGNNNNNNNNNNCCG	GAL4
GCN4	TGANTN	GCN4
GCR1	CWTCC	GCR1
HAP1	CGGNNNTANCGG	HAP1
HSF1	GAANNTCC, GAANNNTCC,	HSE
	TTCNNGAA, TTCNNNGAA	HSE
INO2	ATGTGAAWW	UASINO
INO4	ATGTGAAWW	UASINO
LEU3	CCGNNNNCGG, GGCNNNNGCC	LEU3
MAC1	GAGCAAA	CuRE
MATalpha2	CRTGTWWWW	MATalpha2
MBP1	WCGCGW	MCB
MCM1	CCNNNWWRGG	MCM1
MIG1	CCCCRNNWWWWW	MIG1
MSN2	AGGGG	STRE
MSN4	AGGGG	STRE
NDT80	CRCAAAW	MSE
PDR3	TCCGYGGA	PDR3
PHO4	CACGTK	PHO4
PUT3	CGGNNNNNNNNNNCCG	PUT3
PPR1	TTCGGNNNNNNCCGAA	PPR1
RAP1	RMACCCA	RAP1
REB1	YYACCCG	REB1
RFA1	TAGCCGCCGA	URS1
RFA2	TAGCCGCCGA	URS1
RFA3	TAGCCGCCGA	URS1
RME1	GAACCTCAA	RME1
ROX1	YYNATTGTTY	ROX1
RTG1	GGTCAC	RTG
RTG3	GGTCAC	RTG
STE12	TGAAACA	PRE
SWI4	CNCGAAA	SCB
SWI5	KGCTGR	SWI5
SWI6	CNCGAAA, WCGCGW	SCB/MCB
SUM1	CRCAAAW	MSE
TBP1	TATAWAW	TBP
TEA1	CGGNNNNNNNNNNCCG	TEA1
UME6	CTTCCT, TAGCCGCCGA	UARPHR/URS1
YAP1	TTANTAA	AP-1

Source: Compiled based on information from SCPD, TRANSFAC Database, and Incyte BioKnowledge Libarary (YPD).

Table 3. Estimated Number of Regulatory Interactions (with 6136 ORFs and 46 TFs)

Cut-off	Number of Genes Involved	Number of Interactions Total	Number of Interactions Per Gene	Number of Interactions Per TF
0.70	398	540	1.4	11.7
0.50	1225	2176	1.8	47.3
0.30	2465	7572	3.1	164.6
0.25	2929	10645	3.6	231.4
0.20	3599	15375	4.3	334.2

similar but distinct forms: `GAANNTCC`, `GAANNNTCC`, `TTCNNGAA`, or `TTCNNNGAA` (SCPD: Zhu and Zhang, 1999). Thus, we need to transform the feature matrix X by combining those columns that correspond to the same TF. As a result of this transformation, the columns of X map to distinct transcription factors.

4.2 *Analysis results*

We estimate the overall transcriptional regulatory interaction matrix by averaging over the 658 experiments and then use it to find the network motifs and overall network structure.

4.2.1 Estimated transcriptional regulatory interactions The estimated number of transcriptional regulatory interactions between TFs and genes is a function of cut-off value used. Table 3 shows the results at different cut-off levels.

We found that the weighted probability of regulatory interaction between a TF and its target gene often falls well below 0.5. One explanation is that, an active TF is likely to significantly regulate only a subset of its target genes, depending on specific experiment condition. In other words, a particular target gene of a TF may or may not be significantly regulated by the TF even when the TF is active. For example, our analysis of the α factor synchronized cell cycle data (Spellman et al., 1998) shows that, MBP1p is active in 17 out of the 18 time points, however, the yeast gene CDC2, a known target gene

induced by MBP1p, seems to be significantly regulated (i.e., probability of transcriptional regulatory interaction is ≥ 0.6) by MBP1p only at three time points (t=21, 70 and 77 minutes) with a probability of 0.676, 0.997 and 0.680 respectively. The probability of MBP1p-CDC2 interaction at the other time points is mostly less than 0.1. As a result, averaging over the 17 time points when MBP1p appears to be active brings down the weighted probability of MBP1p-CDC2 interaction to 0.253 (based on analysis of only the 18 α factor synchronized experiments).

We recommend selecting a cut-off such that the intensity of the estimated transcriptional regulatory interactions is comparable to those in published studies. In our analysis, we choose $c = 0.25$ as a cut-off, which is comparable to using $0.01 < p < 0.05$ as a P-value threshold in Lee et al. (2002). A larger and more stringent cut-off could be used, but it may reduce the power of the analysis to detect true TF-gene regulatory interactions.

4.2.2 Network motifs We found 4 autoregulated genes, 34 feed-forward loops, 0 multi-component loops, 23 single-input modules, 168 multi-input modules and 35 regulator chains, based on the estimated transcriptional regulatory interactions matrix for 46 TFs and 6136 genes, at a cut-off value of $c = 0.25$. All the findings are available on the supplement web site (http://www.stat.berkeley.edu/users/bxing/TRN/index.html (under construction)).

To assess the significance of the findings, we compared our results with published results from Lee et al. (2002). Our analysis involves 46 TFs, analysis of Lee et al. (2002) involves 106 TFs. We have 33 TFs in common. However, since the presence of additional TFs affects the finding of almost all the network motifs, particularly the single-input and multi-input modules and regulator chains (a result of the network motif finding algorithm). So the comparison focuses on only autoregulation motif and feed-forward loop motif.

At $c = 0.25$, we found 4 regulator genes (out of 46) that are likely to be autoregulated: ROX1, STE12, PDR3 and NDT80. Among these, STE12 was already identified as autoregulated in Lee et al. (2002) and Ren et al. (2000).

The ROX1 gene encodes a heme-induced repressor of hypoxic genes in yeast. Experiments indicated that ROX1p is capable of binding to its own upstream region and represses its own expression (Deckert et al., 1995). ROX1p was included in Lee et al. (2002), but was not identified as autoregulated.

NDT80p functions at pachytene of yeast gametogenesis (sporulation) to activate transcription of a set of genes required for both meiotic division and gamete formation. There is evidence that NDT80p activates its own transcription through an upstream MSE consensus site (Chu and Herskowitz, 1998; Lindgren et al., 2000).

The yeast PDR3 gene, which encodes a zinc finger transcription factor implicated in certain drug resistance phenomena, is under positive autoregulation by PDR3p. DNase I footprinting analyses using bacterially expressed PDR3p showed specific recognition by this protein of at least two upstream activating sequences in the PDR3 promoter (Delahodde et al., 1995; Simonics et al., 2000).

In addition to STE12, among the 33 common TFs involved in both analyses, SWI4, SUM1 and RAP1 were identified as autoregulated in Lee et al. (2002), but not in our analysis at the 0.25 cut-off level. At a lower cut-off level of 0.20, our analysis suggests SWI4 is autoregulated, but SUM1 and RAP1 are still not. Searching the literature, we did not found significant evidence that SUM1 is autoregulated. RAP1p is capable of binding to its own promoter, but it has been shown that the role of RAP1p in the transcriptional regulation of RAP1 may be very limited (Graham and Chambers, 1994).

We found 34 feed-forward loops involving 28 TFs at the 0.25 cut-off level. Among these, FKH2-ACE2, FKH2-SWI5, MCM1-SWI, MCM1-SWI5, were also identified in Lee et al. (2002).

4.2.3 Overall transcriptional regulatory network We assembled the overall yeast transcriptional regulatory network based on the estimated transcriptional regulatory interactions matrix for 46 TFs and 6136 ORFs. Figure 2 visualizes the overall network structure as a regulator interaction map.

The 31 nodes (boxes) shown are regulator genes that have estimated transcriptional regulatory interaction with either themselves (i.e., auto-regulation)

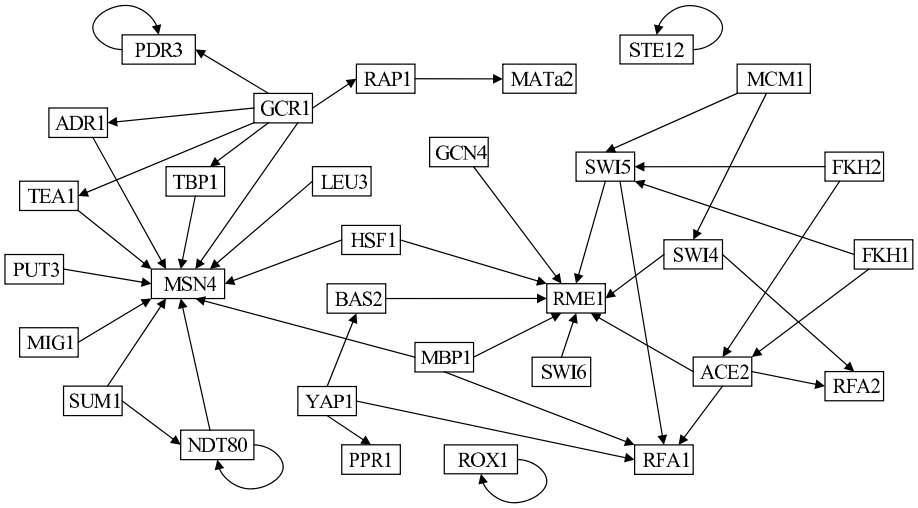

Figure 2. Yeast transcriptional regulatory network. Boxes indicate regulator genes. Arrows indicate the direction of regulatory interactions. Regulators without significant interaction with other regulators are not shown. The potential target genes of each regulator are not shown.

or other regulators. The other 15 regulators that are involved in the analysis but have no transcriptional regulatory interactions with any regulators are not shown. Each of the 46 TFs involved in the analysis has its own set of potential target genes, which are not shown in the graph either to make it clear.

The constructed network shows two sub-network structures: the right hand side part is related to the cell cycle process and the left hand side part is related to the stress-responsive regulation. This is a consequence of the data collection: a big proportion of the microarray experiments used in the analysis are on environmental stress response and cell cycle, and the transcription factors involved in the analysis cover only limited functional areas.

The analysis results show that the proposed statistical approach is capable of identifying important transcriptional regulatory network structures. For example, the constructed transcriptional regulatory network directly connects most of the regulators that are known to regulate the yeast cell cycle

process, such as MBP1, RME1, SWI4, SWI5, SWI6, ACE2, MCM1, FKH1 and FKH2, to form a sub-network for cell cycle regulation. Among the estimated cell cycle related transcriptional regulatory interactions, some have already been experimentally confirmed. For example, SWI5 and ACE2 both induce the meiosis repressor RME1 (Toone et al., 1995; McBride et al., 1999); MCM1 induces both SWI5 and SWI4 (Althoefer et al., 1995; Svetlov and Cooper, 1995; Fitch et al., 2003); MCM1 and FKH2 protein are both capable of binding SWI5 and ACE2 as determined by location analysis (Lee et al., 2002); MCM1 and FKH2 form a transcription factor complex to regulate cell-cycle dependent expression of the CLB2 cluster of genes, which include SWI5 and ACE2 (Boros et al., 2003).

The proposed method can not distinguish competitive binding. But it is capable of revealing the transcriptional regulatory network structure that is not obvious under a single experiment condition. For example, our analysis suggests that SUM1p transcriptionally regulates NDT80, and NDT80 is autoregulated. In fact, SUM1p and NDT80p bind competitively to the MSE sites in NDT80's promoter region and result in very different consequences: NDT80p activates the expression of NDT80, but SUM1p represses the expression of NDT80 (Pak and Segall, 2002). The cross link between SUM1 and NDT80 may not be observed in a location analysis based on only one kind of growth condition.

5. DISCUSSION

We described a comprehensive statistical approach for constructing the transcriptional regulatory network using data on gene expression, promoter sequence, and transcription factor binding sites. Our simulation studies show that the overall and false positive error rates in the estimated transcriptional regulatory network are expected to be small if the systematic noise in the constructed feature matrix is small. Our analysis based on 658 microarray experiments on yeast gene expression programs and 46 transcription factors suggests that the method is capable of identifying important transcriptional regulatory interactions and uncovering the corresponding network structures.

Our method is advantageous over some existing methods at least in the following aspects. The computaional approach is based on available gene

exression and sequence data, so it is time-wise and resource-wise more efficient than the experiment-based methods (e.g., location analysis). It is especially suitable for mining the fast accumulating microarray data on gene expressions under various experiment conditions. The method treats each microarray experiment as a partial realization of the overall transcriptional regulatory network process, which may be more appropriate and effective than the analysis based on perturbation experiments since a TF perturbation experiment does not guarantee that the gene expression changes are attributable solely and directly to the TF being perturbated. Moreover, as compared with the method based on location analysis data, the use of gene expression data may be more appropriate for modeling the transcriptional regulatory network since gene expression data is a direct result of a certain transcriptional regulatory network process while evidence of physical binding may not directly imply transcriptional regulation. Moreover, the location analysis data are typically obtained from a particular growth condition, which may limit the finding of the network structures that are specific to other conditions.

The method has at least two limitations. First, it may fail to estimate the regulatory interactions of a transcription factor that results in only subtle change in the genome-wide gene expression. Second, the method relies on knowledge of transcription factor binding sites. The number of TFs with known consensus binding sites is small and their functional coverage is somewhat limited. However, this may not be a problem when more and more TF binding sites are characterized and added to our knowledge. Also, we may use putative TF binding sites in the analysis. Using putative TF binding sites will increase the error rates in estimation, but the constructed network should suggest more models for further testing.

References

Althoefer, H., Schleiffer, A., Wassmann, K., Nordheim, A. and Ammerer, G. (1995). Mcm1 is required to coordinate g2-specific transcription in saccharomyces cerevisiae. *Mol Cell Biol* **15**, 5917–28.

Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert,

F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A. and Gifford, D. K. (2003). Computational discovery of gene modules and regulatory networks. *Nature Biotechnology* **21(11)**, 1337–42.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Roy Stat Soc Series B* **57**, 289–300.

Boros, J., Lim, F. L., Darieva, Z., Pic-Taylor, A., Harman, R., Morgan, B. A. and Sharrocks, A. D. (2003). Molecular determinants of the cell-cycle regulated mcm1p-fkh2p transcription factor complex. *Nucleic Acids Res* **31**, 2279–83.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A.

Bussemaker, H. J., Li, H. and Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nature Genetics* **27**, 167–171.

Chen, T., He, H. L. and Church, G. M. (1999). Modeling gene expression with differential equations. *Proc. Pac. Symp. Biocomput.* **4**, 29–40.

Chu, S., DeRisi, J., Eisen, M. B., Mulholland, J., Botstein, D., Brown, P. O. and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705.

Chu, S. and Herskowitz, I. (1998). Gametogenesis in yeast is regulated by a transcriptional cascade dependent on ndt80. *Mol Cell* **1**, 685–696.

Conlon, E. M., Liu, X. S., Lieb, J. D. and Liu, J. S. (2003). Integrating sequence motif discovery and microarray analysis. *Proc. Nat'l Acad. Sci.* **100**, 3339–44.

Deckert, J., Perini, R., Balasubramanian, B. and Zitomer, R. S. (1995). Multiple elements and auto-repression regulate rox1, a repressor of hypoxic genes in saccharomyces cerevisiae. *Genetics* **139**, 1149–58.

Delahodde, A., Delaveau, T. and Jacq, C. (1995). Positive autoregulation of the yeast transcription factor pdr3p, which is involved in control of drug resistance. *Mol Cell Biol* **15**, 4043–51.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal*

Statistical Society, Series B **34**, 1–38.

DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686.

D'Haeseleer, P., Liang, S. and Somogyi, R. (2000). Genetic network inference: From co-expression clustering to reverse engineering. *Bioinformatics* **16**, 707–726.

D'Haeseleer, P., Wen, X., Fuhrman, S. and Somogyi, R. (1999). Linear modeling of mrna expression levels during cns development and injury. *Proc. Pac. Symp. Biocomputing* **4**, 41–52.

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**, 71–103.

Ferea, T. L., Botstein, D., Brown, P. O. and Rosenzweig, R. F. (1999). Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc Natl Acad Sci* **96(17)**, 9721–6.

Fitch, M. J., Donato, J. J. and Tye, B. K. (2003). Mcm7, a subunit of the presumptive mcm helicase, modulates its own expression in conjunction with mcm1. *J Biol Chem* **278**, 25408–25416.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000). Using bayesian networks to analyze expression data. *J. Comput. Biol.* **7**, 601–620.

Gasch, A. P., Huang, M., Metzner, S., Botstein, D., Elledge, S. J. and Brown, P. O. (2001). Genomic expression responses to dna-damaging agents and the regulatory role of the yeast atr homolog mec1p. *Mol Biol Cell* **12**, 2987–3003.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D. and Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11**, 4241–57.

Graham, I. R. and Chambers, A. (1994). A reb1p-binding site is required for efficient activation of the yeast rap1 gene, but multiple binding sites for rap1p are not essential. *Mol Microbiol* **12**, 931–940.

Gross, C., Kelleher, M., Iyer, V. R., Brown, P. O. and Winge, D. R. (2000). Identification of the copper regulon in saccharomyces cerevisiae by dna microarrays. *J Biol Chem* **275**, 32310–6.

Hodges, P. E., McKee, A. H. Z., Davis, B. P., Payne, W. E. and Garrels, J. I. (1999). Yeast proteome database (ypd): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.* **27**, 69–73.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. and Friend, S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126.

Keles, S., van der Laan, M. J. and Eisen, M. B. (2002). Identification of regulatory elements using a feature selection method. *Bioinformatics* **18**, 1167–1175.

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

Liang, S., Fuhrman, S. and Somogyi, R. (1998). Reveal: A general reverse engineering algorithm for inference of genetic network architectures. *Proc. Pac. Symp. Biocomput.* **3**, 18–29.

Lindgren, A., Bungard, D., Pierce, M., Xie, J., Vershon, A. and Winter, E. (2000). The pachytene checkpoint in saccharomyces cerevisiae requires the sum1 transcriptional repressor. *Embo Journal* **19**, 6489–6497.

McBride, H. J., Yu, Y. and Stillman, D. J. (1999). Distinct regions of the swi5 and ace2 transcription factors are required for specific gene activation. *J Biol Chem* **274**, 21029–21036.

Ogawa, N., DeRisi, J. and Brown, P. O. (2000). New components of a system for phosphate accumulation and polyphosphate metabolism in saccharomyces cerevisiae revealed by genomic expression analysis. *Mol Biol Cell* **11**, 4309–4321.

Pak, J. and Segall, J. (2002). Regulation of the premiddle and middle phases of expression of the ndt80 gene during sporulation of saccharomyces cere-

visiae. *Mol Cell Biol* **22**, 6417–29.

Qian, J., Lin, J., Luscombe, N. M., Yu, H. and Gerstein, M. (2003). Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics* **19**, 1917–1926.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P. and Young, R. A. (2000). Genome-wide location and function of dna binding proteins. *Science* **290**, 2306–9.

Simonics, T., Kozovska, Z., Michalkova-Papajova, D., Delahodde, A., Jacq, C. and Subik, J. (2000). Isolation and molecular characterization of the carboxy-terminal pdr3 mutants in saccharomyces cerevisiae. *Curr Genet* **38**, 248–255.

Somogyi, R., Fuhrman, S., Askenazi, M. and Wuensche, A. (1997). The gene expression matrix: Towards the extraction of genetic network architectures. *Proc. of Second World Congress of Nonlinear Analysts* **30(3)**, 1815–1824.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* **9**, 3273–97.

Storey, J. D. (2003). The positive false discovery rate: A bayesian interpretation and the q-value. *Annals of Statistics* **31**, 2013–2035.

Sudarsanam, P., Iyer, V. R., Brown, P. O. and Winston, F. (2000). Whole-genome expression analysis of snf/swi mutants of saccharomyces cerevisiae. *Proc Natl Acad Sci* **97**, 3364–9.

Svetlov, V. V. and Cooper, T. G. (1995). Review: compilation and characteristics of dedicated transcription factors in saccharomyces cerevisiae. *Yeast* **11**, 1439–84.

Toone, W. M., Johnson, A. L., Banks, G. R., Toyn, J. H., Stuart, D., Wittenberg, C. and Johnston, L. H. (1995). Rme1, a negative regulator of meiosis, is also a positive activator of g1 cyclin gene expression. *Embo Journal* **14**, 5824–32.

van der Laan, M. J. and Dudoit, S. (2003). Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive

epsilon-net estimator: Finite sample oracle inequalities and examples. *U.C. Berkeley Division of Biostatistics Working Paper Series* **130**.

van der Laan, M. J., Dudoit, S. and Pollard, K. S. (2004). Multiple testing. part iii. procedures for control of the generalized family-wise error rate and proportion of false positives. *U.C. Berkeley Division of Biostatistics Working Paper Series* **141**.

Wang, W., Cherry, J. M., Botstein, D. and Li, H. (2002). A systematic approach to reconstructing transcription networks in saccharomyces cerevisiae. *Proc. Natl. Acad. Sci.* **99**, 16893–98.

Wingender, E., Dietze, P., Karas, H. and Knppel, R. (1996). Transfac: A database on transcription factors and their dna binding sites. *Nucleic Acids Res.* **24**, 238–241.

Yoo, C., Thorsson, V. and Cooper, G. F. (2002). Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational dna microarray data. *Proc. Pac. Symp. Biocomput.* **7**, 498–509.

Zhu, G., Spellman, P. T., Volpe, T., Brown, P. O., Botstein, D., Davis, T. N. and Futcher, B. (2000). Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature* **406**, 90–94.

Zhu, J. and Zhang, M. Q. (1999). Scpd: A promoter database of yeast saccharomyces cerevisiae. *Bioinformatics* **15**, 607–611.